ENGINEERED HIGH-THROUGHPUT DESIGN OF WHOLE-CELL BIOSENSORS FOR MICROBIAL PRODUCTION OF VALUE-ADDED PRODUCTS

A Dissertation

Presented to

The Faculty of the Department of Chemical Engineering

University of Houston

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy

in Chemical Engineering

By

Christopher Scott Frei

December 2015

ENGINEERED HIGH-THROUGHPUT DESIGN OF WHOLE-CELL BIOSENSORS FOR MICROBIAL PRODUCTION OF VALUE-ADDED PRODUCTS

Christopher S. Frei

APPROVED:

Patrick C. Cirino, Ph.D.
Committee Chair
Associate Professor

COMMITTEE MEMBERS:

Navin Varadarajan, Ph.D.
Assistant Professor

Richard C. Willson, Ph.D.
Huffington-Woestemeyer Professor

Timothy Cooper, Ph.D.
Associate Professor
Department of Biology and Biochemistry

Laura Segatori, Ph.D.
Associate Professor at Rice University
Department of Chemical and Biomolecular Engineering

Suresh Khator, Ph.D., P.E.
Associate Dean of Graduate Programs

Michael Harold, Ph.D.
Chair of Chemical and Biomolecular Engineering

ENGINEERED HIGH-THROUGHPUT DESIGN OF WHOLE-CELL BIOSENSORS FOR MICROBIAL PRODUCTION OF VALUE-ADDED PRODUCTS

An Abstract

Of a

Dissertation

Presented to

The Faculty of the Department of Chemical Engineering

University of Houston

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy

By

Christopher Scott Frei

December 2015

ABSTRACT

Humans have developed an overwhelming dependence on synthetic materials, which are often costly and environmentally harmful. Over the past decade, there has been a boom in biotechnology, with one emphasis on engaging easily manipulated microorganisms for relatively inexpensive industrial scale production of pharmaceuticals, polymers, and fuels. This challenging field of research offers invaluable benefits to society and the environment. Rapid and precise screening of large libraries of genetically altered microorganisms for enhanced molecule production is a powerful approach to developing such "microbial factories." Unfortunately, the lack of readily available high-throughput screening techniques inhibits our ability to quickly engineer microorganisms. This limitation is addressed here using engineered whole-cell molecular biosensors based on a family of proteins known as transcriptional regulatory proteins (TRPs). The natural role of many TRPs is to link molecule recognition with gene expression, making them ideal candidates for engineering endogenous molecule biosensors. Through powerful approaches such as directed evolution, TRPs can be altered to recognize targeted value–added molecules and their precursors. Upon recognition of the target molecule, the TRP activates expression of genes with an easily measureable phenotype (e.g., luminescence, cell growth, fluorescence). In the present study, we sought to (i) improve the current screening strategies of TRP libraries, (ii) investigate residue relationships governing the recognition and response of a TRP, and (iii) isolate novel whole-cell biosensors based on the TRP platform.

High-throughput screening techniques for isolating functional clones from large genetic libraries ($> 10^6$ mutants) is pivotal to the continued success of engineering microbial factories. Here, we applied fluorescence-activated cell sorting (FACS) combined with antibiotic selections to dramatically improve the throughput of screening large libraries of AraC, an *Escherichia coli* native TRP. After residue characterization and screening optimization, several functional AraC variants were isolated with desirable specificity and sensitivity toward target molecules, vanillin

and salicylate. As we continue to characterize new biosensors and optimize their design process, the limits of TRP molecule recognition are pushed further, thus allowing us to overcome the restraints imposed by natural TRPs and offering sustainable solutions to engineering microbial factories.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I Engineering microbial production of value-added chemicals and biofuels

I.1) Introduction

This chapter will provide a brief history of the design and applications of *in vivo* molecular biosensors for high-throughput screening and microorganism development, predominantly focusing on bacteria. As a foundation for the work presented in subsequent chapters, the fields of metabolic engineering and protein engineering are reviewed and help to illustrate the current need for molecular biosensors. Specific examples of engineered microorganisms and specialized screening techniques are mentioned throughout the chapter to emphasize advances and applications in each field. Owing to the immense growth in these areas of research over the past two decades, this introduction chapter focuses on those topics that are most relevant to the subsequent chapters in this dissertation. The fields of metabolic and protein engineering are young but constantly growing and developing (Figure I-1), and as new challenges emerge, new technologies will be invented, altering our knowledge and shaping our imaginations as scientists and engineers.

I.2) Metabolic Engineering

Manipulations of microorganisms for the production of natural products (microbial factories) was first termed "metabolic engineering" in the early 1990s (Bailey, 1991; Stephanopoulos and Vallino, 1991). Since then, the field has made great advances towards being directly competitive with classical synthetic chemistry and has the potential to grow beyond the molecular repository accessible by chemical synthesis, owing to the sophisticated reactions catalyzed by enzymes. Despite the growing interest and promising results in metabolic

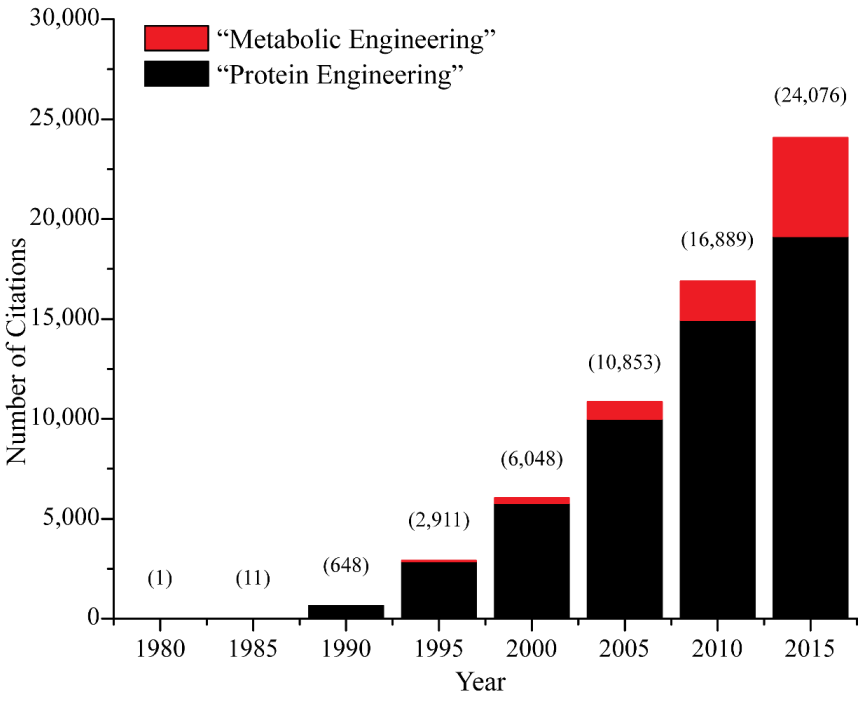


Figure I-1 An increasing interest in microbial engineering is evident by the total number of citations listed on PubMed.gov mentioning either "Protein Engineering" or "Metabolic Engineering" in either the abstract or the title over the past three and a half decades. Numbers in parentheses above bars indicate the total number of citations for "Protein Engineering" and "Metabolic Engineering" combined.

engineering, manipulations of complex cellular networks continues to prove challenging. Therefore, each target product engineered through microbial development requires a balanced consideration of several important factors: (1) the cost and availability of the raw materials; (2) the suitability of a host organism; (3) the methods for genetic manipulations; (4) the degree of genetic control; and (5) the most efficient metabolic pathway (Keasling, 2010). The metabolic engineer must make an effort to address each of these and determine the trade-offs when maximizing product yield. Discussed here is a broad overview of these factors and the delicate links between them. For a more extensive discussion on the field of metabolic engineering, several well-written reviews are available in the literature (Domach, 2015; Hara et al., 2014; Keasling, 2010; Woolston et al., 2013).

In recent years (2010-2013), biopharmaceuticals, defined as recombinant proteins and nucleic acid based pharmaceutical products from microorganisms, have made up approximately a quarter of all new pharmaceuticals approved by the U.S. and Europe (averaging USD 121.5 billion per year). Also, nearly a quarter of all licensed biopharmaceuticals (not including withdrawals) were produced from a single microorganism, *Escherichia coli* (Walsh, 2014) (see Supp. Table A-1 for a complete list of *E. coli* produced biopharmaceuticals as of 2014). In addition to biopharmaceuticals, recombinant DNA technologies have led to increased production of bio-based fuels, chemicals, and materials (i.e., polymers). For example over the past decade, bioethanol production has increased 421% (81,058 Mbbl to 341,414 Mbbl), and biodiesel production increased 4,431% (666.2 Mbbl to 29,522.8 Mbbl). Additionally, a joint venture between BASF and Corbion Purac formed the company Succinity in 2013 for production of succinic acid from *Basfia succiniciprodcens*. Their inaugural plant for commercial production of succinic acid has an annual capacity of 10,000 metric tons and a second plant is projected to produce up to 50,000 metric tons annually. Succinic acid is used to produce biodegradable polymers, polyurethanes, industrial solvents, pigments, and plasticizers. Therefore, the rapidly growing bio-based chemical and pharmaceutical industries provide strong incentives for the scientific community to invest their time, resources, and funding, as is further discussed below.

1.2.1) Bio-based petroleum substitutes

One area of recent concern addressed by metabolic engineering is our strong dependence on petroleum-based products, which has left us vulnerable to the price fluctuations and uncertainties of the petroleum industry. Also, increased awareness of greenhouse gas emission and recent legislation requiring a reduction in these gases over the next decade and a half have also provided a drive for the scientific community to develop alternative fuels that can be easily integrated into our current infrastructure as well as reduce emissions and pollutants. Petroleum-based products are still dominating the market, but the necessity for alternatives is clear and is

currently being addressed through various technological advances. Microbes such as *Saccharomyces cerevisiae* have been a source of attention for thousands of years owing to their ability to ferment ethanol and produce alcoholic beverages, creating two ethanol and two carbon dioxide molecules for every molecule of glucose consumed. The conversion of biomass, predominantly composed of complex sugars and lignin, to bioethanol and its subsequent use as an alternative fuel have been given elaborate thought through several economic, microbial energetics, and environmental impact studies (Condon et al., 2015; Salehi Jouzani and Taherzadeh, 2015). The hygroscopic properties of ethanol, effects on gasoline vapor pressure, and decreased overall fuel economy have led to numerous studies to determine the feasibility of microbial production of biodiesel, butanol, hydrogen and other biosynthetic fuels (Choi and Lee, 2013; Khatri et al., 2014; Lan and Liao, 2011; Petrovič, 2015; Valle-Rodriguez et al., 2014; Wu and San, 2014; Zhang et al., 2011; Zhou et al., 2015).

One of the many challenges associated with bio-based production is the limited quantities and large processing costs that currently cannot compete with chemical based syntheses due to the complex interactions and difficulties in selectively controlling and maximizing the desired bio-product. One recent example, vanillin (4-hydroxy-3-methoxybenzaldehyde), has gained attention for its use as a building block for biopolymers (e.g. polyesters, polyacetals, epoxys) (Fache et al., 2015). Vanilla and vanillin extract production is estimated to be 12,000 metric tons annually, worth approximately three quarter of a billion USD (Gounaris, 2010; Wenda et al., 2011). Vanillin can be naturally extracted from the bean pods of *Vanilla planifolia* orchids (<1% of vanillin production) or from processed lignin. Unfortunately, vanilla pods are a poor source of vanillin due to their slow growth and low titers (up to 2% w/w), so vanillin is predominantly produced chemically from a petroleum byproduct, guaiacol (Figure I-2). However, strict regulations prevents the labelling of vanillin produced using chemical synthesis as "natural vanilla," making it less appealing to consumers (Walton et al., 2000).

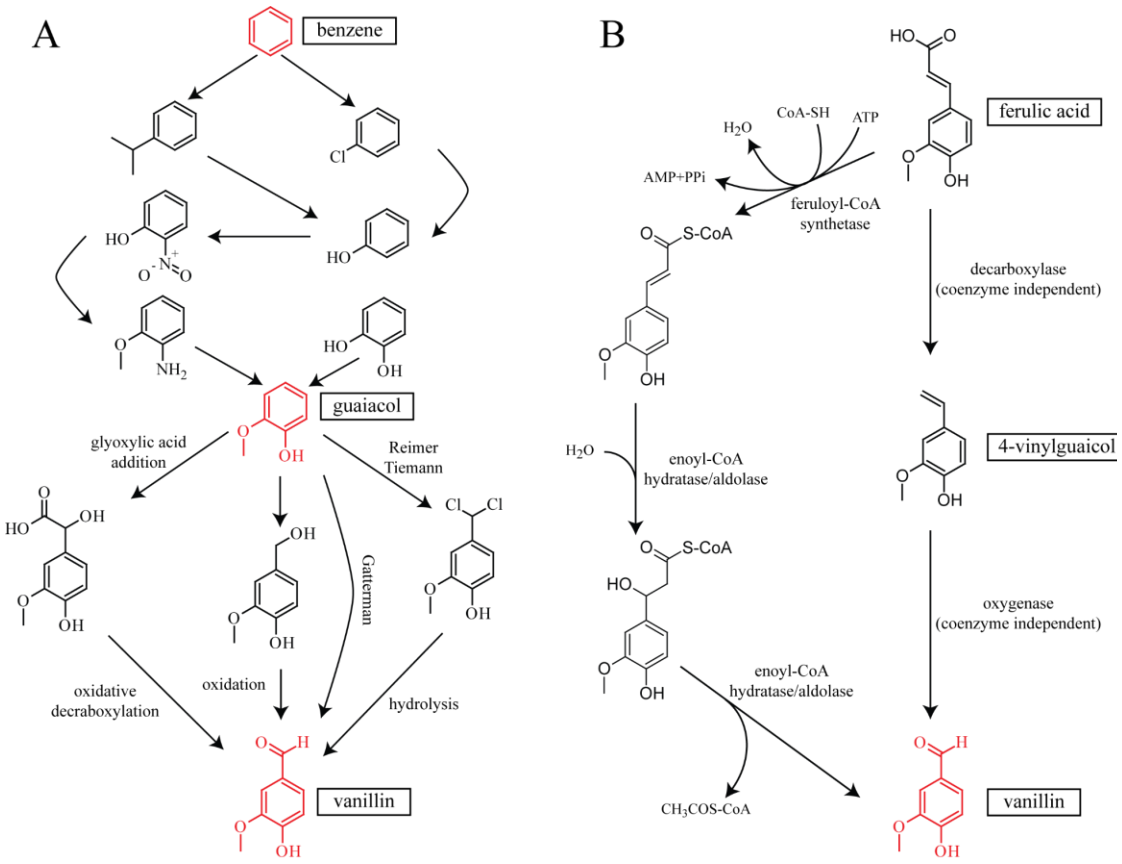


Figure I-2 Flow diagram for production of vanillin from (A) petroleum sources and from (B) metabolic pathways.

The most abundant and available feedstocks for vanillin production from microorganisms are ferulic acid and eugenol. Amongst a variety of host microorganisms that can convert ferulic acid to vanillin, unmodified *Actinomyces* strains have been characterized as efficient producers of vanillin, but they also contain the metabolic machinery to degrade vanillin to vanillic acid or guaiacol. However, microbes such as *E. coli* do not naturally possess the ability to metabolize vanillin, although it has been noted in the literature that some *E. coli* strains will degrade vanillin to vanillic acid or vanillyl alcohol (Barghini et al., 2007). Kim and coworkers were successful in producing vanillin in *E. coli* by expressing heterologous feruloyl-CoA synthetase (Fcs) and enoyl-CoA hydratase/aldolase (Ech) from *Amycolatopsis* sp. Strain HR104 for the three step reaction of converting ferulic acid to vanillin (Yoon et al., 2005). Fcs converts ferulic acid to

feruloyl-CoA and Ech converts feruloyl-CoA to vanillin through a two-step reaction, with intermediate 4-hydroxy-3-methoxyphenyl-β-hydroxypropionyl-CoA. Most likely due to the complex reactions involving coenzymes, they only achieved a titer of 1.1 g L^{-1}. Another group expressed an artificial vanillin pathways in a recombinant strain of *E. coli* but using coenzyme-independent enzymes. A ferulic acid decarboxylase from *Bacillus pumilus* was expressed to convert ferulic acid to 4-vinylguaiacol and a carotenoid oxygenase from *Caulobacter segnis* was coexpressed to convert 4-vinylguaiacol to vanillin in *E. coli* (Furuya et al., 2014). After optimizing the pH, a titer of 1.2 g L^{-1} vanillin was produced. The authors noted that the intermediate, 4-vinylguaiacol, accumulated during the first hour of reaction and was determined to be the rate-limiting reaction for conversion of 4-vinylguaiacol to vanillin. Later, the same group achieved the highest titer recorded in the literature for vanillin production in recombinant *E. coli* (7.8 g L^{-1}) using a two-pot bioprocess, where each reaction was separately cultured and optimized for maximum rate of conversion (Furuya et al., 2015). To date, vanillin has been "bio"-produced from several starting compounds including lignin, guaiacol, eugenol, isoeugenol, and aromatic acids (e.g. ferulic acid) (Wenda et al., 2011). Although vanillin production in *E. coli* is promising, it is in direct competition with companies like Evolva and Solvay, having industrial scale production of vanillin from engineered yeast. Independent of which host organism, "natural" product targets would ideally seamlessly transition into production at current petroleum processing facilities, much like vanillin which is already produced by petroleum industry. Microbial production of these alternative fuels and polymers from biomass is currently one of the most economic and promising technology to alleviate some of our dependence on petroleum-based chemicals, but direct competition with petroleum-based products requires operation with tight and often unfeasible profit margins. Therefore, optimizing the aforementioned oxygenase to have a higher catalytic efficiency would increase the production of vanillin without the use of a costly two-pot system. With the help of a sensitive and selective

endogenous vanillin biosensor, the oxygenase can be subjected to directed evolution techniques, as will be discussed in subsequent sections.

1.2.2) Natural production of active pharmaceutical ingredients

Unlike alternatives to petroleum-based molecules, production of active pharmaceutical ingredients (APIs) from biosynthetic methods do not directly compete with chemical synthesis methods. The complex structures found in this class of molecules directly inhibits economical chemical synthesis. Isoprenoids, alkaloids, and polyketides are among the main classes of APIs currently being investigated by synthetic biologists and metabolic engineers (Dixon and Steele, 1999; Luo et al., 2014; Roessner and Scott, 1996). Polyketides and nonribosomal peptides have garnered particular attention due to their wide range of biological activities, including cytostatic, immunosuppressant, insecticidal, and antibacterial. However, the growing concern of drug resistant bacteria has led to highly publicized debates on the ethics of using these molecules as antibiotics in livestock and insecticides on crops, and has fueled the need for novel APIs with higher potency. For example, polyketides have been targeted due to their bioactivity and modularity. Naturally derived from slow growing and complex organisms, polyketide production has been engineered in native and non-native hosts through recombinant DNA technologies (Asai et al., 2015; Gao et al., 2010; Weber et al., 2015; Yuzawa et al., 2012; Zhou et al., 2008). Polyketides are derived from small molecule building blocks (i.e. acyl-CoA group members) assembled by large enzymatic clusters called polyketide synthases (classified in three groups: Type I, Type II and Type III polyketide synthases) (Shen, 2003). The modularity of the polyketide synthases (PKSs) provides a means to create an innumerable array of molecules inaccessible by chemical synthesis.

Methods such as transformation-associated recombination (TAR) (Larionov and Kouprina, 2008) and linear plus linear homologous recombination (LLHR) (Fu et al., 2012) have been used to isolate PKS pathways from actinobacteria for heterologous expression and easy manipulation

of individual modules using standard recombinant techniques. Using these techniques along with λ red-mediated recombination in *E. coli* (Zhang et al., 1998), a PKS was isolated from *Streptomyces* sp. Tü6071 and heterologously expressed in *Streptomyces lividans* and *Streptomyces coelicolor*, resulting in production of phenalinolactones A and D up to 10 μg L^{-1} (Binz et al., 2008). Some actinobacteria strains have been engineered as "super-hosts" for heterologous expression of PKSs by deletion of unnecessary PKS pathways. *Streptomyces coelicolor* M145 strain was engineered as a "super-host" after deletion of pathways for production of actinorhodin, prodiginine, Type I polyketide CPA, and calcium-dependent antibiotics. Upon heterologous expression of PKSs for production of chloramphenicol and conginocidine, they were able to obtain 20-40 fold higher concentrations in the engineered strain over the parent strain M145 (Gomez-Escribano and Bibb, 2011). Despite the success of engineering PKSs in actinobacteria, the slow growth kinetics and diverse genomic content of actinobacteria strains are less than ideal for engineering polyketide production.

Tang and coworkers successfully expressed a functional minimal Type II PKS from PKS4 of *Gibberella fujikuroi* in *E. coli* leading to the first example of Type II aromatic polyketide production in *E. coli* (Zhang et al., 2008). They were able to obtain a titer of 3 mg L^{-1} of the anthraquinone compound SEK26 after scale-up in a fed-batch fermentation. However, Type II PKS expression in *E. coli* has been difficult, but greater success has been seen with engineering Type I and Type III PKS expression in *E. coli* for the production of alkanes and alkenes, flavonoids, and statins (Leonard et al., 2008; Liu et al., 2015; Xie et al., 2007). Though microorganisms are far more efficient at producing polyketides over chemical synthesis, their complex biosynthetic pathways render them difficult to engineer. Regardless, the availability of interchangeable modules, enzymatic mechanism characterization, and advanced biotechniques are allowing for rapid growth in polyketide production from microorganisms (Hertweck, 2015). In particular, whole-cell biosensors with high specificity towards 2-pyrone derailment products (e.g.,

triacetic acid lactone) of polyketides synthases can be used to engineer individual or multiple enzymes in the pathway, helping overcome low titers and low solubility in an ideal host microorganism.

1.2.3) Host selection for engineering

The abundance of carbon sources, selective pressures, and environmental habitats available in nature has created a vast array of microbes with unique metabolic pathways, however, some microbes are more tolerant to genetic manipulations (as was mentioned in the previous section), more efficient in using specific carbon sources, and less vulnerable to toxic compounds. The choice of a host microbe for the production of a target compound was typically the native producing organism in the early years of metabolic engineering because this organism already contains the proper metabolic pathway, but today, advances in genetic engineering have helped shift the host organism towards a handful of amenable microbes. These microbes have one if not several qualities that make them ideal candidates: maintain survival and growth under a variety of process conditions, are malleable to genetic manipulations, and the genetic tools are available and well characterized. *E. coli* has been the preferred workhorse organism of industry and academic microbiology (Waegeman and Soetaert, 2011), owing to its ability to be cultivated under aerobic and anaerobic conditions, tolerate moderate temperature and pH ranges, grow quickly, and consume a wide variety of carbons sources. This has led to in depth characterization of *E. coli* and the availability of many genetic tools for manipulating *E. coli*. Despite this, *E. coli* is not always the best host organism. For example it lacks machinery for glycosylation, whereas eukaryotic organisms like *S. cerevisiae* do not. *S. cerevisiae* also tolerates toxic effects of certain molecules like ethanol better than *E. coli*. Thus, the selection of a proper host is directly linked with the production of the target compound.

1.2.4) Biotechniques for genetic manipulations

Recombinant deoxyribose nucleic acid (DNA) cloning describes an invaluable set of genetic tools for enzymatic cleaving and recombining of DNA to be expressed in a host organism. The concept of heterologous host expression systems was first realized by Cohen and coworkers in the seminal paper on recombinant DNA cloning in 1973 (Cohen et al., 1973), which birthed genetic engineering and later metabolic engineering. Shortly thereafter, the first licensed drug from recombinant DNA technology, insulin, was produced in *E. coli* and marketed by Eli Lilly in 1982 (Baeshen et al., 2014).

Central to genetic manipulations is the vector for delivery of heterologous genes and the control over expression of those genes. First, delivery vectors (e.g., plasmids, cosmids, bacterial artificial chromosomes, and yeast artificial chromosomes) provide a platform for genetic manipulations. In a classic example of recombinant cloning, DNA containing the genetic element of interest is amplified using a polymerase chain reaction (PCR). The PCR product and the vector are cleaved with restriction enzymes, which result in double stranded DNA fragments possessing compatible ends. The fragments are then ligated together using a ligase enzyme and the final ligated product is transfected into and propagated in a microorganism.

Genes can be expressed directly from the vector(s) or can be integrated into the host genome (Datsenko and Wanner, 2000; Haldimann and Wanner, 2001). Entire pathways with numerous genes can be expressed simultaneously, but often precise control over gene expression is necessary to avoid toxic effects of heterologous protein expression (Bienick et al., 2014). Current genetic tools allow for the tunable expression of genes and gene clusters using synthetic promoters with predictable expression rates (Salis et al., 2009) and inducible allosteric regulation to exogenously control gene expression (Pareja et al., 2006). Recently, complex Boolean-type genetic gates have been designed based on the availability of allosteric regulators to mimic natural and unnatural processes (Stanton et al., 2014). Also, the frequency of the gene in the host

organism affects the expression rate of a gene. Integration into the genome ensures that each cell contains the same frequency of each gene, but when expressed from extrachromosomal DNA (i.e., the vector), stability of the DNA copy number becomes an issue and should be taken into account when designing a microbial factories. As the cell divides, the number of copies could be divided unequally amongst the two progenies, potentially resulting in a biased phenotype. Still, expression from a vector is attractive due to the availability of techniques for easy genetic manipulations of the extrachromosomal DNA. For example, a digital microfluidic DNA assembly chip was recently developed for the precise mixing of DNA fragments (i.e., ~0.2 μL could be accurately dispensed from a reservoir), ligation, and finally electroporation of the ligated product into a host organism (Shih et al., 2015). All three steps were included on a single chip, which minimizes the loss of reagents, DNA template, and enzyme. This technique was used to successfully carry out Golden Gate assembly, Gibson assembly, and yeast assembly.

Even with the recent advances in metabolic engineering, it is a relatively new science and only starting to give viable alternatives to chemical synthesis techniques that have been around for hundreds of years. The boom in biotechnology over recent decades is only the tip of the iceberg. Future discoveries are relying on our ability to characterize and engineer novel pathways with prevailing high-throughput technologies for evolving novel enzymes, microbial behavior, and metabolic fluxes.

I.3) Directed Evolution of Proteins for Rapid Microbial Development

Proteins have been engineered to overcome numerous obstacles presented in metabolic engineering, including poor stability under various process conditions (e.g., high pressure, detergent, pH, temperature), low or no catalytic activity towards a desired substrate, toxicity to the host cells, and thermodynamic and steric restraints. Over the past two decades, numerous techniques have been developed or improved to manipulate proteins and their functions: (1) DNA

sequencing has led to the discovery of countless new enzymes and genetic elements, (2) recombinant technologies now allow us to readily manipulate expression levels, mutations, and even entire synthetic metabolic pathways, (3) growing biotechnology databases (e.g., proteins, DNA, small molecules, literature) provide users with access to a wealth of information to help guide the design of microbial factories, (4) protein structure analysis through computational software and crystallography have shaped our understanding of biomolecular folding and interactions with other biomolecules, and (5) high-throughput screening technologies continue to push the limitations of rapid artificial evolution of genetic elements, cellular phenotypes, and enzymes. Yet even with the advances in biotechnology, microorganisms are the most complex and sophisticated "reactors" available and are not fully understood. Therefore as approaches to engineering proteins become more accessible and customizable, we can continue to develop a deeper understanding of the relative fitness of a microorganism towards production of desired compound. An overview of popular techniques and their successful applications in the field of protein engineering are discussed below.

1.3.1) Directed Evolution

Directed evolution combines Darwinian evolution with iterative rounds of phenotypic-based screening to analyze designed mutant libraries of enzymes and other genetic elements with the ultimate goal of discovering novel desirable traits (Cobb et al., 2013). The concept of artificial evolution has been around since the 1960s, when Lerner and coworkers isolated strains of *Aerobacter aerogenes* with improved utilization of xylitol as a growth substrate, after they chemically induced non-specific mutations in the genome of the strain (Lerner et al., 1964). In recent years, directed evolution has emerged as the dominating method for protein engineering and can be broken down into four parts: (i) determination of the rate-limiting reaction, (ii) *in vitro* or *in vivo* library construction, (iii) transformation into a host organism, and (iv) screening for a desirable genetic trait linked to a measurable phenotype (Figure I-3). Often, directed evolution

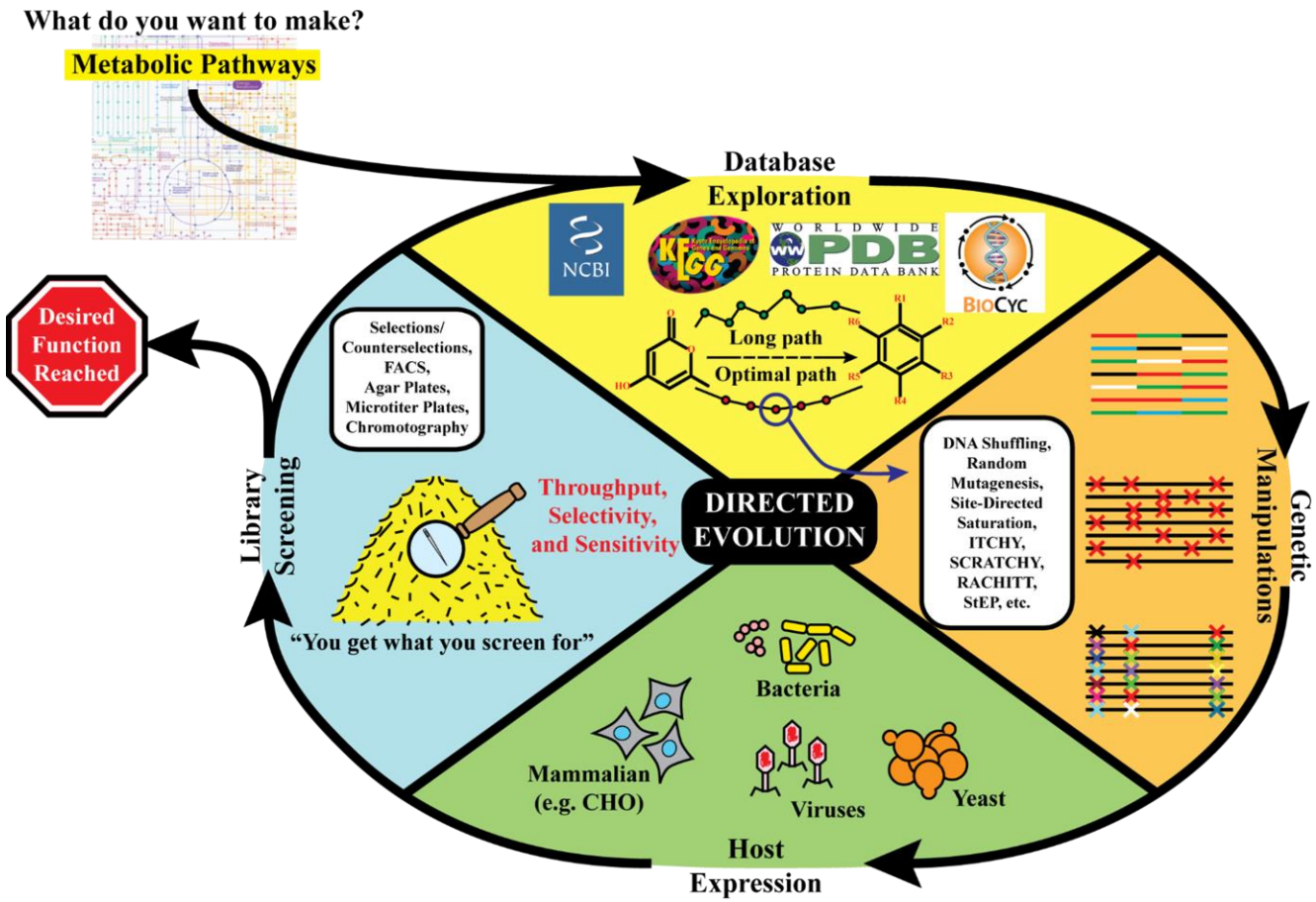


Figure I-3 Overview of combinatorial design and directed evolution of proteins.

starts with selection of a template protein of known structure and has similar function to the target function. Despite some success in the *de novo* design of proteins (Tinberg et al., 2013), constructing artificial proteins remains challenging (Urvoas et al., 2012), thus making engineering novel proteins from naturally occurring proteins more attractive.

1.3.2) Principles governing library design

Two conflicting criteria guide the construction of genetic libraries. First, library members must be conservative in sequence space relative to the native protein. Dramatic changes in the sequence of a protein may result in the majority of clones being non-functional. There are an astronomical number of conformations that a protein could fold into (Levintha.C, 1968), and even a small disturbance in the sequence space could lead to a deleterious effect by altering the stability of the folded protein (Tokuriki and Tawfik, 2009). Second, enough diversity must be introduced into the protein sequence space to sufficiently explore the functional fitness landscape of a protein. This can be best described as a topological map, which relates sequence space with fitness (i.e., desirable function) and elevation represents the level of fitness (Romero and Arnold, 2009). Certain mutations will be deleterious (loss in elevation) and others will be advantageous (gain in elevation), yet superior fitness may result from cooperativity of multiple mutations (both deleterious and advantageous). The study of cooperativity and interactions between residues and how they relate to the evolution of a protein function is called epistasis. An excellent example of conformational epistasis, specifically referring to the repositioning of residues after a substitution has been made, was demonstrated by Ortlund and coworkers (Ortlund et al., 2007). They observed that the evolution of a vertebrate steroid receptor with high specificity was evolved from a promiscuous steroid receptor through multiple groups of interacting residue substitutions. A total of eight residue substitutions were found that together lead to increased specificity, but separately, most lead to deleterious effects. The mutations were broken up into three groups (X, Y, and Z groups). The X and Y groups yield the final specificity towards the steroid, cortisol, but

result in an unstable protein. Group Z substitutions stabilize the protein and allow for a fully functional and highly specific cortisone receptor (Ortlund et al., 2007). This example demonstrates the importance of simultaneous mutations, which should be strongly considered during library design despite the exponential increase in degeneracy of a library.

Statistics govern the effectiveness of library design. Each part of library design (i.e., construction, transformation, and screening of the library) bear weight on the theoretical number of clones observed, as is shown below. A simple model for determining the degeneracy (D) of a library based on a Poisson distribution, assuming all members of a library are equally represented, was proposed by Bosley and Ostermeier (Bosley and Ostermeier, 2005):

$$D = D_{max}(1 - e^{-\frac{T}{D_{max}}}). \qquad \textbf{I-1}$$

The maximum degeneracy (D_{max}) is the total possible members of a library from the library creation technique assuming infinite number of transformants (T). It is important to note the distinction between the degeneracy on the DNA level ($D_{max,DNA}$) and the degeneracy at the protein level ($D_{max,protein}$). A library with five codons completely saturated with NNN sites (N = A, T, G, or C) results in a $D_{max,DNA}$ equal to 64^5 (or 1.1×10^9) unique DNA sequences, whereas $D_{max,protein}$ is equal to 20^5 (or 3.2×10^6) unique proteins. The degeneracy of a library is most often referenced at the DNA level because the response of a particular clone is not only dependent upon the function of the protein variant but also the concentration of accessible protein, as determined by the transcription and translation rates. Microorganisms show bias towards particular codons based on various factors such as tRNA availability and environmental stresses (Gustafsson et al., 2004), which directly affects the translation rate.

In addition to degeneracy, the completeness of a library (P_c) is crucial for evaluating library design (Bosley and Ostermeier, 2005) and is described by the equation

$$P_c = \prod_{i=1}^{D} P_i \,, \qquad \textbf{I-2}$$

where the probability of a particular clone (P_i) is represented as

$$P_i = 1 - (1 - F_i)^T \,. \qquad \textbf{I-3}$$

The library completeness is dependent on the frequency of each clone (i) present in the library (F_i), T, and D, and if every member of the library has equal probability of being represented, P_c can be simplified to

$$P_c = [1 - (1 - F)^T]^{D_{max}} \,. \qquad \textbf{I-4}$$

From this, the minimum number of transformants needed for a level of completeness can be solved for, which should dictate the total number of transformations needed. However, the above equations should be used with the understanding that they are guides to help in library design and are not absolutely representative of the library, because of the complexities and biases introduced through various library design techniques.

1.3.3) Techniques for library creation

Protein mutations are often introduced at the DNA level in the corresponding gene, which takes advantage of the central dogma of biology (DNA is transcribed to mRNA; mRNA is translated to proteins. The introduction of mutations in target genes has become trivial with the advances in synthetic biology. Major classes of library construction include (i) random mutagenesis, (ii) recombination, (iii) semi-rational design, and (iv) scanning mutagenesis. Together, these classes of library construction are termed "combinatorial design" and require little to no prior knowledge of the protein, albeit more knowledge can help design libraries more efficiently. The counter technique is termed "rational design" and requires detailed information about the protein for iterative computational mutagenesis. Both design processes have been instrumental in protein engineering, but this section focuses on the main classes of combinatorial

design. For a more detailed description of methods for library construction, Packer and Liu recently published an excellent review (Packer and Liu, 2015).

Random mutagenesis is a method where substitutions, deletions, or insertions are introduced at random in the target gene, mimicking random mutations introduced during natural evolution, but at an accelerated and tunable rate. Random mutagenesis is particularly appropriate for studying proteins with a lack of structural or functional information. This method can highlight particular "hot spots" in the protein where mutations are most likely to be advantageous to the desired fitness, which can then be targeted for future rounds of mutagenesis. Several techniques have been used to create randomly mutated libraries including, ultra violet light exposure, chemical mutagens, error-prone strains, and error-prone polymerase chain reaction (ep-PCR) (Cirino et al., 2003). The latter has been the dominant method over the past 25 years owing to its specificity, tunability, and accessibility (preassembled mutagenesis kits are available through several biotechnology companies). Both low and high error-rates for ep-PCR have been successful for engineering functional protein variants (Daugherty et al., 2000; Hamamatsu et al., 2006; Kunichika et al., 2002; Takase et al., 2003; Zaccolo and Gherardi, 1999), but the choice of error-rate is a balance between uniqueness and function and is dependent on the protein and the mutagenesis protocol (Drummond et al., 2005).

Another technique that can mimic natural evolution is recombination, where fragments of DNA are swapped between similar genes using homologous recombination. In general, library design through recombination techniques involves the creation of DNA fragments from a single gene or multiple related genes, pooling all the fragments, and piecing back together full length genes. Each full-length gene will be a mixture of the different parts and constitute a unique member in the library. The DNA fragments can be derived from orthologs, genes from the same family, or even genes of previously isolated mutants. Several sophisticated and unique techniques have been developed for recombination library design including DNA shuffling

(Stemmer, 1994a; Stemmer, 1994b), staggered extension process (StEP) (Zhao et al., 1998), random chimeragenesis on transient templates (RACHITT) (Coco et al., 2001), incremental truncation for the creation of hybrid enzymes (ITCHY), and SCRATCHY (a combination of DNA shuffling and ITCHY) (Lutz et al., 2001).

DNA shuffling was originally developed in the mid-1990s and is still popular today. This technique begins with a pool of parent gene fragments randomly digested with DNAse I (50-100 bp in length). The pool of fragments are then subjected PCR in the absence of primers (a short oligonucleotide used to start DNA replication), where homologous or near homologous regions anneal and are extended. The extended fragments lead to full length genes after several cycles of PCR. The full length products are then subjected to PCR once again but with primers to amplify the full length product for subsequent cloning. DNA shuffling and the other recombination techniques are inherently conservative methods of library design, but they are useful for determining deleterious mutations in isolated library members with a higher frequency of mutations from other designs.

Semi-rational design techniques incorporates both combinatorial and rational design. Based on knowledge about hot spots in the protein (e.g., binding pockets, catalytically active sites, dimerization domains), specific residues are selected for mutagenesis due to their importance in the hot spot. Residues are substituted randomly using techniques such as overlap-extension PCR (oe-PCR) and QuickChange mutagenesis (QC-PCR). The former is popular for introducing multiple mutations in the gene. Degenerate oligonucleotides are used as primers in PCR to amplify fragments of the gene. Each amplified fragment contains a region of homology with the appropriate adjacent gene fragment. Fragments are mixed, similarly to DNA shuffling, in the absence of primer and allowed to extend until the full length gene has been amplified. Though this technique has proven successful, it can be tedious with multiple fragments and introduce bias during each round of PCR amplification. Recently with the advances in DNA

synthesis, an entire gene can be synthesized with the proper degenerate sites. The preferred method in academia remains oe-PCR because of the cost of degenerate gene synthesis, but the prices of DNA synthesis is constantly declining and may lead to more attractive pricing when compared with the time and labor involved with in-house library construction.

Saturation mutagenesis is a popular technique for semi-rational design and uses mixed bases at each of the three positions in a codon to make all 20 possible amino acid substitutions. The mixed base code for a degenerate codon used in this technique is typically "NNS" or "NNK" (N = A, T, G, or C; S = G or C; K = G or T; see Table I-1 for a list of the mixed base code), which reduces the possible number of codons two-fold compared to "NNN" and still codes all 20 amino acids (NNN = 4 x 4 x 4 = 64 codons; NNS(K) = 4 x 4 x 2 = 32 codons). The number of sites mutated (n) exponentially increases the diversity of the library (32^n), where five degenerate codons would result in a library with 32^5 (~3.3 x 10^7) genetically unique mutants. The complexity of the hot spots and often unknown cooperative effects of residues require the user to saturate many residue positions, which leads to libraries with immense diversity. As an alternative to saturation mutagenesis, restricted saturation mutagenesis can be implemented to reduce the genetic diversity of a library by using a small subset of the 20 amino acids coded using restricted degenerate codons. The amino acid subsets can be grouped by a variety of different methods including but not limited to: size of the functional group; functional groups having similar chemical properties; functional groups having different chemical properties; residues found in naturally occurring homologous proteins; computationally determined residues. Here, various combinations of mixed bases are used to limit the possible amino acids at a particular position (Kille et al., 2013; Reetz et al., 2008).

Table I-1 Standard mixed base code for equal molar mixing of bases in degenerate oligonucleotides

Mixed Base Code	Base A	Base T	Base C	Base G
R	*			*
Y		*	*	
M	*		*	
K		*		*
S			*	*
W	*	*		
H	*	*	*	
B		*	*	*
V	*		*	*
D	*	*		*
N	*	*	*	*

Finally, scanning mutagenesis systematically determines hot spots by examination of residue substitutions at every residue in the protein. Here, a substitution can be the same amino acid (i.e. alanine scanning) or all/subset of amino acids, but unlike semi-rational techniques described above, only a limited number of residues are targeted per library member (typically 1-2 residues). Alanine scanning involves the single substitution of each residue with an alanine and interrogates loss-of-function owing to alanine's weak chemical interactions and relatively small functional group (Clackson and Wells, 1995; Cunningham and Wells, 1989; Weiss et al., 2000). Deviant alanine substitutions are typically targeted for further rounds of mutagenesis using a semi-rational mutagenesis approach.

The mutagenesis classes discussed above allow for some control over the diversity of the library members: random mutagenesis can be controlled by the PCR conditions; recombination is dependent on the number of starting proteins and the similarities of each protein; and semi-rational mutagenesis relies on the number of target residues set by the user. The amount of diversification in a library is dependent upon the screening method used, the extent in which the user would like to screen the library, and the quality of library construction. The latter can be

evaluated through next generation sequencing and statistical library evaluation parameters after the library has been constructed (Sullivan et al., 2013).

1.3.4) Screening for functional members of a library

The classic adage "you get what you screen for" has become known as the "First Law in Directed Evolution" (Schmidt-Dannert and Arnold, 1999), which implies function has a direct association with the quality and stringency of the screen. Here, functional members in a library are defined as library members exhibiting a minimal desirable trait (e.g., variant of a transcription factor that still regulates transcription). An ideal screening technique would link a genotype with an observable phenotype and allow for direct isolation of the genotype with *all* of the desired functions. This is not always easy because the fraction of functional members in a library possess a range of heterofunctional traits, which may interfere with a particular screen's ability to efficiently segregate the "best" members from the rest in the library. Therefore, determination of the correct screen is of great importance when considering combinatorial approaches.

Library screening is often compared to a funnel where it starts out broad and finishes narrow, but the physical act of screening can be better visualized as a sieve, where libraries are sifted for members with the desirable traits. Directed evolution is a powerful method but is typically bottlenecked by the throughput of available screening techniques (Aharoni et al., 2005b; Arnold, 1996; Fowler et al., 2008; Guntas et al., 2005; Schmidt-Dannert and Arnold, 1999). Often times, library screening techniques are defined as a "screen" or "selection", but these terms are sometimes used interchangeably. To avoid confusion in the subsequent discussion, here, a selection refers to a screening technique where a phenotype is directly linked to cell survival (i.e. qualitative "Yes" or "No" response) and a screen is a screening technique where the phenotype is a measurable output (e.g., quantifiable change in color, growth rate, metabolite concentration, fluorescence). Techniques for library screening i) range from elegant and broadly applicable to complex and specific for a particular experiment; ii) are developed for both *in vivo* and *in vitro*

applications; iii) classified as either low- or high-throughput, depending on the library members observed in a single round; and iv) take advantage of the many biomolecules afforded by nature (e.g. DNA, RNA, proteins). This constantly growing families of screening biotechniques is transforming our ability to evolve novel biosynthetic microbial factories. The general families of screening techniques are compared relative to time and throughput in Figure I-4. The throughput is defined as the number of variants that can be screened in a single experiment. The time refers to the amount of time for a single screening experiment and does not include numerous rounds of screening. The blue box indicates techniques considered to be high-throughput according to both time and throughput. The shape of the highlighted areas represent the variety of parameters and tools that can be used to increase the throughput of a particular family of techniques. For example, FACS can be screened with various nozzle sizes, which affect the sort rate (i.e. time). Highlighted below are examples of whole-cell screening techniques and their relevant applications to protein engineering.

1.3.5) Whole-cell screens

The genotype-phenotype linkage for whole-cell screens typically involves the creation or degradation of a molecule that results in a change in pH, color, fluorescence, or substrate/product concentration. The latter is detectable by chromatography and is considered to be the lowest throughput, but the remaining phenotypes are designed to be visible by the naked eye under proper illumination (e.g., color change or colony size) and are considered medium- to high-throughput depending on their application (Dietrich et al., 2010). Instruments and software are available to increase the throughput of these methods and provide statistical analysis of the screened library, such as the QPix400 Colony Autopicker from Molecular Devices (Molecular Devices, Cat. No. QPix 400 series) which can screen and pick up to 30,000 colonies per day based on fluorescence, color, zone of inhibition (antibiotic resistance), and colony deformation.

Below are examples of unique whole-cell screens developed to isolate variants from agar plates and microtiter plates.

Esterases are a diverse group of hydrolyzing enzymes responsible for the cleavage and formation of ester bonds in animals, plants, and microorganisms. They have been of particular interest for engineering owing to their naturally wide substrate tolerance and high regio- and stereospecificty. The Bornscheuer group has done extensive research on the discovery and engineering of esterases. They were successful in recombinantly expressing and characterizing

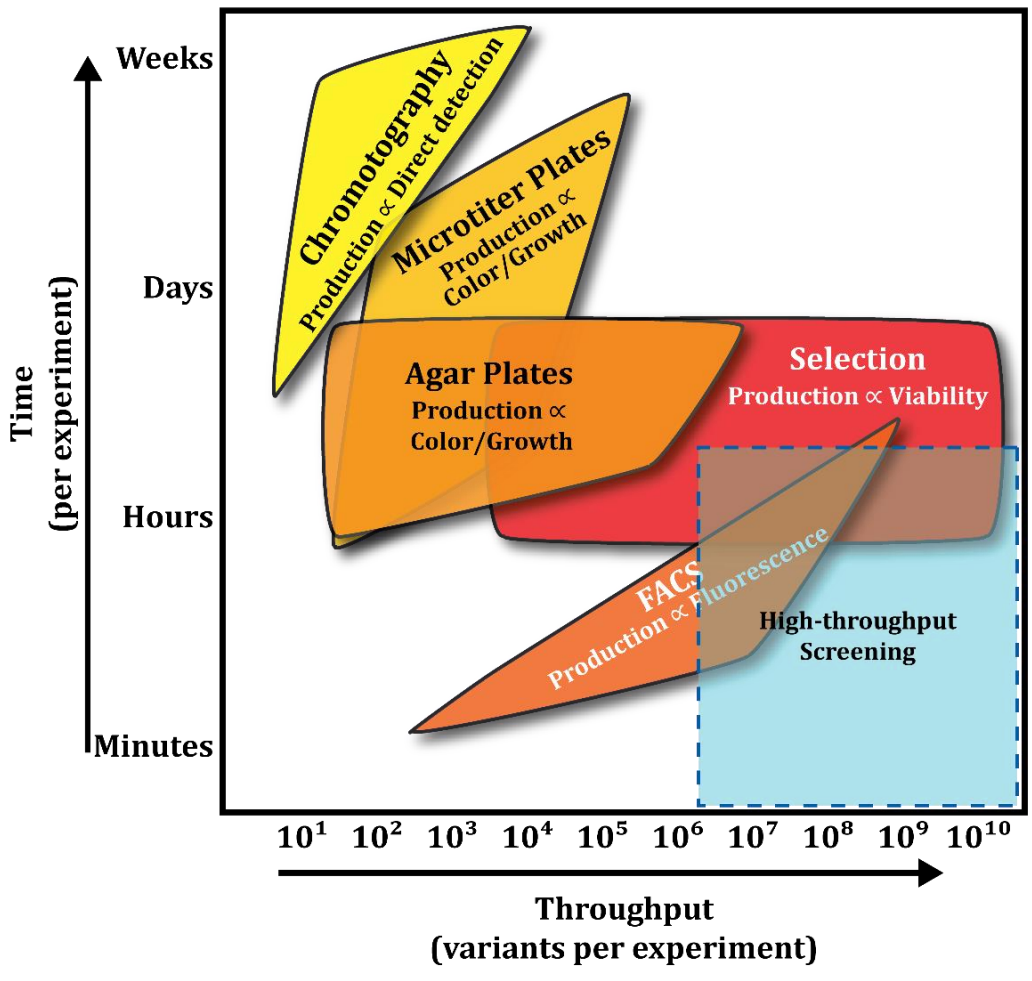


Figure I-4 Diagram depicting the general relationship of throughput and time as they relate to families of screening techniques and was adapted from Dietrich and coworkers (Dietrich et al., 2010).

an esterase from *Pseudomonas fluorescens* (PFE) (Krebsfanger et al., 1998a; Krebsfanger et al., 1998b). They used a unique pH-based agar plate assay to isolate a variant of PFE able to cleave a 3-hydroxy ester that was previously inaccessible to wild-type due to steric hindrance. A library of PFE was created after several growth iterations in the mutator strain *Epicurian coli* XL1-Red and subsequently expressed in *E. coli* DH5α. The clones were replicate plated on agar plates containing crystal violet and neutral red as pH indicators. The plates contained either an ethyl ester or a glycerol ester, each ester bonded to 5-(benzyloxy-)-3-hydroxy-4,4-dimethyl-pentanoic acid, which altered the microenvironment pH upon hydrolysis of the acid and turn the colony red (Bornscheuer et al., 1998). In addition to the color change, the clones that were successful in cleaving the glycerol ester were able to access glycerol as a carbon source and grow faster. Using this assay they were able to isolate a PFE variant with two point mutations (A209D, L181V), showing a moderate selectivity of approximately E=5, where E is the ratio of catalytic efficiency between the two esters.

A unique colorimetric-based assay was developed for detecting the activity of cytochrome P450 BM-3, a fatty acid hydroxylating enzyme, and its mutant F87A. The basic function of a cytochrome P450 enzyme is that of a monooxygenase, which catalyzes the reaction of molecular oxygen with a free C-H bond, making them particularly attractive for pharmaceutical synthesis and pollution management. Schwaneberg and coworkers reported a colorimetric assay based on the NADPH-dependent hydroxylation of p-nitrophenoxycarboxylic acid (pNCA) to a free fatty acid and the chromophore p-nitrophenolate (yellow) (Schwaneberg et al., 1999). In order to incorporate this reaction into a high-throughput screen, suitable changes were made for whole-cell screening in 96-well plates (~3000 variants/day could be screened) (Schwaneberg et al., 2001). The whole-cell assay was optimized by permeabilizing the cells to allow for pNCA diffusion through the cell membrane into the cytoplasm of *E. coli* and determining the non-limiting concentrations of $NADP^+$ and D/L-isocitrate (artificial supply of NADPH). A modified

approach was later employed to screen variants of P450 BM3. The authors used a unique alternative cofactor based on zinc for the electron source and cobalt(III)sepulchrate as the mediator (Schwaneberg et al., 2000), replacing the need for $NADP^+$ and isocitrate. The saturated several residues near the active site and found two double mutants (F87A, R47Y; F87A, R47F) with higher catalytic activity towards 12-pNCA over wild-type and a previously characterized single mutant (F87A), but were "addicted" to the Zn/Co^{III}sep cofactor (Nazor and Schwaneberg, 2006). The single mutant out performed both mutants (4- to 8-fold) in the absence of the alternative cofactor. This strongly supports the aforementioned adage, "you get what you screen for." As our search continues for interesting and more complex biosystems for engineering, the throughput of the screen becomes more crucial.

1.3.6) Whole-cell selections

The throughput of *in vivo* selections dramatically extends beyond the throughput of the screens mentioned above, where the library coverage is no longer dependent upon the screening technique but by the total number of transformants. Here, a mutant library is expressed in a host organism with critical growth requirements relating to the desired enzymatic activity. The clones are then subjected to a specific selective pressure, only permitting the growth of clones with the targeted trait. Selections are exceptionally high-throughput, making them ideal for screening large libraries, but they are often difficult to implement into directed evolution methods. Most industrially interesting enzymes catalyze secondary metabolites that are not essential for cell growth, so strains and optimized growth media must be developed for selections. However, selections have been successfully engineered using either auxotrophic or chemical complementation. The former and most common involves a host organism that is directly growth-dependent on the molecules involved in the targeted enzymatic reaction (e.g., reaction leads to molecules such as essential amino acids, fatty acids, and nucleobases), whereas the latter is indirectly growth-dependent and relies on an additional mechanism to clear an exogenous toxic

molecule. Selections through chemical complementation are best mediated through an engineered whole-cell biosensor, which will be discussed in section (I.4) of this chapter on molecular biosensors. Below are several examples that highlight the use of selections through auxotrophic complementation.

Combinatorial design and subsequent screening is most commonly performed to improve or alter a protein's attributes, but we can also apply these methods to better understand metabolic pathways and the roles of specific pathway constituents. Bi and coworkers recently used a fatty acid auxotroph to characterize an under expressed enoyl-acyl carrier protein reductase, which catalyzes the last step of the bacterial fatty acid elongation pathway essential to membrane fatty acid synthesis (Bi et al., 2014). *Entercoccus faecalis* expresses two enoyl-acyl carrier protein reductases, FabI and FabK. A Δ*fabI* strain has severely limited growth, albeit growth can be restored in the presence of a supplemented fatty acid such as oleic acid. Spontaneous mutations in the upstream region of the *fabK* gene of the Δ*fabI* strain were discovered from colonies able to grow in the absence of oleic acid and were shown to increase the translation rate of FabK. The *fabK* gene and the wild-type promotor region were subjected to random mutagenesis and screened on agar plates in the absence of oleic acid, where only clones resulting in positive mutations should grow. Each of 13 clones isolated after selection contained one of the same mutations seen in the spontaneously mutated clones and 11 of the 13 clones contained addition mutations, either nonsynonymous or synonymous, throughout the *fabK* gene. These results helped in concluding the fatty acid synthesis in a Δ*fabI* strain of *E. faecalis* was growth inhibited not because of the catalytic efficiency of FabK, but by the low translation rate of the enzyme. This highlights the need to explore alternative rate limiting parameters besides the turnover rate, binding affinity, and stability of an enzyme.

A unique selection was designed by Boersma and coworkers to improve enantioselectivity of a lipase for hydrolysis of a butyrate ester (Boersma et al., 2008). A mutant library of lipase A

from *Bacillus subtilis* was expressed in an aspartate auxotroph *E. coli* strain and plated on media simultaneously supplemented with aspartate ester of S-(+)-1,2-isopropylideneglycerol (S-(+)-IPG; desired substrate) and a phosphonate ester of R-(-)-IPG. The latter is a phosphonate inhibitor of S-(+)-IPG and thus cell growth is inhibited when the lipase cleaves the phosphonate ester of R-(-)-IPG. Enatioselectivity of lipase A was inverted and improved for hydrolysis of butyrate esters for the production of enantiopure S-(+)-IPG through three round of selection with increasing stringency. A mutant was isolated with increased enantioselectivity from an *ee* (enantiomeric excess) of -29.6% in favor of R-(-)-IPG to an *ee* of +73.1% in favor of S-(+)-IPG. More recently, Matsui and coworkers engineered a NAD^+-dependent L-tryptophan dehydrogenase (TrpDH), which catalyzes the reversible oxidative deamination of L-tryptophan and the reductive amination of indole-3-pyruvic acid (IPA), to have increased stability and catalytic activity over the wild-type. *trpDH* of *Nostoc punctiforme* was subjected to random mutagenesis and expressed in a L-tryptophan auxotroph *E. coli* strain (Δ*trpB*) (Matsui et al., 2015). The resulting clones were plated on selective media absent L-tryptophan and supplemented with IPA. After a single round of selection, a variant was isolated with four nonsynonymous mutations (TrpDH-4mut, L59F/D168G/A234D/I296N), resulting in approximately a 4.6-fold improvement in stability and nearly a 5-fold increase in catalytic activity.

1.3.7) Fluorescence-activated cell sorting

With the increased availability of flow cytometers over the past decade, fluorescence-activated cell sorting (FACS) has emerged as a powerful and highly quantitative high-throughput screening technique for the directed evolution of proteins. FACS offers strict control over the isolation of population subsets within a library based on multiple parameters of the cell, including fluorescence and size (Link et al., 2007). Flow cytometers are constructed from a light source, a cell delivery mechanism, a detection apparatus, and analysis software (Shapiro, 2003). As shown

in Figure I-5, cells are introduced into a sheath fluid (typically phosphate buffered saline) and hydrodynamically focused by inducing a laminar flow regime with pressurized sheath (typically 25-100 psi) that is forced through a small orifice (typically 50-150 μm). The stream exits the orifice with a diameter (d) of 50-150 μm and often is introduced into the air (called "stream in air"). Cells in the stream are then individually passed through one or multiple incident light beams, typically from lasers of specific wavelengths. As the cell passes through the beam, light is either deflected or induces fluorescence from a fluorophore present in the particular event being measured (an event describes a detectable scatter of light above the background noise). The deflected light is categorized as forward scatter (FSC) and side scatter (SSC), which are generally used to characterize the size and granularity of a cell, respectively. After the cell has reached the incident light source and a fluorophore has been excited, the emitted photon passes through an array of filters in order to determine the specific emitted wavelength. The scattered light and the light emitted from the fluorophore are detected using photomultiplier tubes, which amplify the signal based on the voltage setting, adjustable by the user. The signal is then interpreted through software and typically displayed in either a scatter plot of two parameters or a histogram of a single parameter. If the flow cytometer is equipped for sorting, cells can be sorted based on user-defined gates set in a scatter plot or histogram of the measurable parameters. Droplets are induced in the stream using a piezoelectric transducer with a tunable frequency so the stream becomes predictably unstable (drop formation is stable) and a single cell will be present in a drop. A droplet containing a cell of interest is charged and passed through an electric field, deflecting the droplet into a collection tube. The desirable trait of a cell is usually linked to the presence of a particular fluorescent marker such as an autofluorescent protein (e.g., green fluorescent protein, commonly referred to as GFP) (Shaner et al., 2005) or polypeptide tag that recognizes and binds a small molecule that fluoresces. Fluorescent reporters have proved invaluable across numerous areas of research in biology and engineering (Shaner et al., 2007). They are robust and highly

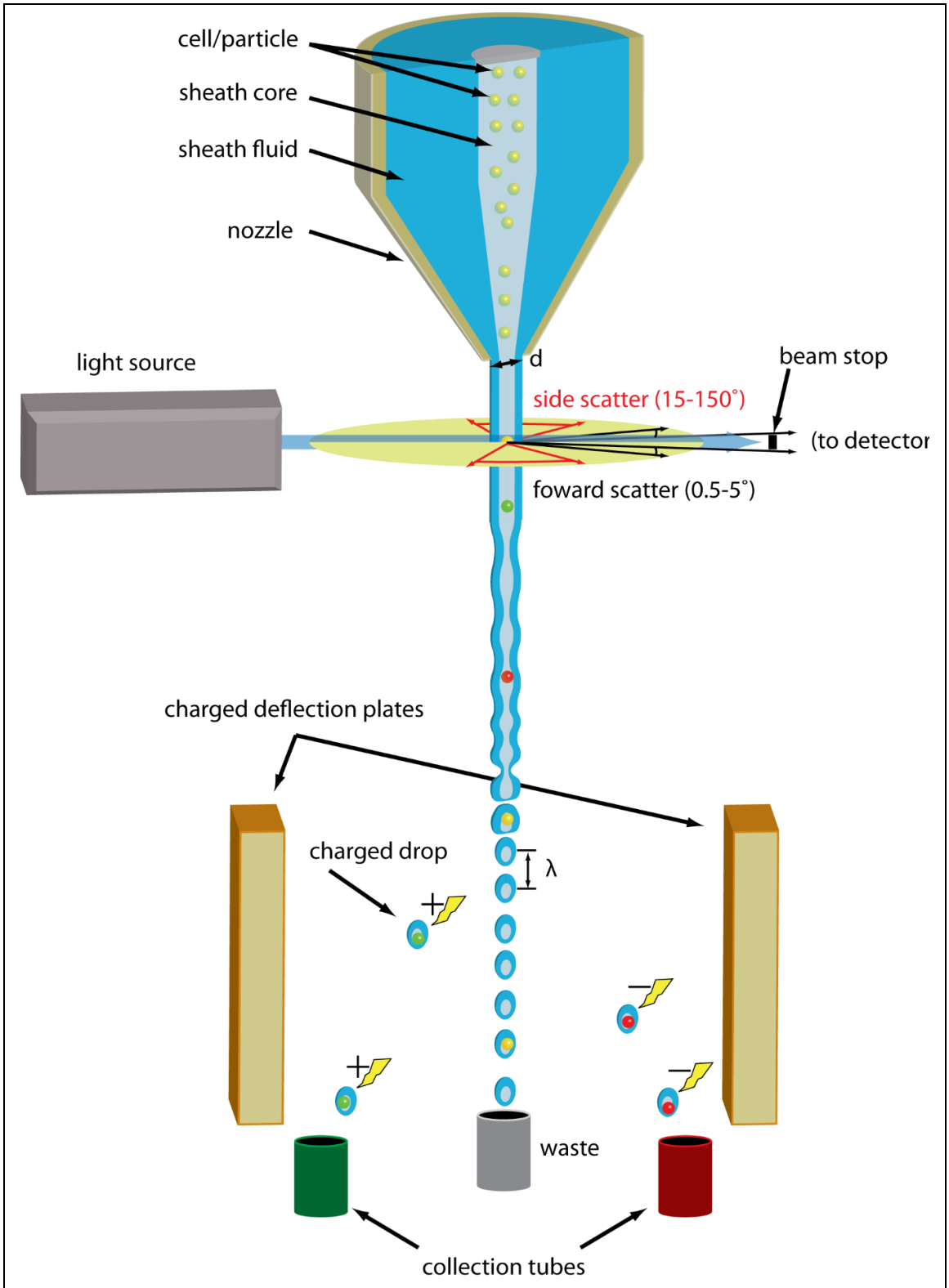


Figure I-5 Schematic of fluorescence-activated cell sorting (FACS).

sensitive in most microorganisms with a range of adjustable parameters including stability (Andersen et al., 1998), excitation/emission wavelengths (Telford et al., 2012), and techniques (e.g., fluorescence resonance energy transfer (FRET) (Jares-Erijman and Jovin, 2003) and fluorescence lifetime image microscopy (FLIM) (Sun et al., 2011)).

The aforementioned FSC and SSC are determined by the amount of light scattered from the incident beam by 0.5-5° and 15-150°, respectively. In general, FSC is used to measure the size of a particle (e.g., a cell or bead) and SSC is used to measure size and granularity of the cell (e.g., organelles and cell surface). FSC can be useful for determining different populations of cells in a heterologous mixture, but can be affected by the other factors, such as the refractive index and different optical designs by different manufacturers. Therefore, cell size should be held with certain skepticism when determined from FCS. In simple and homogeneous samples, FSC provides a robust parameter for defining a trigger to identify an event, but small cells (< 2 μm) can be difficult to distinguish from the background noise on less sensitive instruments. In this case, a fluorescent stain may be a better option for the triggering the detection of an event.

The goal of FACS is to sort as many cells as possible in the shortest amount of time without compromising the quality of the sort. The rate of sorting determines the feasibility of the method, which is dependent upon various instrument parameters such as the pressure of the sheath fluid, the event rate, and the frequency of drop formations. As was mentioned earlier, laminar flow is induced by flow through a small orifice and is crucial for maintaining a stationary and steady flow. The laminar flow regime is related to the fluid velocity and the orifice diameter through the dimensionless variable known as Reynold's number (Re), defined as

$$Re = \frac{d\,\rho\,v_{avg}}{\eta}, \qquad \textbf{I-5}$$

where d is the diameter of the stream (usually assumed to be the orifice diameter in μm), ρ is the density of the fluid (g cm^{-3}), v_{avg} is the average fluid velocity (m s^{-1}), and η is the viscosity (g cm^{-1}

s^{-1}). Pinkel and Stovel found that the fluid velocity is directly proportional to the square root of the sheath pressure (ΔP) and can be estimated with the following relationship (Van Dilla, 1985),

$$v_{avg} = 3.7\sqrt{\Delta P}\,. \qquad \textbf{I-6}$$

Flow is considered laminar when Re < 2,300, so in order to maintain laminar flow and increase the flow rate, the pressure can be increased if the diameter of the orifice is decreased. The higher the flow rate, the greater the number of events that can be observed and thus directly affecting the sort rate.

For example assuming an ambient temperature of 20°C, where water has a density of 1 g cm^{-3} and a viscosity of 0.01 g cm^{-1} s^{-1}, two nozzles with a 100 μm and 70 μm diameter could each have maximum fluid velocities of 23 and 33 m s^{-1}, respectively. These results can be translated into drop frequency using the equation

$$v_{avg} = f\lambda\,, \qquad \textbf{I-7}$$

where f is the frequency of drops (kHz) and λ is the wavelength (distance between drops in μm). The wavelength is dependent on the instrument but is typically observed between 4-8 times the diameter of the stream. So for the same nozzles mentioned above being used on the same instrument, the maximum frequencies would be 38.3 and 78.2 kHz, respectively (λ = 6d). Now as a general guideline while sorting, the event rate should be maintained so that there is approximately one event per five drops as a means to maintain a minimum level of sort purity and decrease the computational load on the instrument. Thus, event rates for a 100 μm and 70 μm nozzle should be maintained at ~7,500 and ~15,500 events s^{-1}, respectively. Finally, if the user needs to look at 10^8 cells in order to cover their library with reasonable confidence, the instrument fitted with a 100 μm nozzle will take almost twice as long as the instrument fitted with

the 70 μm nozzle (3.6 and 1.8 hrs, respectively). This example was presented in order to give the reader a general understanding of the instrument parameters and how they affect library screening. For a more in depth discussion of flow cytometry, an excellent text was written by Howard Shapiro called Practical Flow Cytometry (Shapiro, 2003). Below are several examples of how FACS has been used in library screening.

Perhaps the most basic example of FACS-based screening is when an endogenous enzymatic reaction results in a fluorophore. Many microtiter plate assays have been developed based on the release of a fluorophore post-enzymatic cleavage, but if the resulting fluorophore can remain internalized by the cell, FACS can dramatically enhance the throughput of the assay. For example, a FACS-based screen was developed for the directed evolution of a 2′-deoxynucleoside kinase (dNK), which phosphorylates nucleoside analogs upon penetration of the cellular membrane (Liu et al., 2009). Once phosphorylated, the nucleoside analog remains internalized. By covalently attaching a fluorescent moiety to the target nucleoside analog, the clone with a functional dNK will retain a higher concentration of the phosphorylated fluorescent nucleoside analog. Liu and workers subjected the *Drosophila melanogaster* dNK (*Dm*dNK) gene to random mutagenesis and DNA shuffling and expressed the resulting libraries in *E. coli*. The libraries were screened via FACS for cells exhibiting a high level of fluorescence, indicating the phosphorylation of a fluorescent analog of 3′-deoxythymidine (fddT). The selection pressure was increased over each subsequent round of sorting by lowering the concentration of fddT, decreasing the incubation time with fddT, and incubating in the presence of a high concentration of thymidine (a highly favorable substrate for wild-type *Dm*dNK and an undesirable substrate for the engineered variant). After four rounds of FACS, they isolated several variants for kinetic analysis and found variants with an overall 10,000-fold change in substrate specificity in favor of 3′-deoxythymidine over thymidine. In addition to enzymatic improvements, other protein parameters can be targeted for FACS-based screening, such as stability. Seitz and coworkers

targeted libraries of randomly mutated human the glucocorticoid receptor ligand binding domain (hGR-LBD) for improved stability and solubility in *E. coli* (Seitz et al., 2010). The hGR-LBD was fused with a green fluorescent protein (GFP), where a fluorescent signal would confer proper folding of the fusion protein. After eight rounds of sorting the top fluorescent clones via FACS, four beneficial mutations were identified and when combined increased the thermal stability of the protein by 8°C. Upon introduction of these mutations in to a mouse glucocorticoid receptor ligand binding domain, enough protein was purified to resolve a crystal structure of this protein for the first time.

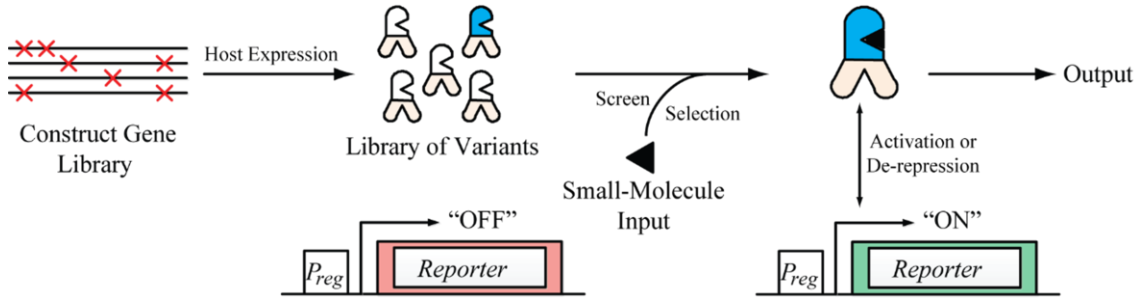
If the fluorescent product from an enzymatic reaction does not remain internalized, the cells can be compartmentalized in microenvironments using water and oil emulsions. The subsequent emulsions, designed to statistically contain a single cell in each emulsion, retain the permeated fluorescent product and can be sorted using FACS. This method was used to screen a library of serum paraoxonase (PON1) for increased catalytic activity towards the hydrolysis of thiobutyrolactones, where both the substrate and product can permeate the cellular membrane (Aharoni et al., 2005a). Both random mutagenesis and a variation of site-saturated mutagenesis (16 residues were targeted randomly using an optimized mixture of degenerate oligonucleotides to generate an average of three mutations in a single variant) were used to generate libraries of PON1, which were subsequently expressed in *E. coli*. The clones were emulsified in oil and supplemented with the substrate (γ-thiobutyrolactone) and a thiol-detecting reagent N-(4-(7-diethylamino-4-methylcoumarin-3-yl)phenyl)maleimide (CPM). Upon hydrolysis of the thiobutyrolactone substrate, the free thiol group of the γ-thiobutyric acid reacts with CPM, creating a fluorescent thiol adduct. The water-in-oil emulsions were passed through a second round of emulsification to form water-in-oil-in-water emulsions, which creates a stable emulsion for FACS screening. Approximately 5×10^7 cells were screened from each of the libraries and were either sorted in a single round with high stringency (top 0.2-0.01% most fluorescent) or

three iterative rounds of FACS with lower stringency. The former screening scheme reduces time spent screening but relies heavily on the sorter to perform efficiently, whereas the latter scheme decreases dependence on the sorter efficiency but increases screening time. Although clones were isolated from a single stringent sort with improved activity, the best clones were isolated from the multiple rounds of sorting, showing ~100-fold better catalytic efficiency for hydrolysis of γ-thiobutyrolactone than the wild-type PON1.

The above examples illustrate the power of FACS for screening large libraries of enzymes and other proteins for improved properties, but it is limited by the lack of readily available genotype-phenotype linkages. Often times, an engineer has to be creative to design elegant and elaborate screening methods that may only work for a handful of reactions in order to use high-throughput methods such as selection and FACS. Recently though, a natural and powerful alternative has been investigated to circumvent this issue and will be discussed in the next section.

I.4) Transcriptional regulatory protein based whole-cell molecular biosensors

As was discussed above, directed evolution is an invaluable method for biological design, but is inhibited by the lack of sensitive, specific, and high-throughput screening techniques. A great amount of effort has been focused on addressing these bottlenecks, but one of the more recent and promising areas of research has been in whole-cell molecular biosensors designed from allosteric transcriptional regulatory proteins (TRPs). These proteins encompass a large and diverse class of proteins ubiquitous throughout nature and are ideal candidates for biosensor design owing to their natural tunable expression, interchangeable domains, and specificity and sensitivity to a wide range of molecules. As depicted in Figure I-6, TRPs recognize and bind a molecule and subsequently undergo a shift towards an alternative conformation, resulting in either a negative or positive change in the transcription rate. Transcriptional regulation by an



Type of Output	Example
Fluorescence	GFP, FRET
Antibiotic Resistance	Chloramphenicol
Colorimetric Reaction	β-Galactosidase
Antibiotic Sensitivity	β-lactamase Inhibitory Protein
Sucrose Sensitivity	SacB
Lack of Fluorescence	No GFP Expression

Figure I-6 General strategy for isolation of biosensors from regulatory proteins. Libraries of the protein are cloned using either random ore site-specific mutations. The library is expressed in an amenable host and screened for desired functionality in the presence or absence of a target effector. Several commonly used outputs are listed in the table to the right.

TRP in the presence of an effector molecule (i.e. ligand) typically follows a sigmoidal response that is dependent on the concentration of the molecule, the fully folded TRP, the promoter, and any inhibitor molecules (Changeux and Edelstein, 2005). TRPs can be classified into three general groups: i) activators, ii) repressors, and iii) dual regulators. Activators bind a ligand, bind to the appropriate operator, and activate transcription from a promoter that otherwise has a little to no expression. Repressors actively block transcription in the absence of a ligand by binding to an operator in the promoter, preventing RNA polymerase from starting transcription. Upon recognition of a ligand, the repressor-ligand complex disassociates from the operator, allowing for transcription to proceed. Finally, a dual regulator will actively repress transcription in the absence of a ligand and will remain bound within the promoter region upon recognition of a ligand, but will activate transcription. Figure I-6 summarizes the general design for engineering a TRP-based biosensor. Their design and biosensor-type applications have recently been reviewed by several groups (Dietrich et al., 2010; Gredell et al., 2012; Li et al., 2011; Michener et al., 2012; Zhang and Keasling, 2011). The subsequent sections describe specific examples of TRPs and their applications as whole-cell biosensors.

1.4.1) Transcriptional regulatory proteins-Activators

Activators play a crucial role in all organisms' gene regulation. The natural design of an activator promotes transcription of a downstream gene upon recognition of the appropriate effector molecule and undergoing an allosteric shift allowing operator binding ("ON"). Without the activator bound, the downstream gene is not transcribed ("OFF").

As a first example of transcriptional activator systems engineered for genetically reporting *in vivo* ligand detection, the de Lorenzo group has developed sensors for toxic or explosive compounds based on transcriptional regulators from *Pseudomonas putida*. Such sensors would be particularly useful in locating the plethora of undetonated landmines plaguing former war-torn countries. This work has recently been reviewed (de las Heras et al., 2010), so here we highlight

the most relevant features. Initially, the researchers used the activator protein NahR, which naturally controls the expression of genes involved in naphthalene degradation through activation of the P_{nah} and P_{sal} promoters (Schell and Poser, 1989). NahR is activated by salicylic acid and several of its structural analogs, yet other similar compounds, such as benzoic acid (lacking a single hydroxyl group) generate no response (Cebolla et al., 1997). To expand the specificity of NahR to include these other molecules, a protein library carrying random point mutations generated by error-prone PCR was screened using *E. coli* strain SAL1 constructed to contain a chromosomal fusion of the native P_{sal} promoter to *lacZ* (encoding β-galactosidase). Colonies exhibiting increased β-galactosidase activity in the presence of ligand as determined by blue/white screening were characterized and two mutants were isolated with increased sensitivity to benzoic acid (from ~0.1-10 mM) (Yanischperron et al., 1985). The mutants also showed increased affinity for 3-chlorobenzoic acid and 3-methyl salicylic acid and maintained responsiveness to salicylic acid.

Witholt and coworkers later developed the NahR system into a simple yet potentially versatile platform for identifying active biocatalysts (Fiet et al., 2006). Their focus was on detecting benzoic acid and salicylic acid, important precursors for the synthesis of various organic compounds. Using *E. coli* as the reporting host, the authors placed the *tetA* gene (conferring resistance to the antibiotic tetracycline, tc) downstream of the P_{sal} promoter and expressed the benzoate-responsive NahR variant described above (Cebolla et al., 1997). Production of benzoate from benzaldehyde by the benzaldehyde dehydrogenase XylC of *P. putida* thus conferred resistance to tc. Therefore, XylC variants having increased catalytic activity could be selected from this reporter strain based on colony size. The success of the NahR platform to identify mutants that recognize new molecules, albeit with relaxed specificity, motivated the de Lorenzo group to use a modified mutagenesis and reporting system for further biosensor development using a platform more suitable for evolving recognition of explosive products in soil. The

homologous transcriptional activators XylR, DmpR, and TbuT, which respond to different sets of aromatic compounds such as m-xylene (XylR), phenol (DmpR), and benzene (TbuT) (Garmendia et al., 2001; Skarfstad et al., 2000), were the basis for this new biosensor. A combinatorial gene library was created by mutation-prone DNA shuffling (Stemmer, 1994a; Stemmer, 1994b) of the genes encoding the N-terminal effector binding domains of the three proteins, and fused to the remaining C-terminal XylR framework. The library was inserted into a transposon delivery vector and conjugated with a specialized *P. putida* strain that allowed positive and negative selection of XylR-regulated Po promoter activity. Effector recognition was first positive selected (active Po promoter expressing the kanamycin resistance gene enabled growth in the presence of kanamycin) on plates in the presence of 2-nitrotoluene, 3-nitrotoluene, 4-nitrotoluene, or biphenyl (2 mM), none of which are natural effectors of the parent proteins, and then negative selected in the absence of effector to eliminate constitutively active clones (active Po promoter expressing *sacB* encoding levansucrase, which inhibits cell growth in the presence of sucrose). Selected clones were re-screened for effector-dependent expression of luciferase (*luxAB*) from the *Po* promoter, and five unique variants were further characterized. Each variant exhibited a broadened effector profile, including responsiveness to DmpR- and TbuT-specific ligands (Garmendia et al., 2001). Sequence analysis revealed that very different mutation patterns among the variants can lead to similar phenotypes and suggested that mutations affecting both the regulatory switching mechanism as well as the binding pocket-ligand interactions can lead to an expanded inducer profile.

Continuing with XylR as a platform for engineering novel ligand specificity, the same group developed a reporter that would respond to previously unrecognized 2,4-dinitrotoluene (DNT), a xenobiotic compound used primarily in the polymer industry (Wegener et al., 2001), but more generally known as a precursor to the explosive trinitrotoluene (TNT). In this case a library of XylR random mutants was generated (with mutations occurring in both the effector binding

domain and the "B connector" to the DNA binding domain, DBD) and subjected to selection and counter-selection based on expression of *pyrF* encoding oritidine-5'-phosphate decarboxylase (controlled by the alternate XylR promoter Pu) in a uracil auxotroph (Δ*pyrF*) *P. putida* strain (analogous to the yeast URA3-based selection) (Galvao et al., 2007). Plate-based selections in the presence and absence of DNT (2 mM) revealed five clones exhibiting the desired DNT-dependent activation of gene expression.

Expression of *lacZ* from promoter Pu in *P. putida* was then used to characterize the whole-cell response of these mutants to various compounds. As a frame of reference, wild-type XylR increased LacZ activity ~24-fold in the presence of the native effector 3-methylbenzylalcohol (3-MBA;1 mM), while LacZ activity increased no more than ~30% in the presence of up to 2 mM DNT (with wild-type XylR). In contrast, several of the selected XylR mutants increased LacZ activity by 2-3-fold in 125-2,000 μM DNT. The variants were not necessarily specific to DNT as they also activated expression up to ~10-fold in the presence of other related compounds at concentrations less than 2 mM (e.g., 3-nitrotoluene, 1,2,4-trichlorobenzene, and 2,4-dichlorophenol). Interestingly, none of the altered amino acids in these variants appeared in the effector binding pocket, and instead are believed to relate to changes in the conformational change following effector recognition. To further improve the response to DNT and reduce effector promiscuity, additional evolution of one mutant ("XylRv17"; XylR-F48I-L222R) was later performed by introducing random mutations via error-prone PCR and positive selecting in the presence of 1 mM DNT (reduced from 2 mM) and then performing two rounds of counter-selection (sensitivity to 5-fluoroorotic acid (FOA) if *pyrF* is expressed from promoter Pu) in the absence of DNT (to eliminate constitutive mutants) and the presence of m-xylene (to eliminate mutants with broad specificity). One variant contained two additional amino acid substitutions (I136T and S174R) that nearly doubled the expression response to 1 mM DNT (de las Heras and de Lorenzo, 2011a). Finally, the reporter strain was further engineered in a number of ways that

included chromosomal integration, elimination of antibiotic markers, and use of luciferase as a reporter. As a result, the more environmentally-friendly strain can emit light in the presence of DNT within a soil sample (de las Heras et al., 2008). Additional work is being undertaken that aims to preserve these bacteria in a deliverable form to maximize their effectiveness as environmental biosensors (de las Heras and de Lorenzo, 2011b).

In a similar effort, Beggah et al. sought whole-cell biosensors for 2-chlorobiphenyl (2-CBP) by engineering the ligand binding domain (LBD) of regulatory protein HbpR (similar to the XylR transcriptional activator), which natively responds to 2-hydroxybiphenyl (2-HBP) but not 2-CBP (Beggah et al., 2008). In this case, rather than using colony-based genetic selections to enrich mutant libraries, the authors used fluorescence activated cell sorting (FACS) to isolate *E. coli* clones harboring HbpR variants that activate *gfp* expression from HbpR-dependent promoter *Pc* in the presence of inducer, and show low fluorescence without inducer. One round of random mutagenesis and screening led to the isolation of variant HbpR-101V-D128N showing low constitutive expression, approximately linear dose-dependent *gfp* expression between 2 and 100 µM 2-CBP, and ~9-fold increased cellular fluorescence at 100 µM 2-CBP. The same variant showed increased sensitivity to 2-HBP as well as responsiveness to 2-bromobiphenyl and the disinfectant Triclosan (10 µM). The authors discuss possible applications of the *E. coli*-based biosensor, as well as use of the regulatory mutant for activating polychlorinated biphenyl degradation in a more appropriate organism (Beggah et al., 2008).

1.4.2) Transcriptional regulatory proteins-Repressors

Repressors constitute a class of regulatory proteins which upon operator recognition, translation of the downstream gene is repressed ("OFF"). Once the corresponding ligand binds with the repressor, the repressor undergoes an allosteric change, thus causing the repressor to disassociate from the operator and subsequently the downstream gene is transcribed ("ON").

The tetracycline-induced Tet Repressor (TetR) and variants thereof have been the subject of numerous studies aimed at understanding and engineering TetR regulation, as reviewed elsewhere (Bertram and Hillen 2008). In tc-producing gram-positive bacteria such as Streptomyces, the presence of tc causes TetR to release the tetO operator, thereby relieving transcriptional repression of tetA encoding an antiporter that actively pumps tc out of the cell and confers tc resistance. Owing to tight repression, high-inducibility of TetR to a variety of tc analogs (e.g., doxycycline (dox) and anhydrotetracycline (atc)), and a reasonable understanding of the relationships between TetR sequence, structure, and function, there exists a solid foundation for applying mutagenesis strategies to alter TetR inducer specificity.

Scholz and coworkers generated a TetR mutant capable of recognizing the non-antibiotic tc analog 4-de(dimethylamino)-6-deoxy-6-demethyl-tetracycline (cmt3) by screening mutant libraries for cmt3-dependent de-repression using β-galactosidase expression as a reporter (Scholz et al. 2003). First, the gene region corresponding to the C-terminus of a previously constructed TetR(BD) chimera (Schnappinger et al. 1998) was subjected to mutagenic DNA shuffling (Stemmer 1994a; Stemmer 1994b). The resulting library was transformed into E. coli and the eight best mutants induced by 0.4 μM of cmt3 were recovered from the β-galactosidase assay. In vitro recombination of these eight mutants, followed by further screening, revealed that mutation H64L confers a change in specificity from tc to cmt3. Additional randomization at position 64 yielded TetR-H64K, showing 9.8-fold higher β-galactosidase expression in 0.4 μM cmt3 compared to 0.4 μM tc or no effector.

Random mutations were next introduced in the tc-binding pocket and surrounding residues (identified from the crystal structure of TetR with tc (Hinrichs et al. 1994; Kisker et al. 1995; Orth et al. 1998; Orth et al. 2000)) yielding several more variants with interesting inducer profiles. Finally, the best variant identified in the first three rounds (TetR-H64K-S135L) was subjected to an additional round of random mutagenesis, yielding three variants fully inducible by

cmt3 and not tc, and ultimately achieving a more than 20,000-fold increase in specificity over tc and 20-fold increase in affinity (Scholz et al. 2003). Variant TetR-H64K-S135L also exhibited relaxed inducer specificity that included dox and atc. To improve the specificity towards tc derivative 4-de(dimethylamino)anhydrotetracycline (4-ddma-atc), while reducing the affinity for atc and dox, two additional residues near the tc binding pocket were randomized and the saturation library was screened for high β-galactosidase activity in the presence of 0.4 μM 4-ddma-atc and low activity with no inducer and with 0.4 μM atc (Henssler et al. 2004). A single variant (TetR-H64K-S135L-S138I) possessed the desired specificity for 4-ddma-atc. Using the same random mutagenesis and screening method, it was also found that single amino acid substitutions were sufficient to reverse the regulatory activity of TetR such that effector molecules are instead required for repression (i.e., binding the tetO operator) rather than de-repression (Kamionka et al. 2004a; Scholz et al. 2004). These TetR systems were adapted to provide simultaneous yet independent control over expression of two different genes in E. coli (Kamionka et al. 2004b; Krueger et al. 2007). Taken together, these results illustrate the range of molecular inputs that can be detected and processed into various cellular outputs, with only a few mutations in a relatively simple genetic reporting scheme.

1.4.3) AraC-a platform transcriptional regulatory protein for biosensor design

The *Escherichia coli* native transcriptional regulator AraC exists as a homodimer within the cytoplasm (Wilcox and Meuris, 1976); each monomer is 292 amino acids and consists of two distinct domains linked by a seven amino acid linker. The crystal structure of AraC was solved by Schleif and coworker, revealing four distinct domains: an N-terminal arm (residues 1-18), a ligand binding/dimerizaiton domain (LBD) (residues 19-167), an interdomain linker (residues 168-175), and a C-terminal DNA binding domain (DBD) (residues 176-292) (Bustos and Schleif, 1993; Eustance et al., 1994; Soisson et al., 1997). The resolved crystal structure revealed an

eight-stranded antiparallel β-barrel with a jelly roll topography and two antiparallel helices making up the ligand binding pocket and the dimerization domain, respectively, within the LBD. The N-terminal arm (residues 7 to 18) encloses one molecule of L-ara within the ligand binding pocket and forms both direct and indirect contact with the L-ara molecule when bound. Specifically, the proline at residue position 8 (P8) stabilizes the arm and makes a direct contact with the hydroxyl group bonded to the anomeric carbon of L-ara (Soisson et al., 1997).

AraC controls the expression of genes responsible for the catabolism (*araBAD*) and transport (*araFGH* and *araE*) of L-arabinose (L-ara), as well as self regulates *araC* (Schleif, 2010). *E. coli* have the metabolic capability to catabolize arabinose as a carbon source through proteins expressed from the *ara* operon. The genes included in the *ara* operon encode for AraC (arabinose DNA-binding transcriptional dual regulator), AraB (L-ribulokinase), AraA (L-arabinose isomerase), AraD (L-ribulose 5-phosphate 4-epimerase), AraE (arabinose/proton symporter), AraF (arabinose ATP-binding cassette (ABC) transporter-periplasmic binding protein), AraG (arabinose ABC transporter-ATP-binding subunit), and AraH (arabinose ABC transporter-membrane subunit). *E. coli* actively transport exogenous L-arabinose through the cell membrane by two distinguishable import systems: (1) a high-affinity ATP-driven system, called the arabinose ABC transporter, consisting of a single AraF, two units of AraG and two units of AraH (Daruwalla et al., 1981; Kolodrubetz and Schleif, 1981; Schleif, 1969) and (2) a low affinity proton-coupled arabinose symporter (AraE) (Daruwalla et al., 1981). AraC tightly regulates these genes through both positive and negative control. The most interesting feature is the mechanism for dual regulation of the P_{BAD} promoter. The "light switch" mechanism was proposed and reviewed by Schleif and coworkers (Schleif, 2010) and is shown in Figure I-7. In the absence of L-ara, the AraC dimer represses the expression of P_{ara} and P_{BAD} by preferentially binding two distant half sites I_1 and O_2, forming a 210 base-pair DNA loop (Martin et al., 1986).

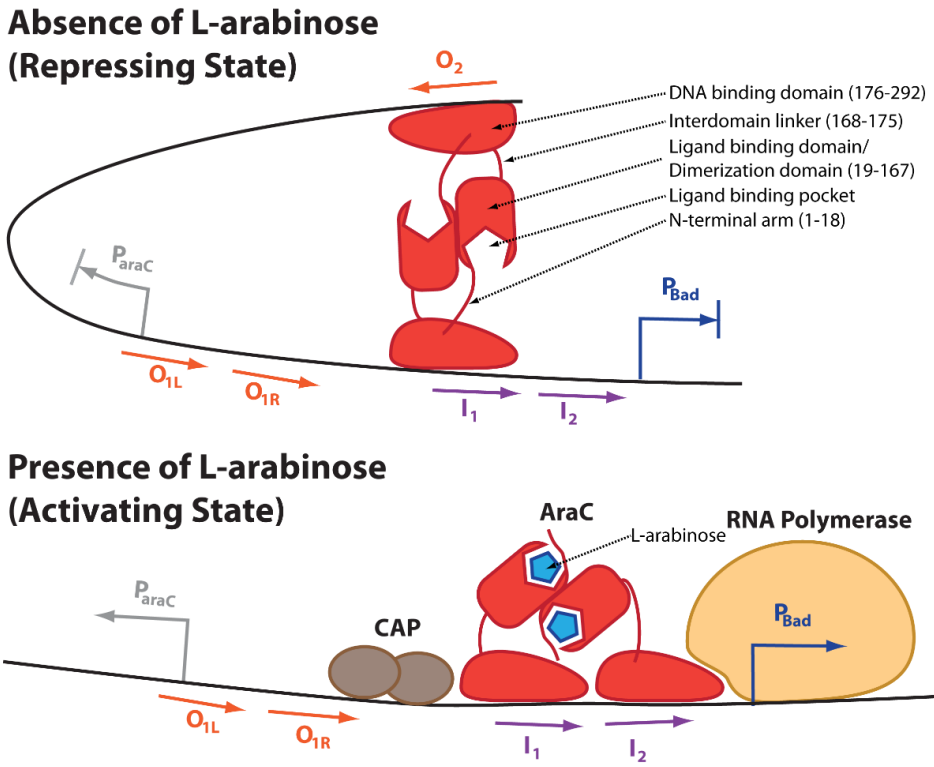


Figure I-7 The "light switch" mechanism of AraC transcriptional gene regulation of araC and the L-arabinose catabolic/uptake genes. The residues corresponding to the domain are presented in parentheses.

The N-terminal arm has been associated with stabilizing the repressive state in the absence of L-ara (Saviola et al., 1998) and directly interacting with bound L-ara. In the presence of L-ara, each monomer of the AraC dimer undergoes an allosteric interaction upon binding of a single L-ara molecule (Rodgers and Schleif, 2012) and the N-terminal arm stabilizes itself over the open end of the LBD. Under this conformation, the AraC dimer binds to half sites I_1 and I_2 and activates expression of P_{BAD} and P_{ara}.

AraC is an ideal platform to model small molecule biosensors because of the tight regulation it provides in *E. coli* and its natural ability to respond to a specific pentose stereoisomer with high sensitivity (µM range detection). However, the native regulation of the *ara* operon leads to two issues: 1) the system is expressed as all-or-none in the presence of L-ara and is not dose dependent and 2) only a fraction of a population will be induced (Siegele and Hu,

1997). These issues were addressed by Keasling and coworkers by constitutively expressing *araE* in the chromosome and deleting the *araFGH* operon. They replaced P_{araE} with a low to medium level constitutive promoters (~0.5-200 Miller units) and observed that the expression from the P_{BAD} promoter was dose-dependent and uniform throughout the population (Khlebnikov et al., 2001). Our lab generated a strain with *araC* deleted from BW27786, called HF19 (Tang et al., 2008). This strain was used throughout the work present in this dissertation.

In our lab, a whole-cell biosensor system was designed to take advantage of the high-throughput capabilities of fluorescence-activated cell sorting (FACS). As was mentioned above, the diversity that can be screened in a single experiment is dramatically improved with increasing throughput. In the original design, a dual plasmid system was implemented. A *gfpuv* gene was cloned downstream of the P_{BAD} promoter on a high copy plasmid, and the *araC* gene was cloned into a medium copy plasmid downstream of the P_{tac} promoter, inducible by the addition of Isopropyl β-D-1-thiogalactopyranoside (IPTG). The AraC system was expressed in the previously designed *E. coli* strain HF19 (Δ*araC*, Δ*araBAD*, Δ*araFGH*) (Tang et al., 2008). We have previously isolated and characterized biosensors from a 5-site saturated library (3×10^7 mutants at the nucleotide level) of AraC for D-arabinose (Tang et al., 2008), mevalonate (Tang and Cirino, 2011), and triacetic acid lactone (Tang et al., 2013). More recently, AraC was engineered to recognize ectoine, an osmoprotectant, and was subsequently used to increase ectoine production from a heterologously expressed pathway in *E. coli* (Chen et al., 2015). A deeper discussion on the applications of these AraC-based biosensors and the importance of each residue targeted for substitution will be discussed in subsequent chapters.

I.5) Summary and Prospective on Biosensor design

Small molecules are essential to everyday life including foods, fuels, and pharmaceuticals. Secondary metabolites make up a diverse set of natural bioactive small molecules derived from

plants and microorganisms that are nonessential for the organism's survival. Unfortunately, they are complex and difficult to extract for mass production, causing high costs for consumers. Heterologous expression of secondary metabolite metabolic pathways in amenable hosts, such as *Escherichia coli* and *Saccharomyces cerevisiae*, can provide cheaper alternatives to the production and discovery of novel secondary metabolites. This has been a major focus of metabolic engineers over the past decade (Zhou et al., 2008).

My research has focused on developing whole-cell biosensors as tools for metabolic engineering. Small molecule biosensors have been reviewed by several groups (Dietrich et al., 2010; Li et al., 2011; Michener et al., 2012; Zhang and Keasling, 2011). Heterologous expression of entire metabolic pathways is not trivial due to many factors including protein solubility, protein folding, and rate limiting reactions. Rational design of these pathways is limited by the information available concerning each reaction within the pathway. Combinatorial approaches require no prior knowledge of the pathway reactions, enabling a viable direction for metabolic engineering of recombinant pathways (Keasling, 2010). One such approach, directed evolution, combines Darwinian evolution with iterative rounds of high-throughput screening to rapidly analyze diverse mutant libraries of enzymes and transcriptional elements with the ultimate goal of discovering novel desirable traits (Cobb et al., 2013).

Directed evolution is a powerful technique but is typically bottlenecked by the throughput of the screening technique employed for isolation of targeted mutants (Aharoni et al., 2005b; Arnold, 1996; Fowler et al., 2008; Guntas et al., 2005; Schmidt-Dannert and Arnold, 1999). The biosensors discussed in this thesis were designed to increase the throughput of library screening by linking endogenous molecule concentration with an easily measurable output. We have concentrated on engineering transcriptional regulatory proteins to recognize specific metabolites and subsequently activate a phenotypic response. In the most basic case, the natural function of a transcriptional regulatory protein is to moderate a dose-dependent expression of downstream

genes through different conformational states, each with a different output signal. The state change is induced by changes in specific environmental factors such as small molecules, proteins, temperature, DNA, RNA, etc.(de las Heras et al., 2010). Common examples of output signals include enzymatic activity, growth, metabolite efflux pumps, luminescence, and fluorescence.

Our lab has already developed several biosensors using the AraC platform and fluorescence-activated cell sorting (FACS) for detecting endogenous D-arabinose (D-ara) (Tang and Cirino, 2010; Tang et al., 2008), mevalonate (mev) (Tang and Cirino, 2011), and triacetic acid lactone (TAL) (Tang et al., 2013). Figure I-8 shows three general scenarios for applications of AraC-based biosensors: (A) in the basic design, the AraC biosensor can be used to detect exogenous small molecules provided they are internalized by the cell; (B) the AraC biosensor can also be used to detect enzymatic reaction products; and (C) enzymatic libraries can be screened for a specific product. In particular, for mev and TAL, we have engineered the metabolic pathways expressed in *E. coli* using our biosensors. However, the best approach for rapidly isolating the best variants from our existing protein libraries was still unclear. Problems we encountered during previous studies included individually isolated clones that were "leaky" or unresponsive after several rounds of rigorous screening.

There is an inherent difficulty in screening transcriptional regulators versus enzymes. Mutations often create non-functional proteins, which can cause leaky expression of the reporter gene. This leads to the isolation and enrichment of false-positives during positive screening. We define a false-positive being a variant causing fluorescence in the presence of the target molecule but does not have a direct correlation with the target molecule concentration. False-positives can be alleviated by implementing a counter-screen or a negative screen. A negative screen highlights a negative attribute of an undesirable response, which can then be selectively screened against. Additionally, decoy molecules are used to impart a selective pressure to the screening, driving the population toward clones with greater specificity. Specificity is a highly desirable

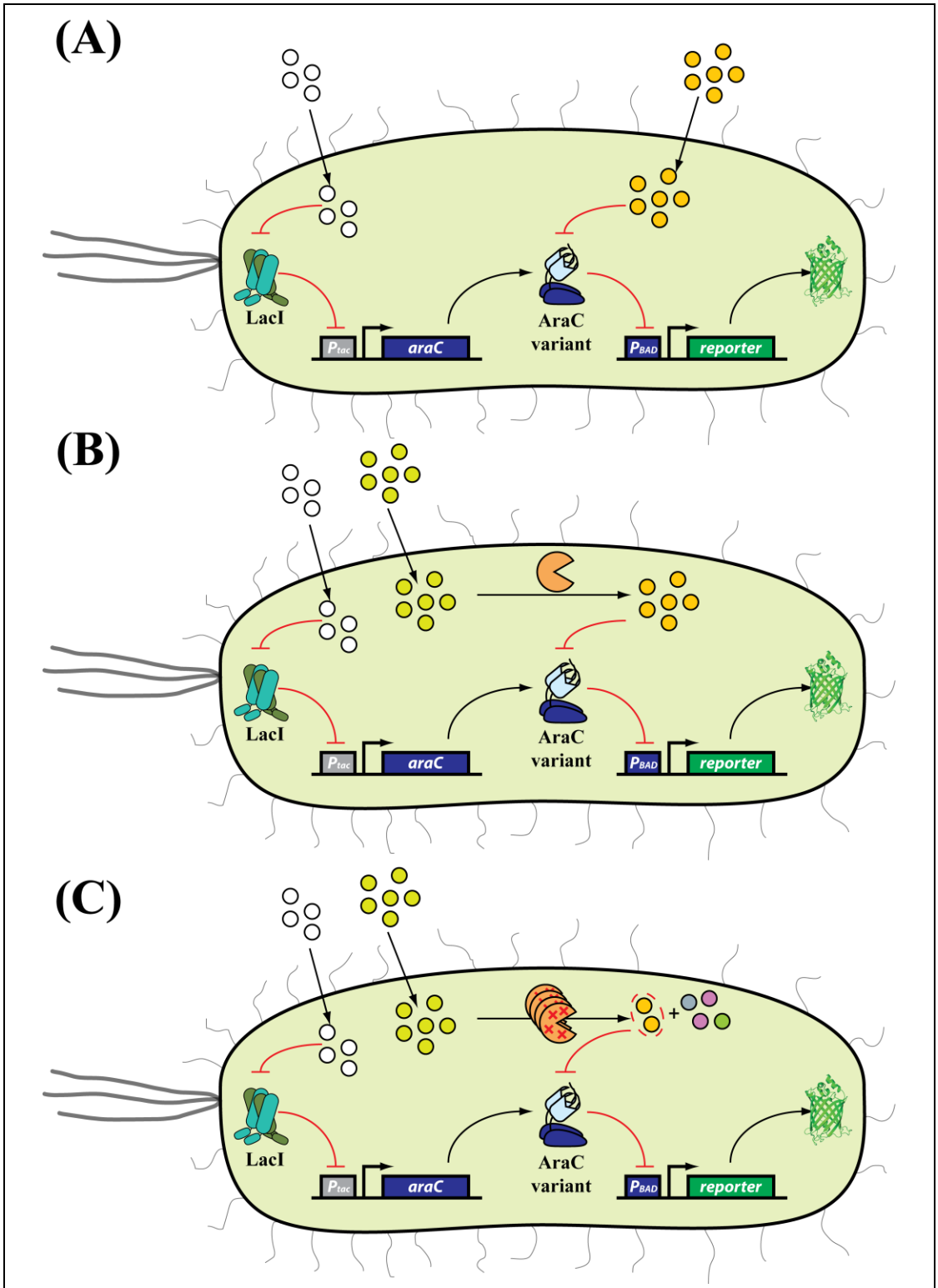


Figure I-8 General scenarios for AraC-based biosensor applications. Symbols: white circles, isopropyl-β-D-galactopyranoside (IPTG); orange circles, compound AraC variant recognizes; all other colored circles represent various small molecules.

trait for biosensor design owing to the complexity of metabolic pathways and the many intermediate molecules with similar structures to our target molecules, which can produce both false-positive or inhibit biosensor function. Also, a mutated enzyme may lead to the production of an unpredictable molecule, which could also inhibit the response of the biosensor. In this instance, the biosensor can be re-engineered for greater specificity using additional rounds of directed evolution.

The work presented in this thesis predominantly focuses on increasing the throughput of biosensor design However, greater throughput is not always necessary. This has been demonstrated by several successful examples of microbial engineering using low-throughput screening techniques, some of which were discussed as examples above. However, as more complex systems are evaluated and engineered, the demand for rigorous, customizable, and broadly applicable screening techniques grows.

My goals are presented in subsequent chapters and are to: i) increase the throughput and quality of past screening strategies to screen larger libraries with greater efficiency (Chapters II and IV); ii) to probe and rationalize optimal sorting schemes for isolating clones with desirable traits (Chapters II and IV); iii) to better understand and characterize the interactions of mutated residues with functional groups within the target effector (Chapter III); and iv) expand the utility of the transcriptional regulator biosensor platform for qualitative and quantitative whole-cell identification of broad range of small molecules (Chapters II and IV).

II Exploration of fluorescence-activated cell sorting sort schemes for rapid isolation of novel AraC-based biosensors

II.1) Introduction

High-throughput screening has been implemented in the isolation of novel antibodies, enzymes, and regulatory proteins for use in numerous research fields such as metabolic engineering and pharmaceutical development. Many different screening methods are available to isolate functional protein variants from large genetic libraries. Recently, fluorescence-activated cell sorting (FACS) has become popular owing to its throughput and increased availability over the past decade. FACS is a powerful tool for the high-throughput screening of large combinatorial libraries (Bonella et al., 2014; Kyte and Doolittle, 1982; Link et al., 2007), but the incredible throughput often allows us to easily overlook how the final functional variants were isolated. By understanding the population behavior throughout the screening process, we can more rapidly and precisely engineer novel protein functions. One of the limitations of FACS is the lack of a universal link between functionality and fluorescence (Golynskiy et al., 2011). As was mentioned in Chapter 1, this issue is currently being addressed using transcriptional regulatory proteins.

The primary focus of this work is to develop biosensors that can be used for metabolic engineering applications. These have been reviewed recently (Dietrich et al., 2010; Gredell et al., 2012; Michener et al., 2012; Zhang and Keasling, 2011). We have elected to use natural regulatory proteins that already successfully link small molecule recognition to a change in structure so that they can control expression of a reporter gene. Similar efforts have been demonstrated using RNA switches as descried by Michener and Smolke (Michener and Smolke,

2012). Recently, numerous molecular biosensors have been synthetically evolved from transcriptional regulatory proteins that activate expression of a reporter gene upon binding of a small molecule. For example, AlkS of *Pseudomonas putida* GPo1 naturally binds to short chain alkanes and activates transcription from the P_{alkB} promoter. After several rounds of directed evolution using FACS for screening an error-prone library of AlkS, the biosensor was improved for a higher sensitivity to C5-C9 alkanes (Reed et al., 2012).

Our group has already engineered endogenous biosensors for D-arabinose (D-ara), mevalonate (mev), and triacetic acid lactone (TAL) (Tang and Cirino, 2011; Tang et al., 2013; Tang et al., 2008). In particular, we used our mev and TAL biosensors to engineer herterologously expressed pathways in *E. coli* to produced higher titers than what had been previously reported in the literature. However, it is still unclear what the best approach is for rapidly isolating these biosensors from our existing DNA libraries. Our current project has three goals: i) to continue isolating novel biosensors for new target ligands; ii) to improve our understanding of the roles of substituted residues; and iii) to investigate the effect of the sorting scheme on the final isolated variants.

Here, we analyze the population behavior from two separate combinatorial libraries of AraC for response to D-ara, mev, and p-coumaric acid (pCA). Both libraries were separately screened in the presence of each compound using FACS and AraC variants were isolated with response to each of the three compounds. In order to understand the importance of alternating positive and negative sorting, each library was screened using eight different sort "paths", all incorporating 4 positive rounds of sorting and differing by number of negative sorts. The frequency of responsive clones and the overall response of the clones are used to compare different sort paths to determine if an optimal FACS-based screening scheme exists. These results are then applied to screen for additional AraC-based biosensors targeting additional value-added products.

II.2) Results

2.2.1) Design of pyramid sorting scheme

An alternative sorting scheme to alternating positive and negative sorting was desirable in an effort to expedite the screening of a combinatorial library using FACS, improve our understanding of the effects the various residues targeted, and investigate the impact of the sorting path on the resulting final population. The sorting protocol was designed to incorporate all possible combinations of positive and negative sorting, with each sort path containing four total positive sorts and no consecutive negative sorts. Consecutive negative sorts were neglected

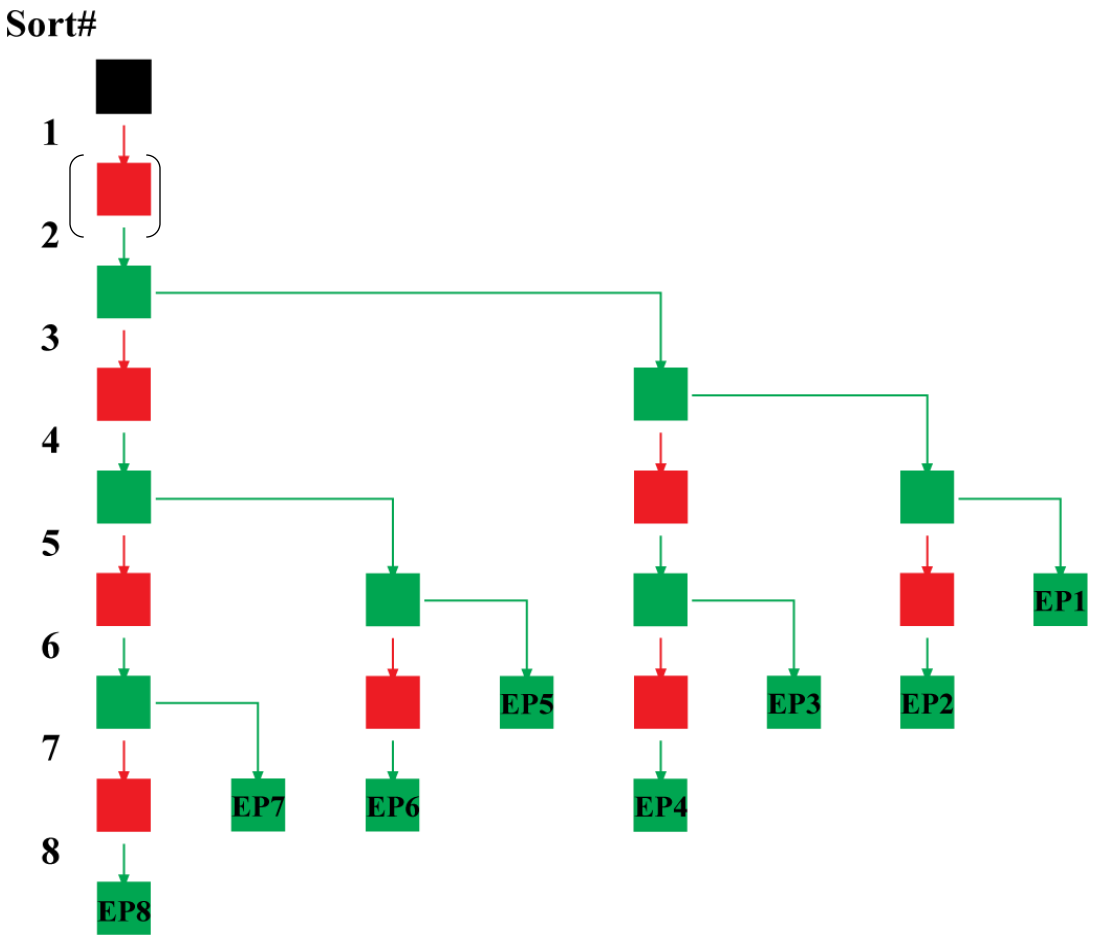


Figure II-1 Schematic of pyramid sorting scheme. The pyramid scheme was designed to include each possible path with four positive sorts and no consecutive negative sorts. The first negative sort in brackets was optional. Abbreviations: Endpoint Populations, EP (number indicates the path). Black box, naïve library; Green box, positive sorted population; Red box, negative sorted population.

Table II-1 List of targeted residues for both libraries (Jlib1 and Slib4) sorted in this project. Each site was saturated at the DNA level with degenerate oligonucleotides containing NNS sites at the targeted residue codon position. Each library had a total of 5 targeted residues.

Library	Proline 8	Threonine 24	Arginine 38	Histidine 80	Tyrosine 82	Histidine 93	Total Sites Targeted
Jlib1		NNS	NNS	NNS	NNS	NNS	5
Slib4	NNS	NNS		NNS	NNS	NNS	5

due to the potential loss in overall signal from a respective population. This layout lead to eight unique sorting paths, together forming a pyramid sorting scheme (Figure II-1). Libraries JLib1 and SLib4 were both separately subjected to the sorting pyramid and treated with three different target compounds. Each library was constructed using overlap extension PCR with degenerate primers, and designed targeting five residues within the AraC ligand binding domain for saturation mutagenesis (Table I-1). Additionally, a pre-negative sort of the naïve libraries was also investigated. Altogether, 8 different pyramids were constructed and analyzed for this study. Compounds, D-arabinose and mevalonate (Figure II-2), were selected as two of the target compounds for this study because previous AraC-based biosensors have been isolated for these compounds. The third compound, p-coumaric acid, was chosen due to the lack of an available biosensor, whether AraC or any other TRP. Using this pyramid sorting scheme we sought to answer the following questions:

1) *Is a first round negative sort advantageous in the sorting scheme?*
2) *Is there a correlation between leakiness and response?*
3) *Does a mutation at residue P8 of the N-terminal arm lead to clones with high leakiness?*
4) *Are the sorts sufficiently stringent for reducing the diversity of the library and enriching the desired functional clones?*
5) *Is there an optimal sorting strategy and does it hold true for different compounds?*

To address these questions, we subjected both of the aforementioned AraC libraries to FACS, via the pyramid sorting scheme, and sorted each library for response to three different

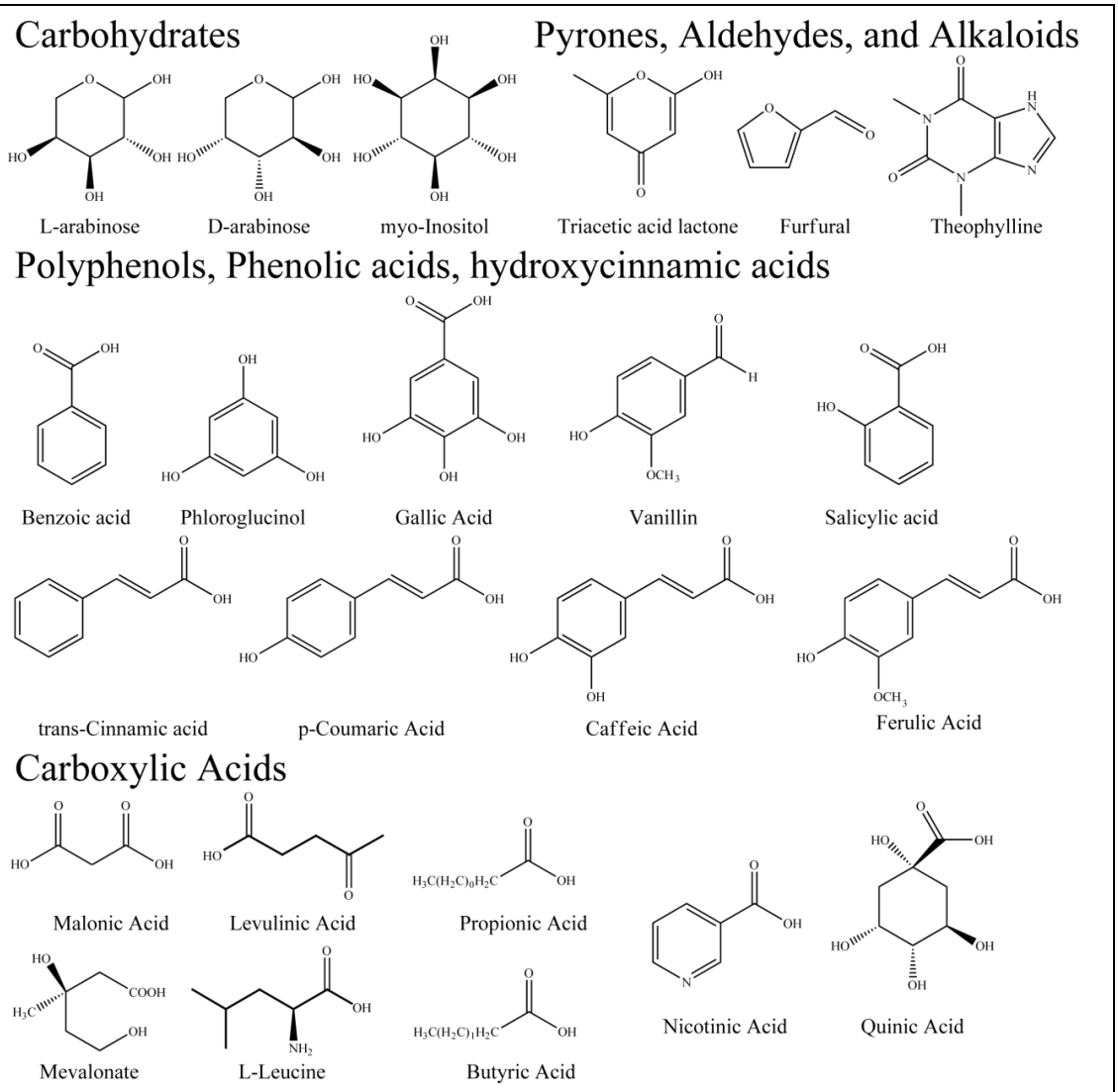


Figure II-2 List of compounds screened in this study. L-arabinose is the native ligand of AraC, whereas AraC has little to no response to the other listed compounds.

target molecules, D-arabinose, mevalonate, and p-coumaric acid. Each of the two libraries discussed above was transfected into HF19 cells harboring the reporter plasmid pPCC442 (P_{BAD}-*gfpuv*). Subsequent calculation of the transformation efficiency confirms the starting libraries have greater than 10^9 total transformants, which is greater than 10-times the naïve library size (3.4×10^7). Subcultures from the resulting transfected cells were mixed with glycerol to a final concentration of 20% and subsequently frozen in liquid nitrogen and stored in a -80°C freezer. By freezing aliquots of the library, we are able to readily inoculate and subject the populations to the aforementioned treatments without having to transform prior to each experiment, and the

starting cultures would be identical from day to day. The wt-AraC biosensor was also treated using the same protocol as the libraries. The wt-AraC biosensor was run concurrently with all populations and presented consistent flow and spectrophotometer data. The frozen stock of the wt-AraC biosensor was viable up to and beyond a year under the conditions described. For this study we did not incorporate the effects of time on the biosensors and their response, so cultures were induced for 16 hrs prior to flow analysis and subsequent sorting. This time was chosen based on previous sorting results in our lab, where the library population showed a noticeable shift in the presence of the target compound after 16 hrs. Sorts were performed on a FACSJazz pre-production model. Sorts for this study were designed to maximize throughput, while maintaining a high level of purity, which was obtained using the sort mode "1.0 drop Yield" after testing each mode offered in the sorting software. Each naïve library, JLib1 and SLib4, was subjected to a negative sort in an effort to determine the influence of false-positives (variants lacking repressibility) on subsequent rounds of sorting and the endpoint population. The negative sort was carried out by applying a gate to the library populations in the fluorescence (Excitation 488 nm; Emission 530/40 nm) histogram relative to the wt-AraC biosensor absent L-arabinose, encompassing 99% of the population. The immediate subsequent sort was a positive sort and the top 1% of the most fluorescent cells in the presence of the target effector were collected. The first round negative sort was only carried out for the D-arabinose pyramids and compared with pyramids starting with a positive sort. The populations initially negative sorted responded to D-arabinose more dramatically in early sorts compared with the same respective population from a pyramid without the first round negative sort. Each subsequent round of negative sorting was performed with the population in the presence of 10 mM L-ara, but using the same sort gate as described above. All positive sorts were collected from the top 1% in the presence of the target compound. Each of the collected samples was diluted in LB (5x the volume collected) supplemented with chloramphenicol and apramycin and grown to saturation. The saturated cultures were mixed with glycerol (20% final concentration) and aliquots were frozen for future

use. Cultures for sorting were inoculated with 1 mL of the thawed culture from the previous sort in 10 mL of LB supplemented with chloramphenicol, apramycin, IPTG, and either no compound (FL_{OFF}), 10 mM L-ara (FL_{ara}), or the target compound (FL_{ON}). Upon isolating the final or "endpoint" population, individual clones were screened in deep well plates, each clone was individually treated with using the same three conditions as described above. The endpoint clones were used to assess the sort path and sorted libraries.

First, we sought to determine if a first round negative sort would help improve the enrichment of a variants with increased repressibility. If the frequency of false-positives in the positive sort gate (top 1%) is high, we risk rejecting functional clones during the positive sort. All endpoint population from each library were pooled to compare the effects of the first round negative sort on the final isolated clones (Figure II-3). We find that the average fold-response (FL_{ON}/FL_{OFF}) from the Pyramid 1 (JLib1; Neg; D-ara) is slightly higher than the respective library population sorted without the first round negative sort (Pyramid 3). All p-values are calculated using the nonparametric Mann-Whitney U-test (populations are not normally distributed) and shown in Supp. Table B-1. Each red sphere in the plots represents a single clone of 192 clones sampled for a pyramid with a first round negative sort and green spheres are clones from pyramids absent the first round negative sort. The black sphere represents the average of all 192 clones from the pyramid. Pyramids 2 (SLib4; Neg; D-ara) and 5 (SLib4; Pos; D-ara) were determined to not have significantly different means (1.4 ± 0.5 and 1.5 ± 0.3 fold-response, respectively), but Pyramid 2 lead to three clones with the highest response among all four pyramids sorted for D-ara response (4.4-, 4.1-, and 3.9-fold response). Each one of these clones has a high background fluorescence compared to wt-AraC (65 ± 14 rfu), and is referred to as "leakiness" ($FL_{OFF,clone}/FL_{OFF,wtAraC}$). Also, we discovered no significant difference between Pyramids 1 and 3, and Pyramid 2 had an average leakiness greater than Pyramid 5 (mean = 36- and 15-fold; median = 2.5- and 2.7- fold, respectfully). The three clones with the highest

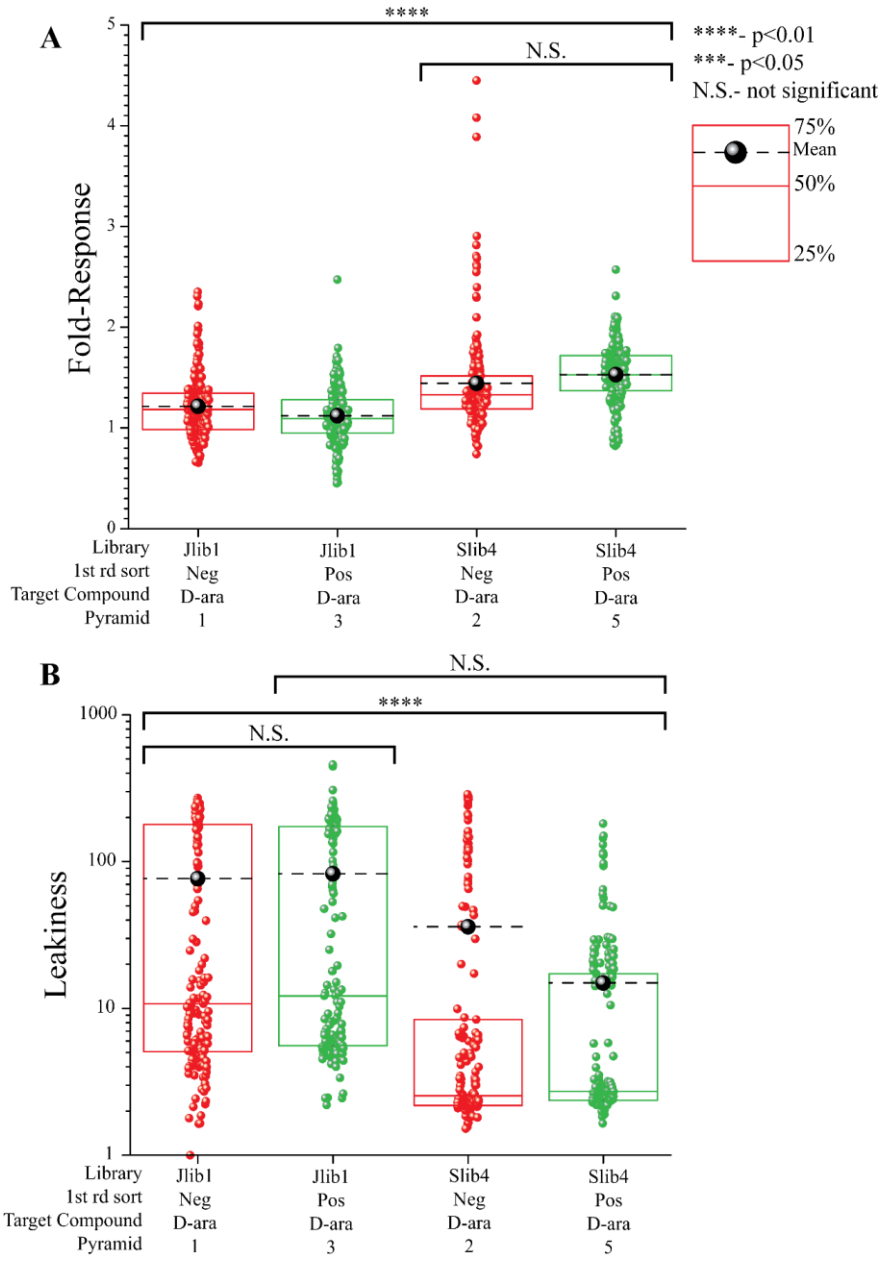


Figure II-3 Box plots comparing (A) fold-response and (B) leakiness of endpoint clones from pyramids with and without first round negative sorts.

leakiness were all isolated from Pyramid 3 (460-, 430-, and 350-fold), which lacks a first round negative sort, and each clone shows no response in the presence of 100 mM D-ara. Interestingly, each of these clones was isolated from sort paths with Pyramid 3 that included at least two rounds of negative sorting, which would suggest that the sort purity was not high during rounds of negative sorting. The purity can be increased by changing the sort modes and sorting at lower

event rates (i.e. number of cells detected per second), as well as increasing the sort stringency. We initially proposed that a first round negative sort would help decrease the frequency of false positives in the endpoint populations, and though the comparison of average leakiness of endpoint populations would suggest the negative sort leads to clones with higher leakiness, the most responsive clones were isolated from a population negative sorted in the first round. Also to note, JLib1 library lacking the mutation at P8 lead to clones with an overall higher average leakiness and lower response than the SLib4 clones, independent of which sort method was used in the first round. As we expanded our observations to include the leakiness of the other Pyramids sorted (4, JLib1, p-coumaric acid; 6, SLib4, p-coumaric acid; 25, JLib1, mevalonate; 26, SLib4, mevalonate), a pattern emerges showing a lower level of leakiness from all pyramids derived from SLib4 naïve library (Figure II-4). Plotting the leakiness against response of all the clones isolated (Figure II-5) shows a trend towards clones with a response to their respective target compound to have a lower level of leakiness. Of all clones showing greater than 2.5-fold response to their respective ligand, the majority have a leakiness below the average leakiness of all observed clones.

The purity of the sort may also be affected by the post sort treatment of the cells. Here, collected cells were simply cultured in rich medium absent induction of the biosensor, but the post-sort viability of the collected cells will affect the growth of the cells and potentially introduce a bias towards healthier cells and not necessarily cells expressing an improved protein variant. Samples of HF19 cells expressing the AraC biosensor system from the dual plasmid system were collected ($N_{collected}$) from sorting the top 1% of most fluorescent cells. The collected samples were directly plated and the resulting colony forming units (N_{cfu}) were counted. Cell survival ($N_{cfu}/N_{collected}$) was 23.6 ± 0.1%, which could negatively impact our population diversity. We assumed that the low survival rate was due to overexpression of recombinant proteins

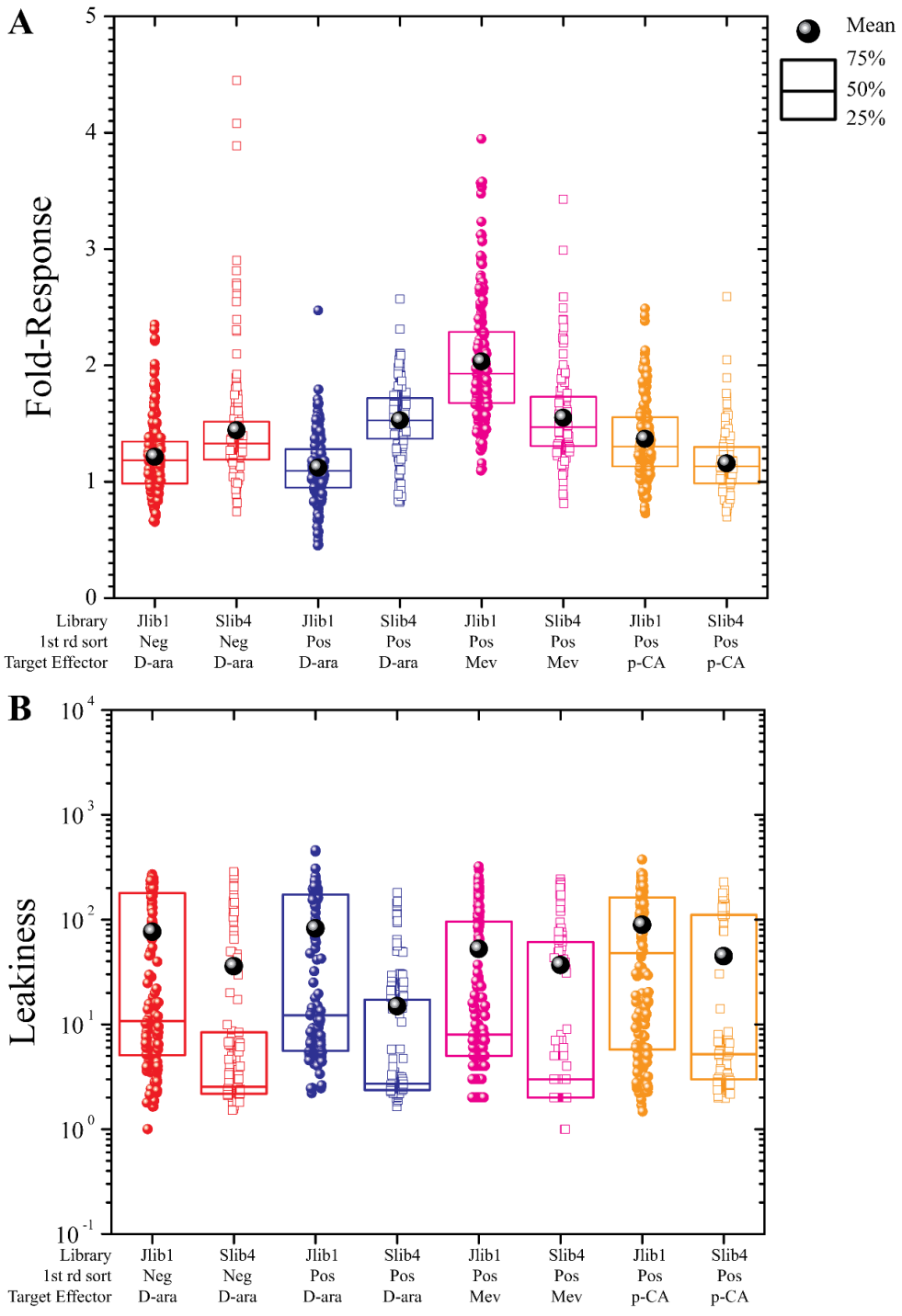


Figure II-4 Box plots for comparing fold-response and leakiness of all pyramid endpoint clones. Symbols: sphere, clone from Jlib1; open box, clone isolated from SLib4; different colors indicates the compound or first round treatment; black sphere, the average of all 192 clones.

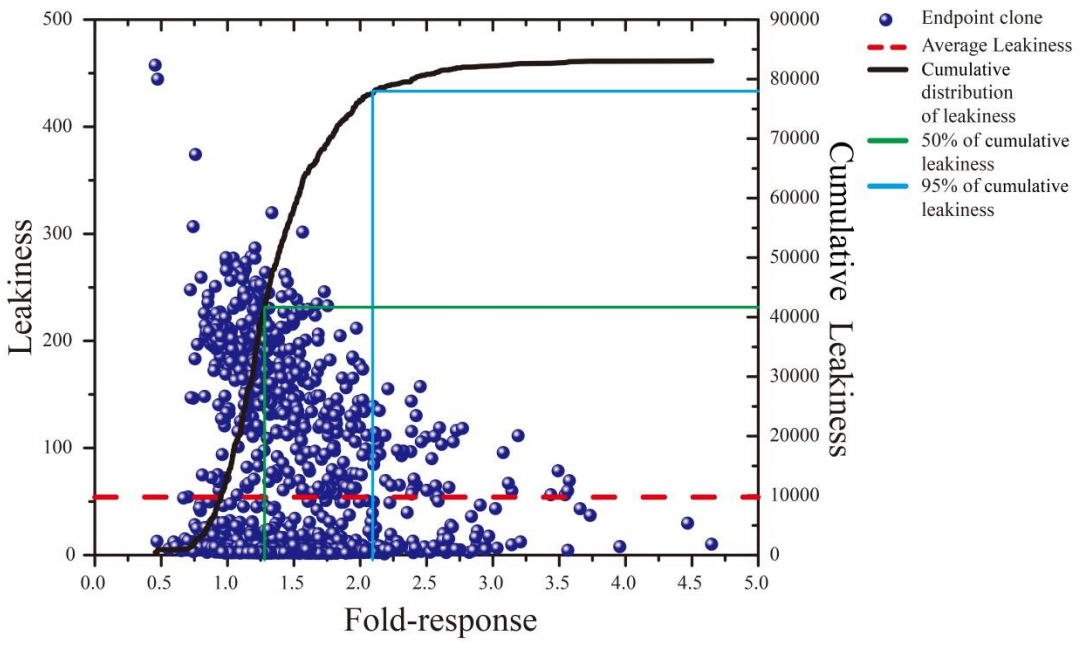


Figure II-5 Scatter plot representing the dependence of fold-response on the leakiness of isolated clones. Each blue sphere represents a single clone isolated from an endpoint assay. All clones isolated from endpoint assays are represented (1,536 clones total).

(Gill et al., 2000), stress of maintaining two plasmids (a medium and a high copy plasmid) (Bentley et al., 1990; Birnbaum and Bailey, 1991), and being sorted from a population that was in late stationary phase of growth (Foster, 2007). Further protocol development is discussed in Chapter IV and addresses the low viability of the positive sorted samples.

So far we have determined a first round negative sort does not reduce the overall leakiness of isolated endpoint clones but does afford responsive clones, there is a trend towards clones with higher response in the presence of their target compound to have lower leakiness but does not imply that clones with low leakiness have a high response, and finally, populations sorted from SLib4 with the saturated N-terminal arm residue (P8) lead to lower leakiness compared to populations subjected to the same respective conditions derived from JLib1. However, SLib4 did not always lead to clone with the highest response to the target compound. One of the main goals we sought to achieve was to discover and optimal sort path using a pyramid of sorting paths with

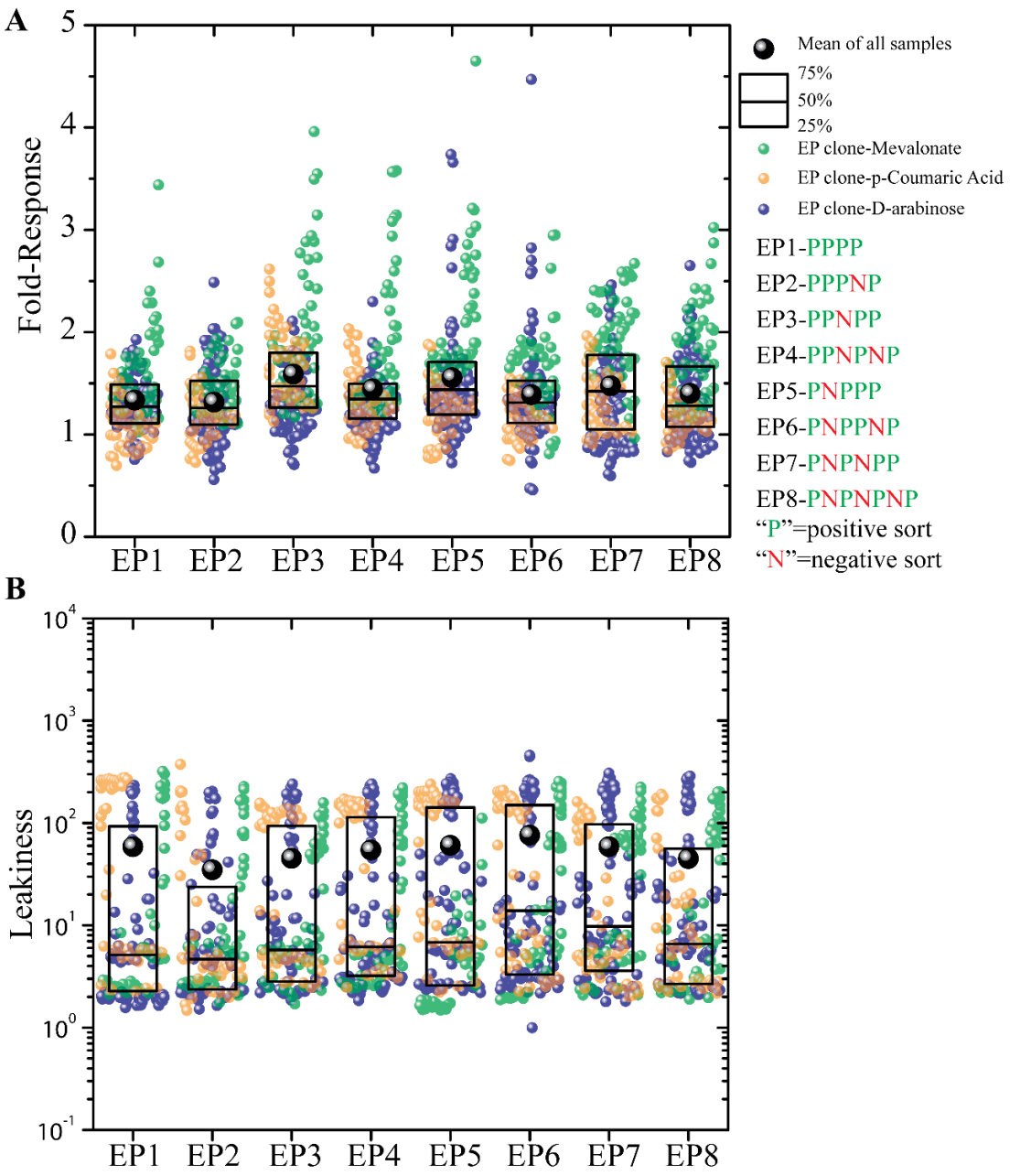


Figure II-6 Box plots comparing fold-response and leakiness of endpoint clones with respect to their endpoint population. Clones are represented as a sphere, which are colored according to the target compound they were screened with.

eight different sort paths. Here we define an optimal sort path as a sorting scheme that leads to the most responsive clones consistently, regardless of target compound, and does so with the least number of sorts. The most responsive clones, regardless of the target compound, were isolated

from endpoints EP3, EP5, and EP7 (Figure II-6). Each of these endpoints contains at least one negative sort in the sort path. However, a clone showing 3.4-fold response to 30 mM mev was isolated from EP1, lacking any negative sorts including a first round sort. Closer examination of the endpoint populations based on their target compound, we find that EP3 produces the highest frequency of p-coumaric acid clones (19% of clones show > 2.0-fold response) and several highly responsive mevalonate clones (8% of clones show > 3.0-fold response), but no D-ara responsive clones greater than 2.5-fold response. Endpoints EP5 and EP7 both result in no p-coumaric acid clones with greater than 2.0-fold response, but several D-ara (5% each) and mevalonate (25% and 6%, respectively) clones were isolated from these endpoints with greater than 2.5-fold response. Also, alternating negative and positive sorting , represented as EP8, do not significantly decrease the average fold leakiness of the clones (40 ± 70) comparted to the majority of other endpoint populations, including EP1 where there were no negative sorts in the sort path (60 ± 90). Note that these pyramid sort paths do not include the first round negative sort. The results of the endpoints suggest responsive variants can be isolated from multiple different sort paths.

Finally, the stringency of the sorts was evaluated for several populations within the pyramids using next generation sequencing (NGS). The plasmid DNA was isolated from select populations and amplified using primers specifically designed for NGS. The samples were run on an Illumina MiSeq NGS sequencer (sequencing by synthesis technology). Briefly, sequencing by synthesis (SBS) is a technology where bridge amplification enriches the local population of a bound sample DNA within a channel, followed by polymerization of the amplified DNA using fluorescently labeled nucleotides. The addition of the fluorescent nucleotide is measured in real time over the entire channel and recorded relative to the position in the channel. This technology measures the forward and reverse sequence of each sample, with samples reads up to ~400 bp. For a more detailed explanation, the Illumina website provides excellent resources (http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html).

For our samples, an average of 1.6 x 10^6 sequences were resolved and analyzed from each sort population. Sequencing of the naïve JLib1 and SLib4 confirmed the quality of the library as assessed by the diversity and the nucleotide saturation frequencies (Figure II-7). For all "NNS" codons, each of the four nucleotides was accounted for at "N"-sites and only C's and G's at S-sites. The presence of additional nucleotides at the "S"-sites (i.e. codons R38 and H80) were investigated and found to be the native nucleotide at that particular position. Furthermore, NGS results of selected populations within Pyramids revealed a high degree of diversity still remaining after three total rounds of positive sorting (64±16% enrichment frequency, defined by the number of reads with greater than 10 repeats per the total number of sequences), with the exception of SLib4 sort in the presence of D-ara and a first round negative sort (88 ± 6% enrichment

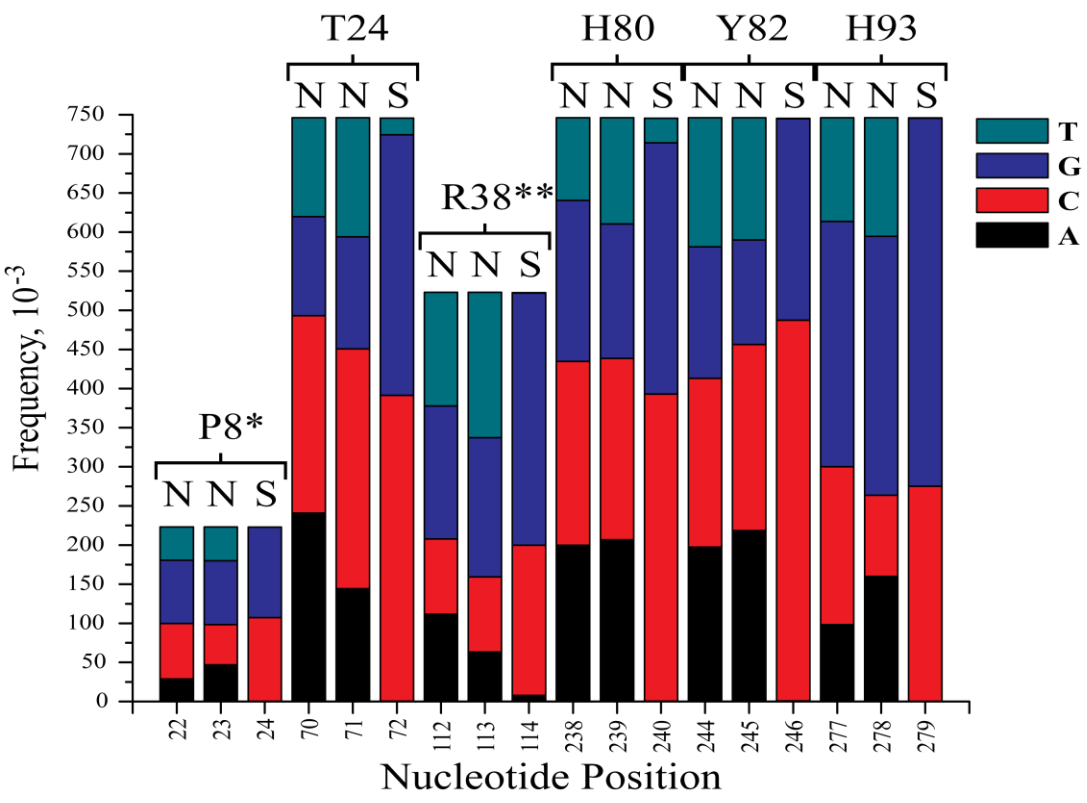


Figure II-7 Next generation sequencing of naïve library confirms high quality library construction, as evident by the distribution of bases at targeted nucletides. The stacked bar graph represents the distribution of mixed bases at each nucleotide targeted for mutagenesis.

frequency). The change in enrichment frequency only increases 1.4 ± 0.4 fold from each positive sort, excluding the first round of positive sorting. The last round of positive sorting (no samples were sent from the endpoint populations because they were thought to be highly enriched) would not dramatically affect the endpoint population enrichment frequency and each endpoint would still have a diverse population. Therefore, further rounds of FACS or other screening methods may be used to better enrich the best clones.

2.2.2) Analysis of isolated clones for D-ara, mev, and TAL

Endpoint populations were all screened in deep well plates, with quadruplicate repeats for each clone screened. The plasmid DNA was first isolated and transfected into freshly prepared HF19 electrocompetent cells harboring pPCC442 to ensure there were no advantageous mutations in the genome of the cells after multiple rounds of sorting. Clones were isolated on solid medium and used to inoculate four replicate starter cultures. These starter cultures were grown to late-exponential phase growth and then diluted in subcultures to an OD_{595} 0.2. Each of the cultures was subjected to three treatments, no ligand, 100 μM L-ara, and the respective target compound (100 mM D-ara, 2 mM p-coumaric acid, or 30 mM mevalonate). Treated cultures were grown for 16 hrs and evaluated for growth and fluorescence. The best clones, determined by fold-response greater than the average fold-response of the endpoint population plus one standard deviation, were isolated and re-cloned into the parent vector (pFG1). The sequences of the best 12 clones from each Pyramid clone were evaluated (Supp. Table B-2). Several clones incorporated deletions, insertions, nonsense and missense mutations. Curiously, none of the clones isolated in response to D-ara had a similar sequence to the previously isolated D-ara variants (Tang et al., 2008). The top two to four responsive clones for each compound were selected and subjected to a range of target compound to determine the dose dependent response (Figure II-8). All of the clones show a higher background than wt AraC, but also every clone has a greater response to the

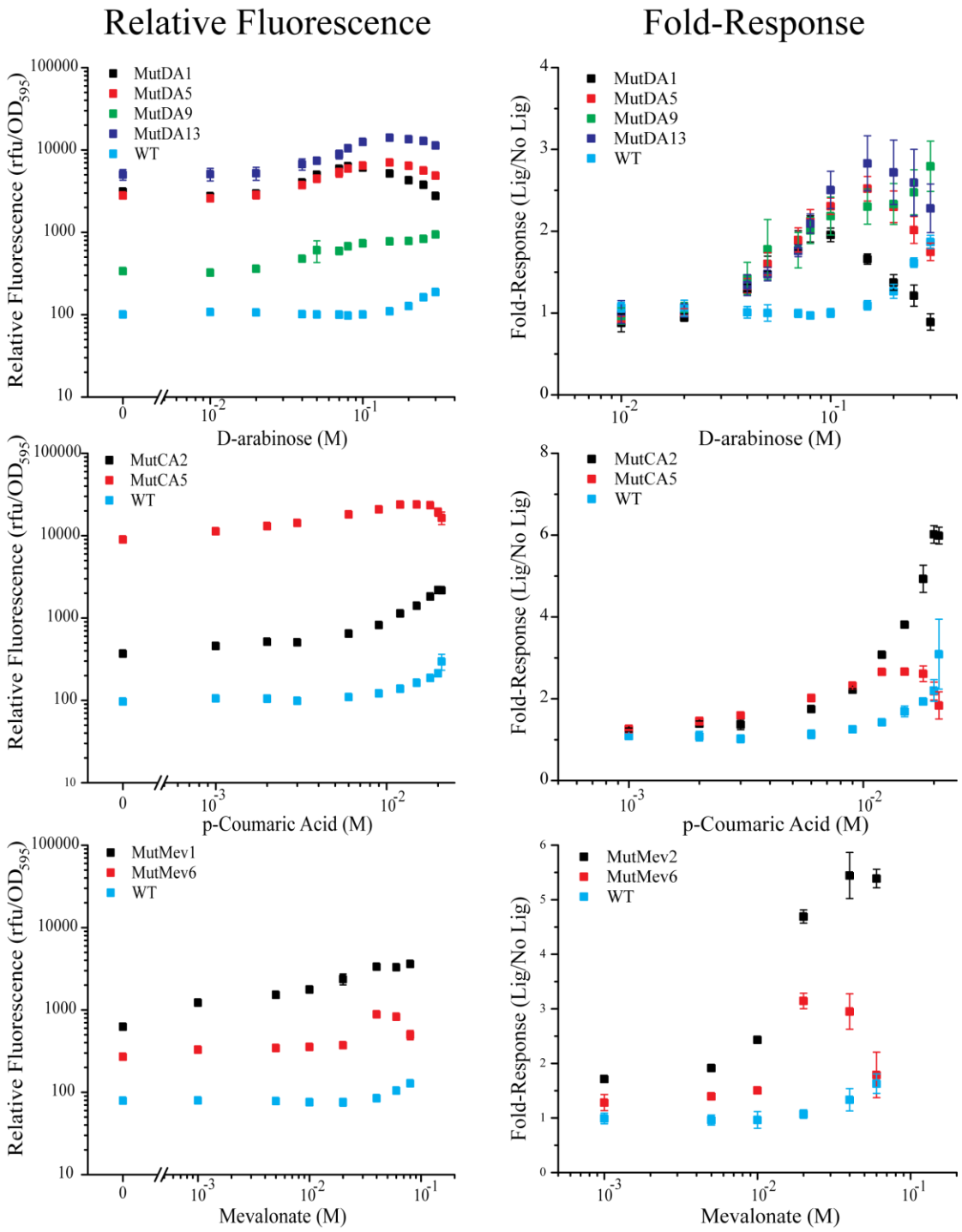


Figure II-8 Dose response curves for best clones for D-arabinose, p-coumaric acid, and mevalonate. All data points were plotted from an average of 4 separate samples. The standard deviations are plotted for all points as well. The wild-type (WT) AraC response was also measured and recorded for each target effector and is platted in blue

target compound than wt-AraC. Clones with a lower level of leakiness lead to the largest response to the target compound (except for the D-ara clones), thereby reiterating the point that highly responsive clones are more often isolated from clones with lower leakiness. Above 20 mM p-coumaric acid, growth was inhibited, so a full dose curve was difficult to establish for wt-AraC and AraC-MutCA2.

2.2.3) Application of new screen to isolate additional biosensors for other target compounds

Additional biosensors were screened for after determining a proper sort path based on the previously discussed results. Functional clones were isolated from EP1, absent any negative sorts and therefore the least time-consuming sort path and the some of the top functional clones for p-coumaric acid and mevalonate were from JLib1. Therefore, all subsequent sorts were simply four consecutive positive sorts starting with naïve JLib1. Ideally, both libraries would be screened, but to explore more target compounds, we forwent screening of SLib4. After a comprehensive search of the literature for bioactive compounds to target for screening, we screened JLib1 in the presence of 2 mM trans-cinnamic acid, 10 mM nicotinic acid, 1 mM theophylline, 10 mM propionic acid, 10 mM butyric acid, 10 mM vanillin, 2 mM ferulic acid, 10 mM salicylic acid, 10 mM levulinic acid, 10 mM gallic acid, 10 mM furfural, 10 mM succinic acid, 10 mM malonic acid, 25 mM myo-inositol, 10 mM quinic acid, 10 mM phloroglucinol, and 2 mM 1-pentinol. Following the endpoint screening assay in deep well plates, several biosensor candidates were isolated and their respective sequences were determined and reported in Supp. Table B-3. The data represented in Table II-2 was taken from the endpoint assay data, where each standard deviation was measured from four replicate cultures. According to the endpoint responses, most clones have almost completely lost response to L-ara and have a moderately high level of leakiness. All of the top clones were re-cloned into the parent vector (pFG1) and the specificity of each clone was tested in a cross-reactivity assay in deep well plates, where each isolated clone

Table II-2 Heat map of clone specificity. The green indicates the highest response; the red represents the lowest response. The first two columns describe the origin of the clone and the third column lists the compound the particular clone was screened with. All reported concentrations used in this assay were the same concentrations used during sorting. Abbreviations: L-arabinose, L-ara; D-arabinose, D-ara; p-coumaric acid, pCA; mevalonate, Mev; trans-cinnamic acid, t-CinnA; propionic acid, PropA; butyric acid, ButA; theophylline, Theo; ferulic acid, FerA; levulinic acid, LevA; succinic acid, SucA; phloroglucinol, Phloro

Endpoint	Clone	Screened	L-ara 10 mM	D-ara 100 mM	pCA 2 mM	Mev 30 mM	t-CinnA 2 mM	PropA 10 mM	ButA 10 mM	Vanillin 10 mM	Theo 1 mM	FerA 2 mM	LevA 10 mM	SucA 10 mM	Phloro 10 mM
-	WT	L-arabinose													
J110	11D	D-arabinose													
J111	9D	D-arabinose													
J152	4	p-Coumaric Acid													
J268	6D	Mevalonate													
J281	3D	Mevalonate													
J235	14D	t-cinnamic acid													
J235	18D	t-cinnamic acid													
J321	21D	Nicotinic Acid													
J280	23D	Propionic Acid													
J280	20D	Propionic Acid													
J307	15D	Butyric Acid													
J307	4D	Butyric Acid													
J309	12D	Vanillin													
J323	21D	Vanillin													
J279	9D	Theophylline													
J263	13D	Theophylline													
J340	11D	Ferulic Acid													
J342	23D	Levulinic Acid													
J357	1D	Succinic Acid													
J357	7D	Succinic Acid													
J348	6D	Phloroglucinol													
J348	24D	Phloroglucinol													

was subjected individually to a panel of compounds. Surprisingly, most clones lost a great deal of response to their respective target compound, such as MutButA1 had a 3.4-fold response before re-cloning and 1.7-fold response after in the presence of 10 mM butyric acid. Most of the clones responded strongly in the presence of 10 mM phloroglucinol and 10 mM propionic acid. The cross reactivity assay was run several time independently and the same results were seen. Why did we not isolate clones all of these clones when we screened for phloroglucinol and propionic acid? This could have been due to the procedure for isolating clones from the endpoint assay, where only the top clones are selected for further analysis. The endpoint population of propionic acid resulted in an average of 2 ± 0.5 fold-response, where 28 of the 48 clones screened had greater than a 2-fold response. Therefore, it is possible these clones were present in the endpoint population, but were not selected because the clones that were selected showed the highest response. Similar results were seen from the phloroglucinol endpoint population (50% of the population had >2-fold response). Clone J111-9D, isolated for a response to D-ara was highly promiscuous, showing a mild response to eight different compounds. This clone may be a good variant to use for making future libraries for isolation of variants with greater specificity. The butyric acid clones are particularly interesting because they show a strong response to 2 mM t-cinnamic acid and a weaker response to 2 mM p-coumaric acid, where t-cinnamic acid can be directly converted to p-coumaric acid using the enzyme trans-cinnamate 4-monooxygenase.

II.3) Discussion

Initially, we sought to characterize and optimize the sorting strategies for isolating variants of AraC with response to a target effector from a large combinatorial library. Our exploration of residue substitutions in AraC for detection of small metabolites has been limited to six residues in the LBD but has produced several novel biosensors, indicating that AraC can be used as a versatile platform for biosensor discovery. Using a target-based screen, such as the one presented here, limits the throughput of analyzing the complete fitness landscape (Eggert, 2013), but a

highly efficient screening protocol could alleviate some of the burden. The more targets we can screen, the better we can determine which small molecules are best to target with our AraC-based biosensor design. Here, access to previously reported AraC variants and their parent library afforded us the unique opportunity to optimize our screening protocol to more effectively screen libraries for novel biosensors. Therefore, we investigated the total number of sorts and the sort stringency. The initial sorting scheme was designed to incorporate all possible combinations of negative and positive sorting, without any consecutive negative sorts. Previous AraC clones were isolated after sorting the top 0.1% of a positive population (Tang and Cirino, 2011; Tang et al., 2013; Tang et al., 2008). For this study, we adjusted this to the top 1% of most fluorescent cells, so we didn't discard variants with a lower induced signal that may also present lower background. The decrease in sort stringency would also lead to a lower enrichment. This can be overcome by further rounds of screening using different screening methods, which is highlighted in Chapters III and IV. Caution was held when determining the degree of stringency. If the stringency was decreased too much, there was a risk of maintaining a high level of diversity despite the number of rounds of sorts that were performed. Despite our efforts to maintain a high level of sort stringency, high-throughput sequencing of the sort populations within the pyramids show minimal enrichment (see Appendix E).

For a clone to be desirable, the clone must possess both an ON signal in the presence of the target compound (high fluorescence) and an OFF signal in the absence of the target compound (low background fluorescence). An AraC variant may be in three general states in the absence of a target compound: i) fully repressive, ii) partially repressive, or iii) lacking any repression. In the presence of a ligand, the same states exist, but are in terms of activation instead of repression. It should be noted that a mutation in the TRP could also render a TRP response to an endogenous molecule that is not preferentially targeted, but these variants should be discarded during negative

sorts. Also, simply collecting the top most fluorescent clones could lead to the isolation of a population consisting of clones lacking any repression.

The results presented here show that there is a loss in response to L-ara for the majority of AraC variants isolated, despite the sorting scheme, which suggests that negative sorts in the presence of L-ara are not crucial. However, this does not imply that negative sorts would not be necessary for increasing the specificity in the presence of target-like compounds. For example, trans-cinnamic acid is a precursor to p-coumaric in the pathway leading to resveratrol and only differs by a single hydroxyl group. As such, it may be desirable to ensure that no clones that were isolated for p-coumaric acid respond to trans-cinnamic acid. The specificity may be further improved by negative screening against a "cocktail" of target compound analogs. However, the increased stringency could cause us to lose variants that respond to the target compound in early round of directed evolution. Therefore, a variant with a promiscuous response (but still responds to the target compound) may be desirable after the first round of evolution, and subsequent rounds of evolution can focus on increasing the specificity.

Though the goal of this study is to determine an optimal sorting scheme, the data suggests there are multiple solutions. As many experiments are destined to have alternative and unexpected outcomes, lessons are learned. Sorting was not as efficient as expected, which indicates the need for further analysis of the screening protocol and is the subject of Chapter IV. The toxicity of the target compounds induced false responses or reduced responses. Further exploration of culturing conditions may help reduce these false responses. However, the findings presented in this study convey a lack of information about FACS-based screening and its true throughput. These results establish a foundation for reevaluating the use of FACS as a standalone screen and leads to the studies discussed in subsequent Chapters III and IV.

III Analysis of amino acid substitutions in AraC variants that respond to triacetic acid lactone

III.1) INTRODUCTION

Transcriptional regulatory proteins (TRPs) induced by small molecules have emerged as useful molecular reporting tools in whole-cell screening (Dietrich et al., 2010; Eggeling et al., 2015; Hansen and Sorensen, 2001; Schallmey et al., 2014). Here, the natural link between molecular recognition and gene expression is used to report the presence and production of a metabolite of interest. For cases where there is no known TRP that responds to a desired compound, an existing TRP may be engineered to exhibit altered specificity toward the compound of interest (Gredell et al., 2012). In previous studies we engineered variants of the *Escherichia coli* regulatory protein AraC, natively induced by L-arabinose (L-ara), to instead specifically activate gene expression in response to D-arabinose (Tang et al., 2008), mevalonate (Tang and Cirino, 2011), and triacetic acid lactone (TAL) (Tang et al., 2013).

TAL (4-hydroxy-6-methyl-2-pyrone) and other 2-pyrone lactones are derailment products of polyketide synthases (PKSs) and serve as precursors to many higher value products (Chia et al., 2012); hence, a sensitive and specific TAL biosensor would be of value in optimizing polyketide producing strains. In a previous study, we isolated our TAL-responsive AraC variant by screening a combinatorial AraC library constructed by simultaneously randomizing five codons corresponding to five residues (P8, T24, R38, Y82, and H93) located within the AraC ligand binding domain (LBD) (a library of ~34 million variants). This AraC library was expressed in *E. coli* and TAL-induced expression of GFP from the P_{BAD} promoter was screened via multiple rounds of fluorescence-activated cell sorting (FACS), resulting in isolation of a

single TAL-responsive variant, "AraC-TAL1." To our knowledge, no natural or other artificial TRPs responding to TAL or similar 2-pyrone lactones have been identified.

Selection of the five residue positions for mutagenesis was based on prior structural and mutational analyses. Crystal structures of the wild-type AraC (wt-AraC) LBD in the absence of and in complex with L-ara were previously solved. The L-ara complexed structure revealed primary contacts between a single L-ara molecule in the LBD and residues P8, T24, R38, Y82, and H93, as well as several other residues indirectly interacting with L-ara through water-bridged hydrogen bonds (Soisson et al., 1997). In addition, substantial conformational changes in the wt-AraC N-terminal arm (NTA; residues 1-20) upon ligand binding were observed (Reed and Schleif, 1999; Saviola et al., 1998). Substitutions at residue F15 dramatically affect the response to L-ara, resulting in constitutive and non-inducible AraC variants. Residues P8 and L9 are believed to contribute the strongest individual interaction energy between the NTA and L-ara(Damjanovic et al., 2013). Substitutions examined at residues 6-18 largely resulted in variants with loss of repressibility (i.e. constitutive), whereas substitutions at residues T24, R38, H80, and Y82 led to repressible but non-inducible variants (Ross et al., 2003).

With the goal of designing AraC-TAL variants that respond specifically to 2-pyrone lactones of interest (e.g., a compound reflecting altered starter- or extender-unit specificity of a PKS variant), here we aim to gain insights into molecular recognition by AraC-TAL1, and variants thereof. From additional screening of a library of AraC variants using alternate protocols, we describe the isolation and characterization of a variety of new AraC-TAL variants (each having four to five amino acid substitutions), from which patterns of amino acid substitutions were observed. Since single amino acid substitutions can dramatically alter the behavior of wt-AraC, we examined the individual and combined contributions of amino acid substitutions in AraC-TAL1 gene expression control to determine if this variant would be subject to a similar level of rigidity. Finally, we solved the AraC-TAL1 LBD structure by X-ray crystallography to

gain further insights into the sequence-to-function relationships that may help guide further design and screening efforts to identify TRPs for new targets of interest.

III.2) RESULTS

3.2.1) Isolation and analysis of new AraC-TAL clones

AraC-TAL1 was isolated after 11 rounds of FACS sorting, and during those sorts cells were induced by TAL until late-stationary phase prior to sorting (Tang et al., 2013). Subsequent to this study, we optimized our AraC library screening protocol for isolating new variants responding to various small molecules (Chapter IV). The new protocol includes: enriching FACS endpoint populations using selections and screening in microtiter plate assays after fewer rounds of sorting, screening cells after shorter growth periods in the presence of the desired inducer ligand, and optimized cell recovery and media/growth conditions. For the case of TAL as the inducer, we discover that different sorting strategies lead to the isolation of different TAL-responsive variants, which we discuss below. Here we describe nine new AraC-TAL variants isolated from different sorting strategies but the same AraC library as AraC-TAL1, containing NNS (N = A,T,G, C and S = G, C) sites at codon positions relative to residues P8, T24, H80, Y82, and H93 (SLib4) (Tang and Cirino, 2011). Library screening was based on green fluorescent protein (GFP) expression controlled by the AraC cognate promoter P_{BAD} (P_{BAD}-*gfpuv*). Our optimized screening protocols and FACS were used to screen the library as described below.

After five rounds of sorting, two distinct populations (endpoints EP1 and EP2) emerge from different sort paths, each showing enhanced expression of GFP in the presence of TAL (Figure III-1). From these endpoints, we discovered three unique TAL-responsive variants previously not isolated (AraC-TAL2-4). Interestingly, the original AraC-TAL1 was not found in either endpoint population, despite having a similar response to the newly isolated clones. Only 4 out of 48 clones screened from EP1 and EP2 showed a response to 5 mM TAL. Owing to this,

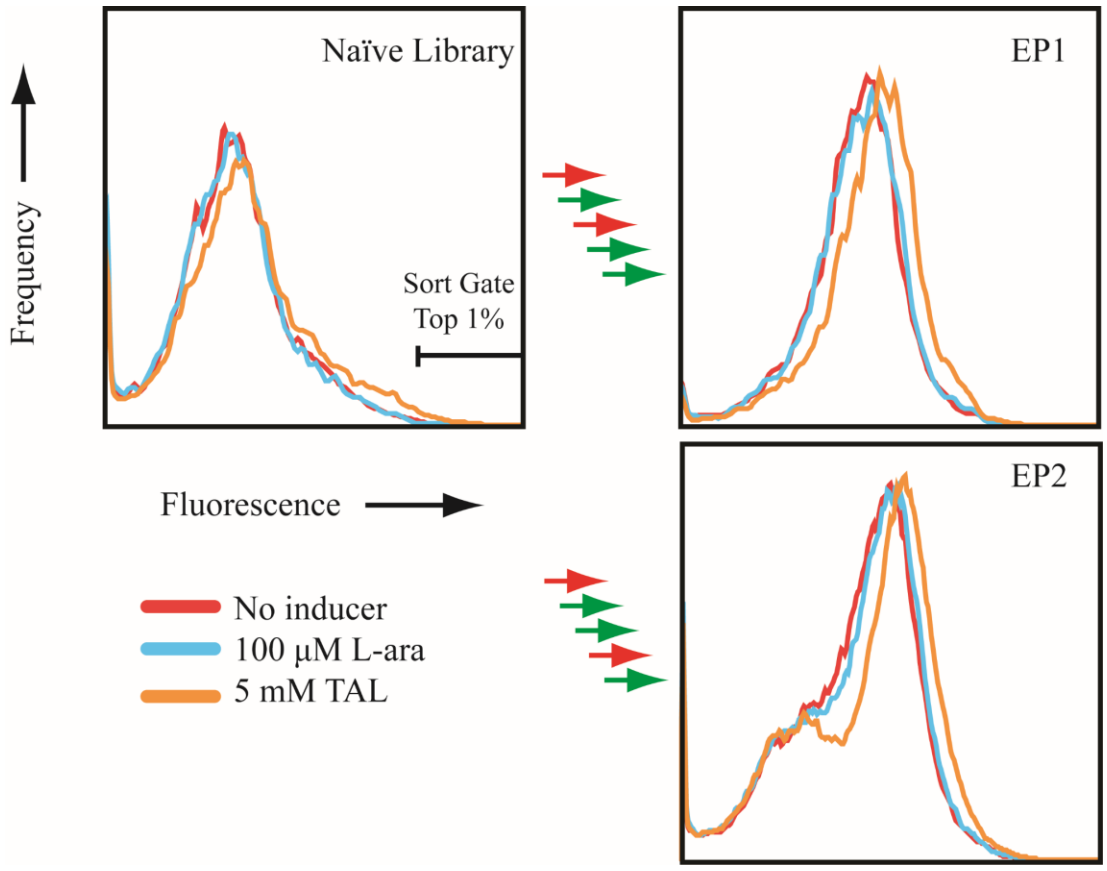


Figure III-1 Histograms of flow cytometry data from endpoint populations after five rounds of FACS screening. The naïve library was sorted using two different sort schemes. (green arrows, positive sort; red arrows, negative sort) and led to endpoint populations EP1 and EP2. Fluorescence is reported on a log scale.

we reasoned that these endpoint populations still retained high levels of sequence diversity and the populations required further enrichment to enhance the frequency of responding clones. To address this, we incorporated a single round of selection after isolating the endpoint populations. P_{BAD}-*bla* (β-lactamase) was integrated into the chromosome of HF19 and confers resistance to ampicillin upon AraC-mediated activation. The resulting populations after selection led to the discovery of six additional unique AraC-TAL variants (AraC-TAL5-10) and the isolation of the original AraC-TAL1. The amino acid substitutions of each AraC-TAL variant are reported in Table III-1. Further optimization strategies of AraC library screening, including strategic placement of selection steps, media optimization, and gene copy number are the topic of a forthcoming manuscript.

Table III-1 Residue substitutions of isolated AraC-TAL variants. The respective codons are also reported. AraC-TAL variant 1 is the original AraC-TAL. AraC-TAL2-4 were isolated form the new screen. AraC-TAL5-10 were isolated after a final round of selection.

	Codon					Residue				
Clone	8	24	80	82	93	8	24	80	82	93
WT-AraC	CCC	ACG	CAT	TAC	CAC	P	T	H	Y	H
AraC-TAL1	GTG	ATC	GGC	TTG	CGC	V	I	G	L	R
AraC-TAL2	GGG	CAC	CAC	AAG	CTG	G	H	H	K	L
AraC-TAL3	TCC	ATC	GGC	ATC	AGG	S	I	G	I	R
AraC-TAL4	AGC	CTG	GGC	CTC	CGC	S	L	G	L	R
AraC-TAL5	ATC	TTG	GGC	ATC	CGG	I	L	G	I	R
AraC-TAL6	GGG	TTG	CAC	AAG	GTC	G	L	H	K	V
AraC-TAL7	GTG	CTC	GGC	CTC	CGC	V	L	G	L	R
AraC-TAL8	GGG	CTG	CAC	AAG	TTC	G	L	H	K	F
AraC-TAL9	ACG	ATC	GGG	CTC	CGG	T	I	G	L	R
AraC-TAL10	GGC	CTG	GGC	ATC	CGC	G	L	G	I	R

We were curious as to why these new TAL-responsive clones (AraC-TAL2-10) were not isolated previously. Lower affinity for TAL and/or reduced activation may have led to these clones being discarded under our previous stringent sort conditions. Therefore, we investigated the dose-dependent responses of all AraC-TAL variants (Supp. Figure C-1). Above 25 mM TAL, cell growth is dramatically inhibited, preventing the measurement of a saturated response for most clones. However, each variant showed a dynamic range of response over 1-20 mM TAL and less than 2-fold response in the presence of up to 100 mM L-ara (data not shown). Fluorescence (GFP expression) in the presence of TAL (5 mM) and uninduced background fluorescence are reported in Table III-2. The uninduced background fluorescence ("leakiness") of all variants is significantly greater than wt-AraC but similar to the previously isolated AraC-TAL1. AraC-TAL9 has a single amino acid substitution compared to AraC-TAL1 (V8T) and shows higher leakiness than all other variants. Induced fluorescence with AraC-TAL9 in the presence of 5 mM TAL was proportionally higher, leading to a fold-response similar to those of the other AraC-TAL variants. This result is consistent with earlier findings, suggesting that

Table III-2 AraC-TAL variant responses to various treatments. The fluorescence per OD_{595} is reported for each clone in the absence of any ligand ("Background") and 5 mM TAL. The data was collected from three independent experiments and the averages are reported. The standard deviations were less than 20% of the average unless otherwise indicated. The fold-response of each clone in the presence of each ligand is reported as the fluorescence in the presence of the ligand divided by the background fluorescence.

Clone	Background (leakiness)	5 mM TAL	TAL Fold Response	$K_{d,app}$ [a] (mM)	$Max_{1/2}$ [b] (mM)
WT-AraC	29 ±6	22 ±11	0.8	-	-
AraC-TAL1	110	2000	18.2	-	16.1
AraC-TAL2	130	1300	10.0	-	17.6
AraC-TAL3	140	1900	13.6	-	18.7
AraC-TAL4	130	1800	13.8	17.6	16.1
AraC-TAL5	140	2000 ±500	14.3	16.5	14.9
AraC-TAL6	130	1100	8.5	-	20.1
AraC-TAL7	90	1200 ±260	13.3	-	17.6
AraC-TAL8	150 ±30	1500	10.0	-	16.6
AraC-TAL9	260	4100	15.8	12.9	9.1
AraC-TAL10	80	1900	24	8.9	10.0

a-$K_{d,app}$ was only calculated for samples with complete dose response curve before reaching the toxicity limit of TAL in the media

b-$Max_{1/2}$ was determined by finding the maximum fold-response for each clone and extrapolating the TAL concentration at half the maximum fold-response

substitutions at P8 in the NTA more strongly influence repression than response to inducer (Ross et al., 2003; Saviola et al., 1998; Tang and Cirino, 2010). Due to the similarities in TAL response for all variants, we believe the individual responses do not explain why these variants were not previously isolated.

Despite their similar responses to TAL, two patterns of amino acid substitutions are present among the AraC-TAL variants. AraC-TAL variants 1, 3, 4, 5, 7, 9, and 10 show highly conserved amino acid sequences among the substituted residues: T24I or T24L, H80G, Y82I or Y82L, and H93R. Among these variants, AraC-TAL10 shows the highest fold-response which is associated to its low background fluorescence in the absence of TAL. A second pattern emerged from the AraC-TAL variants (AraC-TAL2, 6, and 8) and shows weaker responses to TAL than those with the first pattern. These variants all contain: i) at least two positively charged amino

acid substitutions (all others only contain one; Supp. Table C-2); ii) have the least changes in substituted residue hydrophobicity (Supp. Table C-2); and iii) were the only variants that do not include the substitution H93R. TAL is negatively charged at neutral pH (deprotonated at the 4-hydroxyl), and the positively charged substitution(s) may directly interact with the hydroxyl group. Meanwhile, substitutions with more hydrophobic amino acids should promote stronger interactions with the lactone ring and methyl group of TAL, as compared to the more polar pyranose ring of L-ara. A direct correlation was seen between the increase in amino acid substitution hydrophobicity and response to TAL (Supp. Figure C-2), which could indicate a less specific response to molecules with hydrophobic functional groups.

In addition to TAL-responses, we looked into the specificity of each variant by measuring their response to structurally similar compounds phloroglucinol and 2,6-dimethyl-4-pyrone (Supp. Figure C-3). No response to these compounds was detected (up to 25 mM for both). These results suggest selectivity towards TAL, though additional screening with TAL analogues will provide further insights into specificity and suggestions for decoy compounds in counter screens to evolve AraC variants that respond specifically to TAL or TAL analogues. Finally, it should be noted that the dose-dependent response of AraC-TAL1 to TAL was found to be unaffected by the presence of up to 10 mM L-ara (Supp. Figure C-4). This suggests L-ara does not bind in the ligand binding pocket, as opposed to it binding and not activating transcription.

3.2.2) Amino acid substitutions in AraC-TAL variants reveal mostly cooperative interactions

Sequence analysis of AraC orthologues indicates that the amino acid substitution patterns of the LBD in AraC-TAL variants are important to both ligand binding and the *on/off* switch (Damjanovic et al., 2013; Schleif, 2010; Soisson et al., 1997). To better understand the roles of

Table III-3 Substitution analysis of the targeted AraC ligand binding pocket residues. The 32 clones represent all combinations of residue substitutions between wt-AraC andAraC-TAL1. The fluorescence, measured in relative fluorescence units per OD_{595}, is reported for each clone. The data was collected from three independent experiments and the averages are reported and standard deviations are less than 20% of the average unless otherwise indicated. The fold-increased fluorescence response of each clone in the presence of ligand is reported as the fluorescence in the presence of the ligand divided by the background fluorescence.

						Fluoresence (rfu/OD)			Induction Fold	
	Residue					Background	500 µM	7 mM	500 µM	7 mM
Clone	8	24	80	82	93	(leaky)	L-ara	TAL	L-ara	TAL
WT-AraC	P	T	H	Y	H	72	20500 ±5200	77 ±16	280	1.1
Mut10000	V	T	H	Y	H	300	13700	330 ±90	46	1.1
Mut01000	P	I	H	Y	H	100 ±22	73 ±16	100	0.7	1.0
Mut00100	P	T	G	Y	H	150	81	110	0.5	0.7
Mut00010	P	T	H	L	H	100	73	120 ±27	0.7	1.2
Mut00001	P	T	H	Y	R	110	84	120 ±28	0.8	1.1
Mut11000	V	I	H	Y	H	750	610 ±170	740	0.8	1.0
Mut10100	V	T	G	Y	H	540 ±140	530	600 ±130	1.0	1.1
Mut10010	V	T	H	L	H	160	180	200	1.1	1.3
Mut10001	V	T	H	Y	R	450	450 ±110	550	1.0	1.2
Mut01100	P	I	G	Y	H	90 ±20	68	110 ±30	0.8	1.2
Mut01010	P	I	H	L	H	100	76	120	0.8	1.2
Mut01001	P	I	H	Y	R	700	610	750	0.9	1.1
Mut00110	P	T	G	L	H	260	170	240 ±60	0.7	0.9
Mut00101	P	T	G	Y	R	160	120	180 ±50	0.8	1.1
Mut00011	P	T	H	L	R	120	91	130 ±30	0.8	1.1
Mut11100	V	I	G	Y	H	1400	1300	1200	0.9	0.9
Mut11010	V	I	H	L	H	320	240	330	0.8	1.0
Mut11001	V	I	H	Y	R	280	240	290	0.9	1.0
Mut10110	V	T	G	L	H	1500	1400	1800 ±450	0.9	1.2
Mut10101	V	T	G	Y	R	400	340	510	0.9	1.3
Mut10011	V	T	H	L	R	300 ±80	240	360	0.8	1.2
Mut01110	P	I	G	L	H	190	130	180	0.7	0.9
Mut01101	P	I	G	Y	R	100 ±20	68 ±14	110	0.7	1.1
Mut01011	P	I	H	L	R	130	93	130 ±30	0.7	1.0
Mut00111	P	T	G	L	R	170	150	230	0.9	1.4
Mut11110	V	I	G	L	H	680	670 ±150	820	1.0	1.2
Mut11101	V	I	G	Y	R	140	96	160 ±30	0.7	1.1
Mut11011	V	I	H	L	R	280	240	340	0.9	1.2
Mut10111	V	T	G	L	R	300	270 ±90	1500	0.9	5.0
Mut01111	P	I	G	L	R	140	97	400	0.7	2.9
AraC-TAL	V	I	G	L	R	380	420	5900	1.1	15.5

the five AraC-TAL1 amino acid substitutions and assess their potential cooperative effects, we investigated the TAL- and L-ara responses of 32 AraC variants representing all combinations of wt-AraC or AraC-TAL1 residues, at the five target residues (Table III-3). Variants are labeled according to the five residue positions carrying the wt-AraC ("0") or AraC-TAL1 ("1") amino acid. For example, "Mut10000" indicates that the wt-AraC residue at position P8 was changed to the AraC-TAL residue (P8V). Other than wt-AraC and AraC-TAL1, only three variants retain partial responses (>15% of wt-AraC or AraC-TAL to L-ara or TAL, respectively). Such intolerance to single substitutions in AraC-TAL1 points to cooperative interactions among these amino acids toward the gene expression response.

Weakened interactions between the AraC NTA and adjacent DNA-binding domain is expected to weaken gene repression at P_{BAD} (Rodgers and Schleif, 2009; Ross et al., 2003; Saviola et al., 1998). Variants with elevated leakiness resulting from the P8V substitution was therefore not unexpected. The induced expression response to L-ara is also much less affected by substitution P8V (Mut10000) compared to all other single substitutions (Mut01000, Mut00100, Mut00010, and Mut00001). A similar effect with a single substitution P8R was previously noted (Tang and Cirino, 2010). The AraC-TAL1 variant with no substitution at P8 (Mut01111) also shows significantly reduced background fluorescence compared to AraC-TAL1, along with a dramatic, though not complete, loss in induced response to TAL. Mut10111, AraC-TAL variant with a single wt-AraC substitution at residue 24, is the only other variant retaining a substantial response to TAL (> 15% of AraC-TAL1 response).

In a similar analysis we created and tested alanine-substitution variants of AraC-TAL1 to determine the contribution of each residue relative to a comparatively inert and small amino acid (rather than that of native wt-AraC amino acids). As shown in Table III-4, substitution V8A retains a significant response to TAL, again supporting a stronger role of this residue in repression and arm switching, compared to ligand recognition. Alanine substitution at residue 24

Table III-4 Response of AraC-TAL1 variants with single alanine substitutions. Fluorescence per OD595 is reported for each clone. The fold-induced fluorescence response of each clone in the presence of ligand is reported as the fluorescence in the presence of ligand divided by the background fluorescence. The data was collected from three independent experiments and the averages are reported and standard deviations were less than 20% of the average unless otherwise indicated.

	Fluorescence			Induction Fold	
	Background (leaky)	500 µM L-ara	7 mM TAL	500 µM L-ara	7 mM TAL
AraC-TAL	400	330	4800	0.8	12.0
V8A	140	100	860	0.7	6.1
I24A	890	860	1700 ±490	1.0	1.9
G80A	200	160	490	0.8	2.5
L82A	160	100	170	0.6	1.1
R93A	1800 ±370	1400	1700	0.8	0.9

also retains some response to TAL. Interestingly, variant G80A also shows response to TAL, while Mut11011 (AraC-TAL1 with Histidine at position 80) shows none. The larger histidine might crowd the binding pocket and exclude ligand binding. Finally, L82A and R93A show no response to TAL.

3.2.3) X-ray crystal structure of AraC-TAL1 LBD reveals the similarities in the ligand binding pocket

Structural determination of AraC-TAL1 LBD in complex with TAL using X-ray crystallography was sought to illuminate details of the ligand-protein interactions. Conditions supporting crystal growth in the presence of TAL were not found. We were however able to solve the apo AraC-TAL1 LBD structure at a resolution of 2.6 Å. There are three monomers in the asymmetric unit and the electron density is well defined and continuous for residues 17 through 168. The resulting crystal structure is shown in Figure III-2A overlaid with apo wt-AraC. Not surprisingly, the structure of the NTA (residues 1 to18) was not completely resolved, which plays a crucial role in the transcriptional regulation of AraC (Reed and Schleif, 1999; Rodgers and Schleif, 2009; Ross et al., 2003; Saviola et al., 1998; Seabold and Schleif, 1998).

The crystal structure of apo AraC-TAL1 is similar to those of apo and holo wt-AraC, with a root-mean-square deviation (RMSD) of 0.93 Å and 0.85 Å, respectively. The similarities to both the apo and holo structures of wt-AraC were not surprising due to the low RMSD (0.63 Å) between the two wild-type structures, differing significantly in only the NTA position. Still, the apo form of AraC-TAL1 is slightly better aligned with the holo form of wt-AraC. This could indicate that AraC-TAL1 is in a partially activated state and would explain the leakiness seen by all AraC-TAL variants. The resolved substituted residues, T24I, Y82L, and H93R (P8V is part of

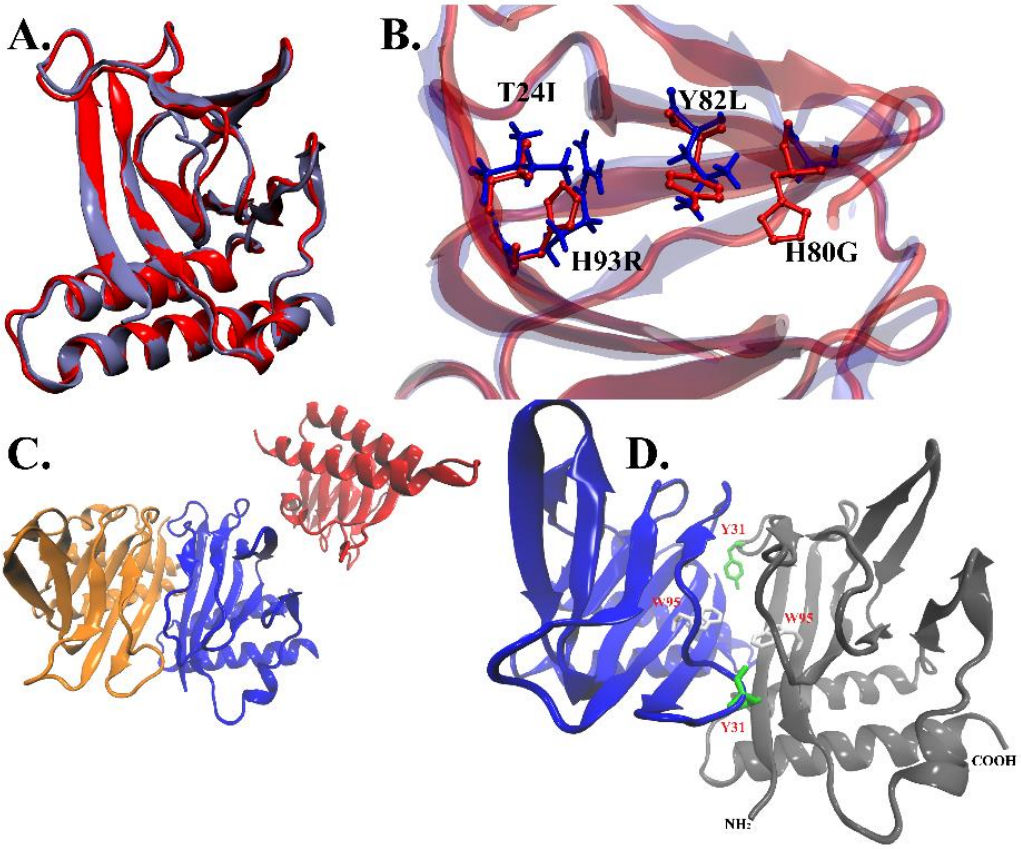


Figure III-2 Comparison of AraC-TAL1 crystal structure with wt-AraC. (A.) Overlay of the apo structures of wt-AraC (red) and AraC-TAL1 (blue). (B.) The substituted residues of AraC-TAL1 (blue) are oriented similarly to the native residues of wt-AraC (red). (C.) Each asymmetric unit of the AraC-TAL1 crystal structure contained three monomers. (D.) The β-kiss of the two monomers. Residues Y31 and W95 are highlighted.

the NTA and was not in an ordered region of the structure) protrude into the ligand pocket (Figure III-2B) and alter the binding pocket properties with minimal changes to the backbone positions. Substitution H80G does however shift the position of beta sheet β2 by 2.5 Å. Also

despite the substitution of more hydrophobic amino acids (T24I, H80G, and Y82L), the ratio of solvent accessible surface area of hydrophobic and non-hydrophobic residues in the ligand binding domain remains relatively unchanged (1.284 and 1.281, respectively).

Two of the three monomer chains in the asymmetric unit of the LBD formed a dimer through interactions between the N-terminal β-barrels (Figure III-2C and 2D). wt-AraC also exhibits a dimer through a "β-kiss" interface but only in the apo form, where a tyrosine (Y31) of the adjacent monomer fills the ligand binding pocket and interacts with W95 (Soisson et al., 1997; Weldon et al., 2007). However, the β-kiss of AraC-TAL1 is slightly rotated relative to the apo wt-AraC β-kiss, which prevents Y31 from filling the ligand pocket and interacting with W95. Overall the structure of apo-AraC-TAL1 suggests the regulatory mechanism is similar to that of wt-AraC but with a modified binding pocket that accepts different substrates. Resolving the NTA and structure in the presence of TAL remains an important factor in understanding the mechanism of AraC-TAL variants.

III.3) DISCUSSION

AraC tightly regulates gene expression at promoter P_{BAD}, and the AraC-P_{BAD} regulatory system is an invaluable tool in applied molecular biotechnology and metabolic engineering. By altering AraC effector specificity we developed this system to act as a reporter of TAL. Whereas attempts to isolate AraC variants with altered effector specificity from randomly mutated libraries were unsuccessful (Tang et al., 2008), simultaneously targeting multiple positions within the binding pocket yielded several functional biosensor variants. In this study, we describe a new set of TAL-responsive AraC variants that were isolated as a result of a modified library screening protocol. Though a single AraC variant was previously isolated with response to TAL using FACS alone, inadequate endpoint population screening or over-screening of the combinatorial library prevented identification of other variants with similar responses to TAL. From the new

variants discovered here, two distinct patterns emerged among the five amino acid substitutions. Both patterns contained amino acid substitutions with positively charged side chains, which could form electrostatic interactions with the C4-hydroxyl group of TAL, which is deprotonated under physiological conditions. These results led us to probe the contributions of individual residues toward the TAL-dependent gene expression response, via amino acid substitution analyses. The absence of response from the majority of the AraC-TAL variants substituted with wt-AraC amino acids or alanine highlights the importance of a semi-rational design approach targeting multiple residues simultaneously.

While it is not known to what extent the various substitutions affect each variant's stability, overall fold, or solubility, our results collectively demonstrate non-additive and cooperatively acting amino acid substitutions within a given variant. The crystal structure of an apo AraC-TAL LBD variant is also shown to be nearly identical to that of wt-AraC. However, solving the crystal structure of AraC-TAL LBD in complex with TAL remains a work in progress. Based on the structure of wt-AraC LBD in complex with L-arabinose, we expect the NTA to be folded over the ligand binding pocket in the AraC-TAL1 holo complex. This structure would help confirm the role of substitutions at position P8, which seem to be primarily involved with regulating repression. These results are consistent with previous studies indicating substitutions in the NTA weaken repression in the absence of a ligand (Cox et al., 2002; Reed and Schleif, 1999; Saviola et al., 1998). Understanding the orientation of TAL within the binding pocket will also help to understand molecular recognition and how to better design for selectivity. Notably, molecular docking studies with the current apo-structure of the AraC-TAL1 LBD have been inconclusive, in that many potential TAL orientations show similar binding energies and are within the error of the energy calculations in the docking protocol (results not shown). This is not unexpected, given the relatively low sensitivity of all AraC-TAL variants to inducer TAL. However, the amino acid sequence-function data sets provided here are useful for training and

validating future AraC modeling and ligand docking studies, which in turn should help guide rational design approaches to fine-tune specificity.

We aim to ultimately develop new AraC variants that respond specifically to various TAL analogues. Such variants can serve as biosensors to report on altered substrate specificity of select polyketide synthase variants. Results from this study will help to guide the design of this next generation of biosensors. The discovery of different AraC variants depending on the screening protocol demonstrates the importance of parallel screening and placement of selection steps in combination with FACS (the topic of a Chapter IV). The structural similarities between wt-AraC and AraC-TAL1 indicate tolerance to amino acid substitutions in the ligand pocket and encourage exploration of additional residue substitutions for improved biosensors.

IV Rapidly evolved AraC-based biosensor for vanillin and salicylic acid from combinatorial libraries using enhanced ligand-induced combination screening

IV.1) INTRODUCTION

Directed evolution is a valuable approach for synthetic biology and has successfully demonstrated applications in engineering proteins for improved binding affinity, specificity, catalytic efficiency, toxicity tolerance, thermostability, and expression (Bornscheuer et al., 2012; Denard et al., 2015; Wang et al., 2012; Yoshida et al., 2014). (Bornscheuer et al., 2012; Denard et al., 2015; Wang et al., 2012; Yoshida et al., 2014). In a typical whole-cell directed evolution experiment, there are four primary steps: (i) identification of limiting reactions; (ii) genetic library construction; (iii) expression in a host microorganism; and (iv) screening for variants with improved properties. Variants with the most promising performance are isolated and can be subjected to additional rounds of directed evolution if the desired function has not been achieved. Despite the successes of directed evolution approaches, it is bottlenecked by the lack of adaptable and readily available genotype-phenotype linkages and high-throughput screening (HTS) methods that balance both throughput and purity.

Transcriptional regulatory proteins (TRPs) are emerging as powerful tools for interpreting intracellular metabolite concentrations via activation of a phenotypic response upon recognition of the specific metabolite (Dietrich et al., 2010; Goodey and Benkovic, 2008; Gredell et al., 2012; Michener et al., 2012; Zhang and Keasling, 2011). Unfortunately, a universal allosteric TRP does not exist for detecting various compounds, nor is there one available for every compound. Thus, there is a need to engineer TRP-based biosensors. Albeit, success has been limited with

engineering molecular biosensors from TRPs due to the lack of a universal TRP and the current state of available screening techniques.

The *Escherichia coli* native transcriptional regulator AraC has high selectivity and sensitivity to its native ligand, L-ara. AraC represses in its native state ("OFF") and activates transcription ("ON") from its cognate promoter, $P_{BAD,}$ in the presence of L-arabinose (L-ara). Schleif and coworkers extensively studied the structure of AraC and the roles of various residues for the function of its "light switch mechanism" (Lobell and Schleif, 1990; Ross et al., 2003; Soisson et al., 1997). Using combinatorial design of AraC and fluorescence-activated cell sorting, we have developed endogenous molecular reporters for D-arabinose (D-ara) (Tang et al., 2008), triacetic acid lactone (TAL) (Tang et al., 2013), and mevalonic acid (mev) (Tang and Cirino, 2011). The natural behavior of AraC requires both transcriptional functionality and target recognition. This property is different from screening of combinatorial libraries of enzymes, wherein enzyme screening is conceptually a "Yes" or "No" output, requiring enrichment of the positive population (baring any desirable specificity). However, a TRP may possess the ability to retain functional activation absent a ligand (i.e., loss of repressibility). Therefore, the positive population may be riddled with false positives and continuous enrichment of the most "On" samples may lead to all constitutively expressing clones. Previously, alternating positive and negative fluorescence-activated cell sorting (FACS) was employed to reject false-positives and enrich only functional variants. The native ligand, L-ara, was used as a decoy ligand for the negative sorted populations to increase the stringency of the sort towards variants with at least minimal selectivity against L-ara. As is shown in Chapter II, various sorting schemes lead to functional low-responding clones, but the previously reported isolated D-ara and mevalonate variants were not isolated from any of the sorting schemes. Through next generation sequencing, results showed a high level of diversity carried over between rounds of FACS despite the stringency of the sort (Appendix E). Therefore in order to explore a broader sequence space of

AraC variants for recognition of target compounds, we aim to improve the throughput and efficiency of previous screening strategies.

Here we present a high-throughput method for screening large libraries of TRP-based biosensors, specifically focusing on AraC, by combining both FACS and selection. We investigate several strategies to enhance the signal-to-background ratio (fold-response), including biosensor expression and plasmid copy number. By improving the fold response, the quality of the assay increases, making it more applicable to HTS. The isolation of multiple AraC variants with novel altered specificity from the combined screening method and enhanced fold-response demonstrates the plasticity of a single TRP to recognize various small molecules. The mutational exploration of additional TRPs could help provide a platform for whole-cell biosensor design and alleviate key limitations of directed evolution of enzymes for metabolic engineering.

IV.2) Results and Discussion

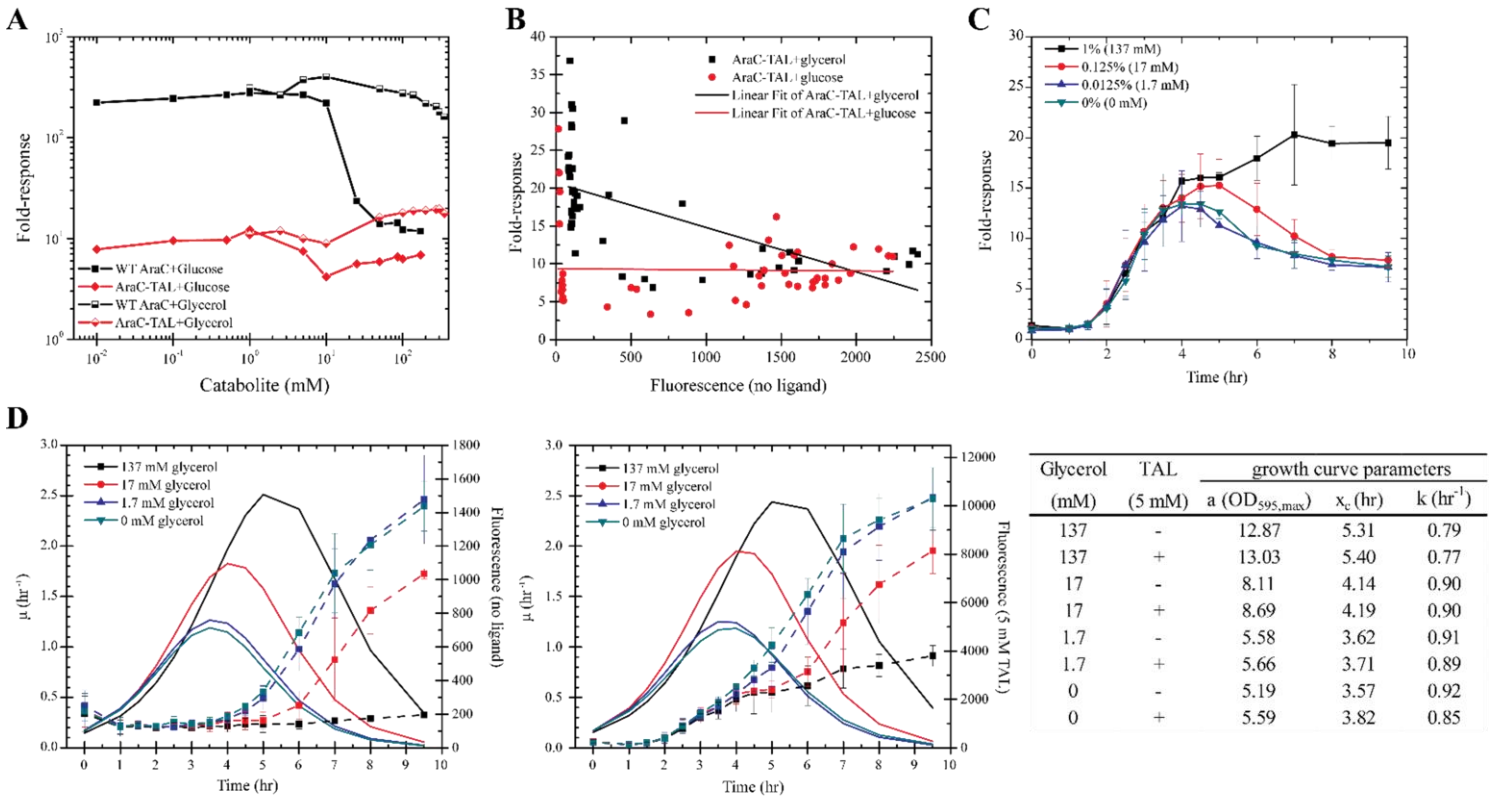
4.2.1) Enhanced response by catabolite repression

We previously determined that the background fluorescence (i.e., "leakiness") of various engineered biosensors exceeded wt-AraC (Tang and Cirino, 2010; Tang et al., 2008), despite the incorporation of stringent negative screens to discard leaky clones. High leakiness will potentially lead to a decrease in signal-to-background (fold-response) for isolated clones. Several methods have been used to enhance transcriptional regulation including promoter optimization and construction and screening of DNA binding domain libraries (Alper et al., 2005; Chusacultanachai et al., 1999).

The natural diauxic growth of *E. coli* presents a powerful means for repressing CRP-regulated gene expression. Naturally, cAMP receptor protein (CRP)-regulated promoters (i.e., P_{BAD}) depend on the phosphorylation of the glucose specific enzyme II (EII^{Crr}) and activation of cyclic adenosine monophosphate (cAMP) intracellular production, which binds to CRP to

activate the respective metabolic pathways for non-glucose carbon source metabolism related to the phosphotransferase system (Neidhardt et al., 1987). Many CRP-regulated promoters are co-regulated by additional TRPs (Martínez-Gómez et al., 2012), including AraC/L-ara regulation of the P_{BAD} promoter (Ogden et al., 1980). During catabolite metabolism (i.e., glucose) cAMP production is inhibited, thus preventing expression of CRP positive regulated genes. Glucose has been shown to rapidly and adversely affect expression of genes from the P_{BAD} promoter (Guzman et al., 1995).

Here, we explored the effects of glucose and glycerol on the AraC-based biosensor to determine if a better fold-response could be achieved by decreasing the background fluorescence (P_{BAD} is dual regulated by AraC and the cAMP receptor protein (CRP)). Both wt-AraC and previously engineered AraC-TAL1 biosensors were grown in a rich medium containing either 0 to350 mM glycerol or 0 to 171 mM glucose. Upon reaching stationary growth phase, the background (FL_B) and induced (FL_I) fluorescence (100 µM L-ara or 5 mM TAL, respectively) were measured. The average values from three independent experiments were used to calculate the fold-response (FL_I/FL_B). The AraC-TAL1 background decreases by approximately 2-fold in the presence of high concentrations of glucose. However, glucose elicits a decrease in fold-response for both biosensors (Figure 1A). In contrast, the presence of 100 to 300 mM of glycerol decreases the background leading to an increase in the fold-response for AraC-TAL1. wt-AraC fold-response is unaffected over this range of glycerol. We suggest this occurs due to the strong repression and low background of wt-AraC being negligibly affected by the presence of the catabolite. From this data, we elected to use 137 mM glycerol (1% v/v) in all culture media for biosensor response, which corresponded with the range in which the maximal fold-response was observed for AraC-TAL1 and wt-AraC was unaffected.



Glycerol (mM)	TAL (5 mM)	growth curve parameters a ($OD_{595,max}$)	x_c (hr)	k (hr^{-1})
137	-	12.87	5.31	0.79
137	+	13.03	5.40	0.77
17	-	8.11	4.14	0.90
17	+	8.69	4.19	0.90
1.7	-	5.58	3.62	0.91
1.7	+	5.66	3.71	0.89
0	-	5.19	3.57	0.92
0	+	5.59	3.82	0.85

Figure IV-1 Effects of glycerol and glucose on the AraC-based biosensor response. (A) The fold-response of the wt-AraC and AraC-TAL were compared over a range glycerol and glucose concentrations. (B) All individual data points from the three independent experiments of (A) were plotted for their fold-response relative to the fluorescence in the absence of ligand. A linear regression line was fitted to each set of data to determine the correlation between fold-response and background fluorescence. (C) The effect of glycerol on the fold-response of the AraC-TAL biosensor over time.(D) The effect of a range of glycerol on the growth rate (left axis and solid lines) and the background fluorescence (right axis and dotted lines) of the AraC-TAL biosensor in response to 5 mM TAL was plotted. The growth rates were determined by fitting the average measured growth curves from three independent experiments with a sigmoidal curve and solving the first derivative. The parameters to the fitted curves are listed in the table.

Bettenbrock and coworkers demonstrated a negative correlation between the growth rate (μ) and cAMP intracellular levels (Bettenbrock et al., 2007). They found that in the presence of glucose as compared to other carbon sources, the growth rates of *E. coli* strains were elevated and the $EIIA^{Crr}$ phosphorylation level was minimal along with cAMP synthesis (the phosphorylated form of EII^{Crr} activates adenylate cyclase). However, *E. coli* in the presence of glycerol, an energy-poor carbon source, showed a reduced growth rate compared to growth with glucose which consequently increased the level of $EIIA^{Crr}$ phosphorylation and cAMP concentrations. Other experimental and computational models also support the negative relationship between growth rates and expression of heterologous proteins (Alper et al., 2005; Bienick et al., 2014; Scott et al., 2010). To determine if the growth rate affects the expression of GFPuv, we measured the growth curves of a strain expressing the AraC-TAL1 biosensor in the presence of different concentrations of glycerol with and without 5 mM TAL. The apparent growth rate (μ_{app}) was plotted against time in Figure IV-1C and overlaid with the respective relative fluorescence over time. The presence of 5 mM TAL did not sufficiently affect any of the growth curve parameters (Figure IV-1D). Thus, we assume there is no undesirable amplification of fluorescence from slower growth in the presence of TAL. An appreciable increase in fluorescence occurred shortly after reaching the maximum growth rate, but increasing glycerol concentrations helped maintain a low background fluorescence. A similar affect was seen for the cultures induced with 5 mM TAL. The fluorescence of samples in the presence of higher concentrations of glycerol (17 mM and 137 mM) appeared to have reached a plateau, but the growth rate slowed and there was a spike in fluorescence. Despite this spike in fluorescence, Figure 1C shows that the fold-response remained relatively constant after reaching the maximum growth rate (late exponential phase/early stationary phase). Whereas, all other concentrations of glycerol lead to a decline in response after the maximum growth rate was achieved. Therefore, all samples for sorting were taken after the cultures had reached late exponential growth phase (approximately 6 hr).

4.2.2) Improved biosensor response using a single plasmid system

Previously, a dual plasmid system was used to express the AraC-based biosensor (P_{BAD}-*gfpuv*, pPCC442; RS1030 modified origin for high copy plasmid (Phillips et al., 2000); P_{tac}-*araC*, pPCC423; pBR322ΔROP medium copy origin). Although the dual plasmid system was helpful in isolating variants with altered specificity, this system was prone to poor growth and difficulties during post-screening variant characterization. Therefore, we sought to use a single plasmid

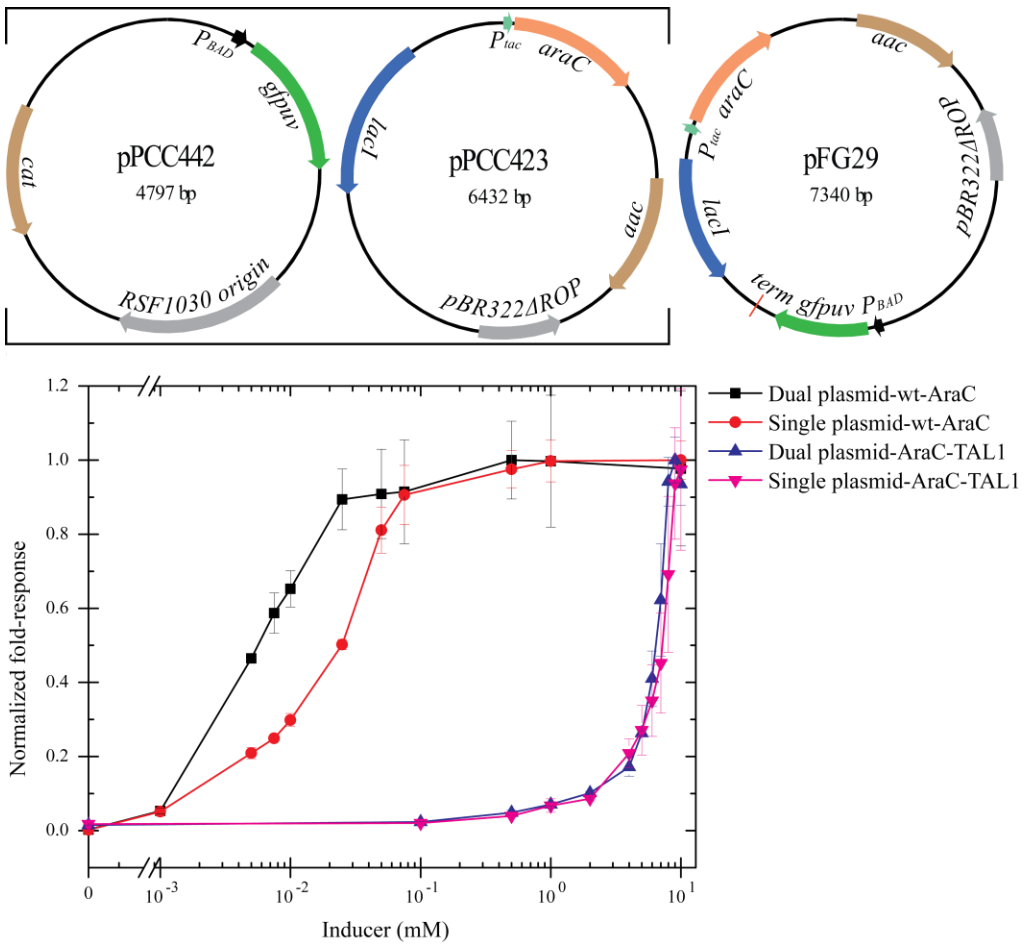


Figure IV-2 Plasmid maps of dual (in parentheses) and single plasmid biosensors. The RSF1030 origin of replication was modified to have a high copy number (Phillips et al., 2000). Aminoglycoside 3-N-acetyltransferase (*aac*) confers resistance to apramycin and chloramphenicol acetyltransferase (cat) confers resistance to chloramphenicol. P_{tac} is the LacI cognate promoter. P_{BAD} is AraC cognate promoter. Terminator (term) sequence downstream of gfpuv was cloned to prevent read through of unwanted open reading frames. Comparison of single and dual plasmid biosensors. Dual plasmid system: black line with square, wt-AraC; blue line with triangle, AraC-TAL1; single plasmid system: red line with circle, wt AraC; pink line with inverted triangle, AraC-TAL1.

system, containing both P_{tac}-*araC* and P_{BAD}-*gfpuv* on the same medium copy plasmid (pFG29; pBR322ΔROP origin). As shown by Figure IV-2, the half maximal signal increases for the single plasmid system. The monoclonal population from the dual plasmid system shows a high coefficient of variance (CV = 97% ± 5%) as measured by background fluorescence, whereas the CV of the single plasmid system was significantly lower (CV = 76% ± 6%, $p < 0.05$). Benefits of the single plasmid system also extend to the growth of the harboring cells. Growth of cells expressing high levels of GFPuv ($OD_{595,GFPuv}$) compared to growth of cells absent GFPuv production ($OD_{595,OFF}$) showed hindrance with the dual plasmid system but not from the single ($OD_{595,GFPuv}/OD_{595,OFF}$ = 0.63 ± 0.07 and 1.01 ± 0.10, respectively).

Finally, the Z'-factor is a widely used statistical tool to evaluate the quality of a screen by comparison between two control population distributions (i.e. means and standard deviations). Here, AraC-TAL1 is used to compare fluorescence of the cells induced with 5 mM TAL and not induced. The results show there is improvement of the single plasmid system (Z'-factor = 0.88) over the dual plasmid system (Z'-factor = 0.37). The dual plasmid system is considered low quality (Z'factor less than 0.5). However, the single plasmid system improves the spread of the distribution statistics and increases the Z'-factor signifying a high quality screen. Our analysis successfully shows that the addition of glycerol to the media and a single plasmid system for biosensor expression improves the AraC-based biosensor system for more efficient screening of combinatorial libraries.

4.2.3) Isolation of AraC variants with altered specificity using combination of FACS and selection

Currently, two of the most effective whole-cell HTS techniques require fitness be linked to cell survival (selection) or concentration of a fluorescent marker and detection using FACS. However, selections are highly qualitative and false-positive prone (Umeno et al., 2005), and

FACS lacks the throughput of selection, requiring several sorts per round of directed evolution. Therefore, there is a strong need for a technique that addresses both the genotype-phenotype linkage and the throughput. Combining multiple screening techniques to screen large libraries allows for greater throughput with fewer rounds of screening and provide a simple and robust method for discarding false-positive variants. Previously, Shultz and coworkers reported using both FACS and selection to isolate an aminoacyl-tRNA synthetase with altered specificity toward desired tyrosine analogs (Santoro et al., 2002). After a single round of positive selection and negative FACS screening, they successfully isolated a variant with selective properties for unnatural amino acids and no natural amino acids, but this method was highly specific for engineering an aminoacyl-tRNA synthetase. Often, enzymes are the target in protein engineering because they catalyze diverse reactions and lead to the production of specific and predictable metabolites. Other groups have also been successful in combining different screening techniques for engineering enzyme (Feldhaus et al., 2003; Santoro et al., 2002)

Previously, we cloned and screened a 5-site saturated library (SLib4; P8, T24, H80, Y82, H93) using the dual plasmid system for TAL and mevalonate biosensors (Tang and Cirino, 2011; Tang et al., 2013), but we wanted to 1) perform screening using the single plasmid system and 2) expand the sequence space screened. The original SLib4 library was cloned into the pFG29 vector (SLib4s) using restriction sites flanking the plasmid DNA encoding for the LBD. Additionally, we cloned a 5-site saturated library (JLib1s; T24, R38, H80, Y82, H93) and an error-prone library (CLib2s; 1.6% nucleotide error rate in LBD; 8.8 mutations/gene) into the pFG29 vector for the following screening strategy. The targeted arm mutation in SLib4s was not targeted in JLib1s due to known correlations with N-terminal arm mutations and decreased repressibility (Saviola et al., 1998). Residue R38 was targeted instead, which directly forms a bidentate interaction with O4 and O5 of L-ara (Soisson et al., 1997). CLib2s was a library of wt-AraC randomly mutated from 11-537 bp (amino acids 4-179), where the average amino acid

mutation rate was 5.9 amino acid substitutions per protein giving a maximum degeneracy of 2.8 x 10^{15} potential protein variants according to the equations provided by Bosley and Ostermeier (Bosley and Ostermeier, 2005). We recognize that the subsequently described experiments for screening TRP-based biosensors does not completely cover the screening of all possible variants with the addition of an error-prone library containing a high frequency of mutations mostly due to the limitations of the transformation efficiency.

The enhanced ligand-induced combination screening (ELICS) protocol described below is outlined in Figure IV-3. Briefly, individual naïve libraries were initially negatively screened using FACS. The resulting populations were pooled and subjected to positive selection in the presence of the target compound. The resulting clones were then screened with two rounds of negative and positive sorting, followed by a final round of positive selection (as was reported in Chapter III). The endpoint populations from the final round of selection were screened in deep-well plates for response to the target effector. If there were no responsive clones, the initial round of selection was repeated with a higher concentration of the target compound. JLib2s, CLib2s, and SLib4s were separately transformed into HF19 electrocompetent cells (total transformants > 10^8) and individually negative sorted using FACS based on their population frequency in a fluorescence gate (530/40 nm channel) relative to 99% of the wt-AraC population in the absence of L-ara (84%, 85%, and 75% of the total library populations, respectively). This effectively reduced the fluorescent geometric mean of each population by 5-38%, decreasing the frequency of clones lacking repressibility. The plasmid DNA was isolated from the three resulting populations and pooled together (CLib5s), which was used as the starting population for screening in the presence of each compound tested (Figure IV-3). CLib5s was transformed into SQ12 electrocompetent cells (total transformants 3 x 10^8) and plated on selection plates supplemented with ampicillin and one of the following 5 mM TAL, 3 mM p-coumaric acid, 10

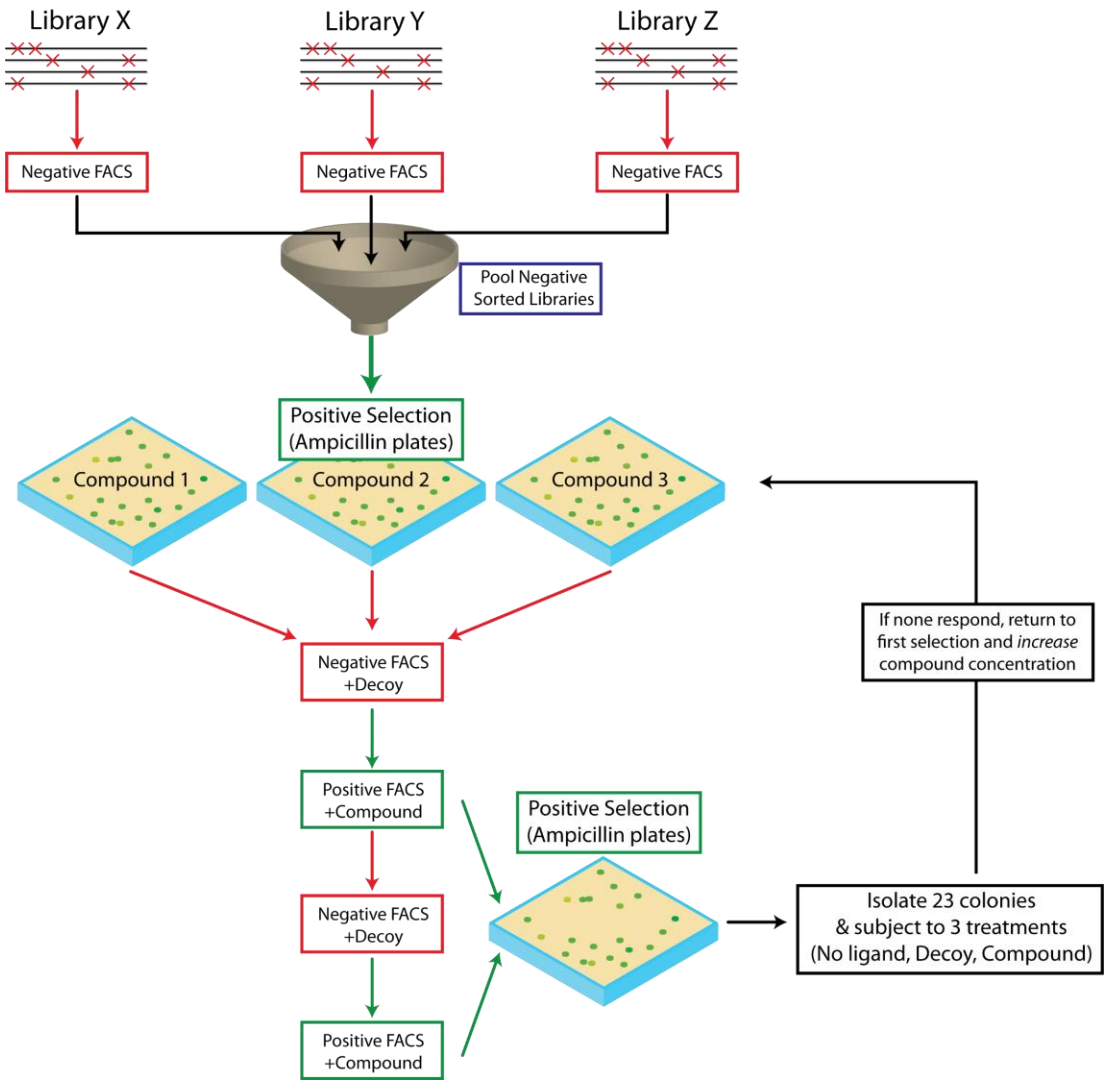


Figure IV-3 Enhanced ligand–induced combination screening (ELICS) flow diagram for screening combinatorial protein libraries.

mM p-coumaric acid, 2 mM t-cinnamic acid, 5 mM t-cinnamic acid, 2 mM vanillin, 2 mM nicotinic acid, 5 mM nicotinic acid, 2 mM salicylic acid, 5 mM phloroglucinol, 5 mM quinic acid, 5 mM gallic acid, 5 mM shikimic acid, or 5 mM gluconic acid lactone. These compounds are not metabolized by our *E. coli* strains. Colony-forming units were counted and the average survival for all the samples was 0.1 ± 0.4%, with the highest being in the presence of 2 mM

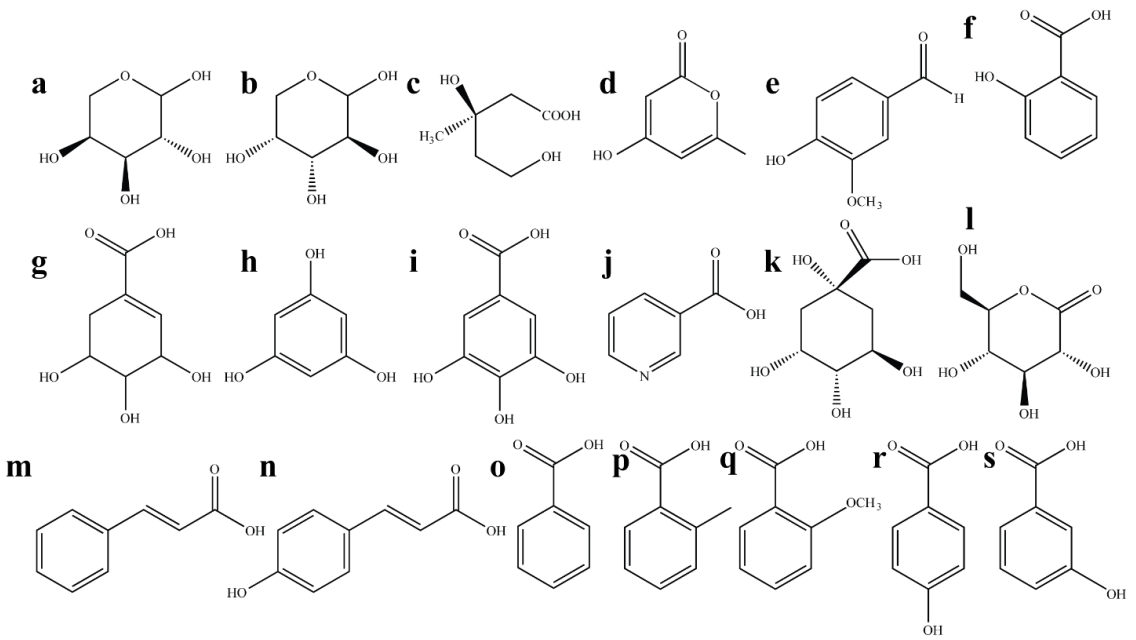


Figure IV-4 Compounds observed in this study (a) L-arabinose (b) D-arabinose (c) mevalonate (d) triacetic acid lactone ($pK_{a,enol}$=5.1) (e) vanillin (f) salicylic acid ($pK_{a,COOH}$=2.97) (g) shikimic acid ($pK_{a,COOH}$=4.76) (h) phloroglucinol (i) gallic acid ($pK_{a,COOH}$=4.11) (j) nicotinic acid ($pK_{a,COOH}$=4.75) (k) quinic acid ($pK_{a,COOH}$=3.58) (l) gluconic acid lactone (m) trans-cinnamic acid ($pK_{a,COOH}$=4.44) (n) p-coumaric acid ($pK_{a,COOH}$=4.34) (o) benzoic acid (p) o-toluic acid (q) 2-methyoxybenzoic acid (r) 4-hydroxybenzoic acid (s) 3-hydroxybenzoic acid. Wild-type AraC responds to a (L-ara). Previous AraC variants have been found to respond to compounds b-d. In this study, compounds e-n were screened for AraC variant recognition. AraC variants were found for compounds e and f. Compounds o-s were used in specificity assays for the newly isolated AraC variants.

vanillin at 0.78 ± 0.01% on plates supplemented with 100 μg/mL ampicillin. Therefore, the CLib5s population was significantly reduced, thus decreasing the overall load on the cytometer for screening the remaining populations.

To ensure the need for additional screening, single colonies were selected from each selection plate and tested in liquid culture for response to their respective compound. None of the selected clones showed a response, except for two clones responding to salicylic acid as discussed below. Because the frequency for responsive clones was low, we concluded that additional enrichment through further rounds of screening was necessary.

All of the remaining colonies on the selection plates were scraped and collected; the plasmid DNA was isolated from each population and subsequently transformed into HF19

electrocompetent cells. Though the SQ12 strain is adequate for biosensor expression of GFPuv (data not shown), we did not want to include the additional stress of β-lactamase expression. Post-selection populations were negative sorted in the presence of 100 μM L-ara via FACS (sort gate was setup as previously described). Subsequent populations were then introduced to their respective compounds at the same concentration as was present in the solid medium for the selections and positive sorted for the top 5% most fluorescent cells. The collected population was subjected to direct plasmid recovery according to the protocol designed by Ramesh and coworkers (Ramesh et al., 2015). Direct recovery of the plasmid DNA reduced bias during outgrowth and improved the population behavior. Each population was subjected to one more rounds of negative and positive sorting. Both of the positive FACS screened populations were subjected to one further round of selection in strain SQ12 on ampicillin supplemented plates.

Clones from each population after selection (endpoint clones) were screened in deep well plates for a response to their respective compound. The results of each endpoint are listed in Table IV-1. The majority of endpoint clones were tested and did not respond to 100 μM L-ara with (<2-fold response) (data not shown). Despite their growth on the selection plates, clones isolated from several populations in response to their respective compounds (p-coumaric acid, t-cinnamic acid, nicotinic acid, phloroglucinol, quinic acid, gallic acid, shikimic acid, and gluconic acid lactone) were unresponsive in liquid cultures and therefore no clones from these populations were selected for further testing. The lack of response from these populations could be due to 1) the compound degrades or reacts under culture condition or 2) the library lacks any responsive clones. The former is most likely the case for phloroglucinol and gallic acid due to the dramatic color change observed after the cultures were grown, which was verified by HPLC analysis and the presence of multiple peaks. Compounds such as p-coumaric acid and t-cinnamic acid may simply hinder allosteric interactions due to their spatial requirements exceeding the ligand binding pocket volume. Conversely, the high frequency of positive endpoint clones responding

to TAL, vanillin,

Table IV-1 List of endpoint population responses to their respective target compound. Endpoint populations are listed by the compound and its concentration that they were screened in the presence of. The frequency is a measure of the fraction of clones responsive to the target compound per the number of clones screened in the deep-well plates.

Compound/ Concentration	Frequency ($N_+/N_{screened}$)[a]	Average Response	Highest Response[c]
5 mM TAL	0.76	5.9 ± 4.8	11.1 ± 1
3 mM p-Coumaric acid	0	1.1 ± 0.2	1.2 ± 0.1
5 mM p-Coumaric acid	0.04	1.8 ± 0.2	2.2 ± 0.2
2 mM t-Cinnamic acid	0	1.1 ± 0.2	1.6 ± 0.1
5 mM t-Cinnamic acid	0	1.1 ± 0.4	1.7 ± 0
2 mM Vanillin	0.57	11.4 ± 11.2	32.9 ± 2.4
2 mM Vanillin[b]	1.00	11.8 ± 2.7	21 ± (na)
2 mM Salicylic acid[b]	0.67	19.6 ± 21.4	43.3 ± (na)
2 mM Nicotinic acid[b]	0	1.0 ± 0.0	1.0 ± (na)
5 mM Nicotinic acid	0	0.9 ± 0.1	1.1 ± 0.1
5 mM Phloroglucinol	0	0.9 ± 0.1	1.0 ± 0.1
5 mM Quinic acid	0	0.9 ± 0.1	1.0 ± 0.1
5 mM Gallic acid	0	0.8 ± 0.4	1.7 ± 0
5 mM Shikimic acid	0	1.0 ± 0.1	1.2 ± 0.2
5 mM Gluconic acid lactone	0	1.0 ± 0.1	1.1 ± 0.1

[a] N_+ is the number of clones with >2-fold response; $N_{screened}$ is the total clones screened

[b] Clones were isolated directly from selection plates and were not subjected to replicate culturing

[c] Standard deviation of highest response clone was calculated from replicate cultures

and salicylic acid fortified the benefits of the additional endpoint selection to enrich positive clones (if there are any present in the population). The sequence of each clone is reported in Table IV-2.

Of the 31 vanillin clones (AraC-Van) sequenced, there were eight different clones (AraC-Van1-8) at the amino acid level. AraC-Van3 was represented three times with different DNA sequences but the same residue substitutions (AraC-Van3a-c). Each of the three clones showed a statistically similar response. The majority of the AraC-Van clones originated from JLib1s, but two of the eight were derived from SLib4s (AraC-Van2 and AraC-Van8). Additionally, AraC-

Table IV-2 List of the resulting clones isolated for TAL, vanillin, and salicylic acid. The codons and residues are reported according to their residue substitution position.

	Codon						Residue					
Clone	8	24	38	80	82	93	8	24	38	80	82	93
WT-AraC	CCC	ACG	CGA	CAT	TAC	CAC	P	T	R	H	Y	H
AraC-TAL	GTG	ATC	CGA	GGC	TTG	CGC	V	I	R	G	L	R
AraC-TAL11	TGC	TTG	CGA	GGG	TTG	CGC	C	L	R	G	L	R
AraC-Van1	CCC	CTC	GTG	GGG	ATC	TGC	P	L	V	G	I	C
AraC-Van2	GGG	TTG	CGA	CAC	AAG	TTG	G	L	R	H	K	L
AraC-Van3a*	CCC	CTC	GTG	AGC	GTC	GCG	P	L	V	S	V	A
AraC-Van3b*	CCC	CTC	GTC	TCG	GTC	GCG	P	L	V	S	V	A
AraC-Van3c*	CCC	CTG	GTG	TCG	GTG	GCG	P	L	V	S	V	A
AraC-Van4	CCC	GCC	ACG	TCG	TGC	TTC	P	A	T	S	C	F
AraC-Van5	CCC	CTG	GTC	GGG	GCG	GCC	P	L	V	G	A	A
AraC-Van6	CCC	GCG	GTG	GCC	TGC	TTC	P	A	V	A	C	F
AraC-Van7	CCC	GCG	CTG	GCG	GTC	TTC	P	A	L	A	V	F
AraC-Van8	GGG	CTG	CGA	CAC	AAG	TTC	G	L	R	H	K	F
AraC-Sal1	CCC	ACG	CGA	CAT	TAC	CAC	P	T	R	H	Y	H
AraC-Sal2	CCC	ACG	CGA	CAT	TAC	CAC	P	T	R	H	Y	H
AraC-Sal3	GGG	CTC	CGA	GTC	CGG	TTC	G	L	R	V	R	F
AraC-Sal4	GTG	ATC	CGA	GGC	TTG	CGC	V	I	R	G	L	R

*-Clones have same residues but different codons

Van8 had an identical sequence to the previously isolated AraC-TAL8. Salicylic acid responding endpoint clones were most intriguing because the highest responding clone (AraC-Sal4) had an identical sequence to AraC-TAL1. Also, two AraC-Sal clones were isolated and screened from direct selection of naïve CLib5s, each clone originating from CLib2s. No responsive clones were isolated using this method for any other compound. Both of these clones, AraC-Sal1 and AraC-Sal2, contained a substitution at residue E149 with either a glycine or lysine, respectively. Substitutions at residue E149 have been shown to promote constitutivity (Dirla et al., 2009), but only one AraC-Sal1 showed a high level of background expression (200 ± 70 rfu). Each clone contained two silent mutations and three missense mutations. The missense mutations for AraC-Sal1 and AraC-Sal2 were M42I, P86Q, and E149G and G30C, V56A and E149K, respectively.

4.2.4) Variant analyses reveals dynamic and promiscuous responses

After isolation of AraC-Sal4 and AraC-Van8 with identical sequences to AraC-TAL1 and AraC-TAL8, respectively, we tested each of the isolated variants of AraC-Van and AraC-Sal for responses to their target compound and various analogs of vanillin and salicylic acid. The data from the dose responses were fitted using the Hill equation and the Hill parameters were used to calculate the range of response and apparent K_d ($K_{d,app}$), reported in Table IV-3. The range of the vanillin and salicylic acid biosensors for their respective compounds spanned only approximately a single order of magnitude on average (Figure IV-5). The maximum fold-response (S/B_{max}) of the top variants (AraC-Van1, 2, and 6; AraC-Sal1, 3, and 4) show similar maximum fold-response as wt-AraC.

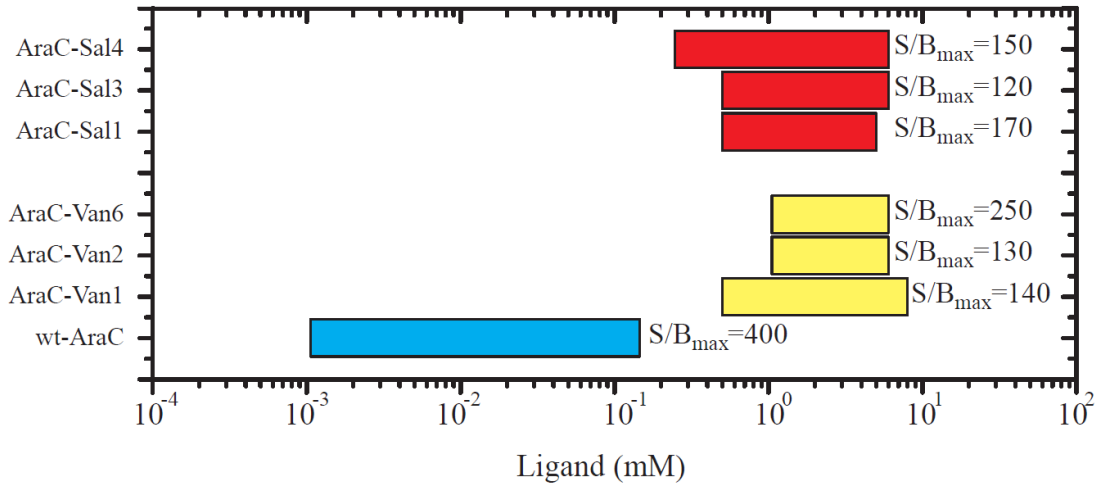


Figure IV-5 (A) Floating bar graph representing the range of response for the best vanillin and salicylic acid variants compared with the range of response of wt-AraC to L-ara. The clones are listed along the y-axis and ligand concentration along the x-axis. The bars represent response range of the biosensor in presence of salicylic acid (red), vanillin (yellow), and L-ara (blue). The solid bars represent the dynamic range of response. The maximum fold response of the variants ($Fold_{max,var}$) is represented as the percent of maximum fold-response of wt-AraC ($Fold_{max,wt}$).

Table IV-3 Response of isolated biosensors to L-ara, TAL, vanillin, and salicylic acid

Clone	Background (leakiness)	100 µM L-ara	5 mM Vanillin	5 mM Salicylate	5 mM TAL	Fold-Response			$K_{d,app}$ (mM)[a]	
						Vanillin	Salicylate	TAL	Vanillin	Salicylate
WT-AraC	32	4400	27	29	19	0.9	0.9	0.6	--	--
AraC-Van1	38	30	3000	61	28 ± 8	78	1.6	0.7	5.4	--
AraC-Van2	40	31	4400	30	41	109	0.8	1.0	4.2	--
AraC-Van3a*	27	33	1400 ± 700	28	24 ± 6	52	1.0	0.9	6.1	--
AraC-Van3b*	31	31	1600	30	28 ± 6	52	1.0	0.9	6.1	--
AraC-Van3c*	34	30	2100	30	25 ± 7	62	0.9	0.7	5.0	--
AraC-Van4	34 ± 7	28	1800 ± 800	29	24	54	0.9	0.7	4.7	--
AraC-Van5	34 ± 8	34 ± 8	3200	45	33 ± 13	94	1.3	1.0	5.4	--
AraC-Van6	32	43 ± 15	8500	39	33 ± 14	268	1.2	1.0	4.1	--
AraC-Van7	34	34	220 ± 44	39	37 ± 14	6.5	1.1	1.1	7.0	--
AraC-Van8	32	30 ± 7	270	35	120 ± 36	8.5	1.1	3.8	4.2	--
AraC-Sal1**	47	8500	72 ± 17	7900	36	1.5	167	0.8	--	2.6
AraC-Sal2**	43	1300	45	650	34	1.0	15	0.8	--	6.8
AraC-Sal3	41	35 ± 12	46	2500	36	1.1	61	0.9	--	4.0
AraC-Sal4	35	38 ± 13	44	3200	320	1.2	90	9.0	--	3.5
AraC-TAL11	26	--	--	--	97 ± 22	--	--	3.8	--	--

*-Clones have same residues but different codons

**-Clones were isolated from error-prone library

a-$K_{d,app}$ was only calculated for samples with complete dose response curve before reaching the toxicity limit of compound in the media

-- indicates that the data was not collected or results could not accurately calculate the parameter

Each vanillin and salicylic acid clone was grown in liquid rich medium supplemented with several concentration of compounds L-ara (**a**), TAL (**d**), benzoic acid (**o**), o-toluic acid (**p**), 2-methyoxybenzoic acid (**q**), 3-hydroxybenzoic acid (**r**), and 3-hydroxybenzoic acid (**s**). The responses of the clones to the compound at the highest concentration tested are listed in Table IV-4. A response to the native ligand L-ara was only present for the clones isolated from the error-prone library, AraC-Sal1 and AraC-Sal2. The response to 20 mM TAL was strong for AraC-Van2, AraC-Van8 (AraC-TAL8), AraC-Sal1, and AraC-Sal4 (AraC-TAL1). AraC-Van2 and AraC-Van8 were the only vanillin clones isolated from SLib4s and each contained the same substitutions except at position H93, where a leucine and phenylalanine were substituted, respectively. The data therefore suggests that either a substitution at R38 is ideal for vanillin response or a substitution at residue P8 is not necessary for a strong response to vanillin. Interestingly, AraC-Sal3 was highly specific towards salicylic acid, showing no response to isomers of salicylic acid, 3- and 4-hydroxybenzoic acid. Except for AraC-Van2 and AraC-Van8, the AraC-Van clones were also highly specific towards their target compound, vanillin.

Since there was overlap with the TAL variants, we tested the response of the previously isolated TAL variants for a response to salicylic acid and vanillin, and surprisingly, they all either responded strongly to salicylic acid or vanillin (none of them showed a strong response to both compounds). Therefore, there should be little surprise that vanillin and salicylic acid variants had the same sequence as some TAL variants. To further improve the specificity of isolated variants, future screening strategies will include a "cocktail" of decoy ligands, where multiple decoy ligands will be introduced into the culture simultaneously. This would be more applicable to enzyme engineering, where compounds with similar structures will be present in the assay and could potentially inhibit or induce a false-positive response.

Table IV-4 Table of results from specificity assay for vanillin and salicylic acid variants

	Fold-Response								
Clone	5 mM Vanillin	5 mM Salicylate	10 mM L-ara	20 mM TAL	7 mM BA	7 mM 2-MBA	7 mM 3-HBA	7 mM 4-HBA	7 mM o-TA
WT-AraC	0.9	0.9	394	1.2	0.7	1.1	0.7	1.0	1.0
AraC-Van1	78	1.6	0.8	1.2	1.4	1.7	1.1	1.3	1.1
AraC-Van2	109	0.8	0.9	**18**	0.7	1.1	0.8	1.2	0.9
AraC-Van3a*	52	1.0	1.1	0.6	0.6	1.0	0.5	1.3	1.0
AraC-Van3b*	52	1.0	1.2	0.7	0.8	1.0	0.8	1.4	0.8
AraC-Van3c*	62	0.9	0.9	0.7	0.7	0.9	0.8	1.5	0.8
AraC-Van4	54	0.9	0.9	0.8	0.6	1.1	0.7	1.5	0.8
AraC-Van5	94	1.3	1.0	1.1	0.7	1.0	0.8	1.1	0.8
AraC-Van6	268	1.2	0.9	0.9	0.8	0.8	0.8	1.2	0.8
AraC-Van7	6.5	1.1	1.0	0.9	0.8	1.0	0.7	1.1	0.8
AraC-Van8	8.5	1.1	0.8	**53**	0.8	0.9	0.8	1.0	0.8
AraC-Sal1**	1.5	167	**329**	**12**	**53**	**2**	1.1	**4**	**22**
AraC-Sal2**	1.0	15	**53**	0.9	0.9	1.2	0.7	1.1	0.8
AraC-Sal3	1.1	61	0.8	1.2	0.8	1.0	0.8	1.2	0.8
AraC-Sal4	1.2	90	1.0	**78**	0.8	1.3	0.7	1.2	0.9

	Fold-Response	
Clone***	5 mM Vanillin	5 mM Salicylate
AraC-TAL1	0.7	**99**
AraC-TAL2	**16**	0.7
AraC-TAL3	1.1	**71**
AraC-TAL4	1.3	**89**
AraC-TAL5	0.4	**33**
AraC-TAL6	**39**	0.2
AraC-TAL7	1.3	**20**
AraC-TAL8	**13**	0.3
AraC-TAL9	1.9	**17**
AraC-TAL10	**3**	**53**
AraC-TAL11	1.1	**56**

*-Clones have same residues but different codons

**-Clones were isolated from error-prone library

***-Clones were not measured with a degradation tag present on the GFPuv

Abbreviations: L-ara, L-arabinose; TAL, triacetic acid lactone; BA, benzoic acid; 2-MBA, 2-methyoxybenzoic acid; 3-HBA, 3-hydroxybenzoic acid; 4-HBA, 4-hydroxybenzoic acid; o-TA, o-toluic acid

4.2.5) Single amino acid substitutions in AraC increase response

Two of the four salicylic acid variants (AraC-Sal1 and AraC-Sal2) described above originated from the random mutagenesis library (CLib2s). Each variant contained five total mutated codons, two silent mutations and three missense mutations (Table IV-5). Both of them contained a mutation at residue E149. Dirla and coworkers previously reported constiutivity of AraC with a single amino acid substitution at residue E149 (E149F) (Dirla et al., 2009). These two variants did not show as strong of a response as the variants isolated from the site-saturated libraries, but this was the first time we have isolated variants from a random mutagenesis library. Therefore, we sought to determine the importance of the missense mutations in confiring a response to salicylic acid. The wt-AraC and AraC-Sal4 variants were substituted with each of the missense mutations one at a time and tested for their response to L-ara, TAL, and salicylic acid. The results are outlined in Figure IV-6. The wt-AraC retained a strong response for all substitutions except for E149K. However, the E149K wt-AraC variant shows a response to 5 mM salicylic acid that is almost 2-fold greater than AraC-Sal4. The E149G wt-AraC variant also shows a strong response to salicylic acid. This may be due to the negative charge of glutamic acid under physiological conditions, which would repel the negatively charged hydroxyl group of salicylic acid. Also, the substitution M42I in AraC-Sal4 shows a reduced response and a high level of leakiness, but when the GFP tagged with the degredation tag is used, the background was

Table IV-5 Table of substituted residues of salicylic acid variants originating from the error-prone library (CLib2s). Red letters indicate a missense mutation and green letters indicate a silent mutation.

		Clone	Residue										
		Name	30	42	56	86	122	133	140	149	151	Total	Total
Compound	Original	WT-AraC	G	M	V	P	P	L	A	E	L	Missense	Silent
Screened	Library	AraC-TAL	G	M	V	P	P	L	A	E	L	Mutations	Mutations
Salicylic Acid	Clib2s	AraC-Sal1	G	I	V	Q	P	L	A	G	L	3	2
Salicylic Acid	Clib2s	AraC-Sal2	C	M	A	P	P	L	A	K	L	3	2

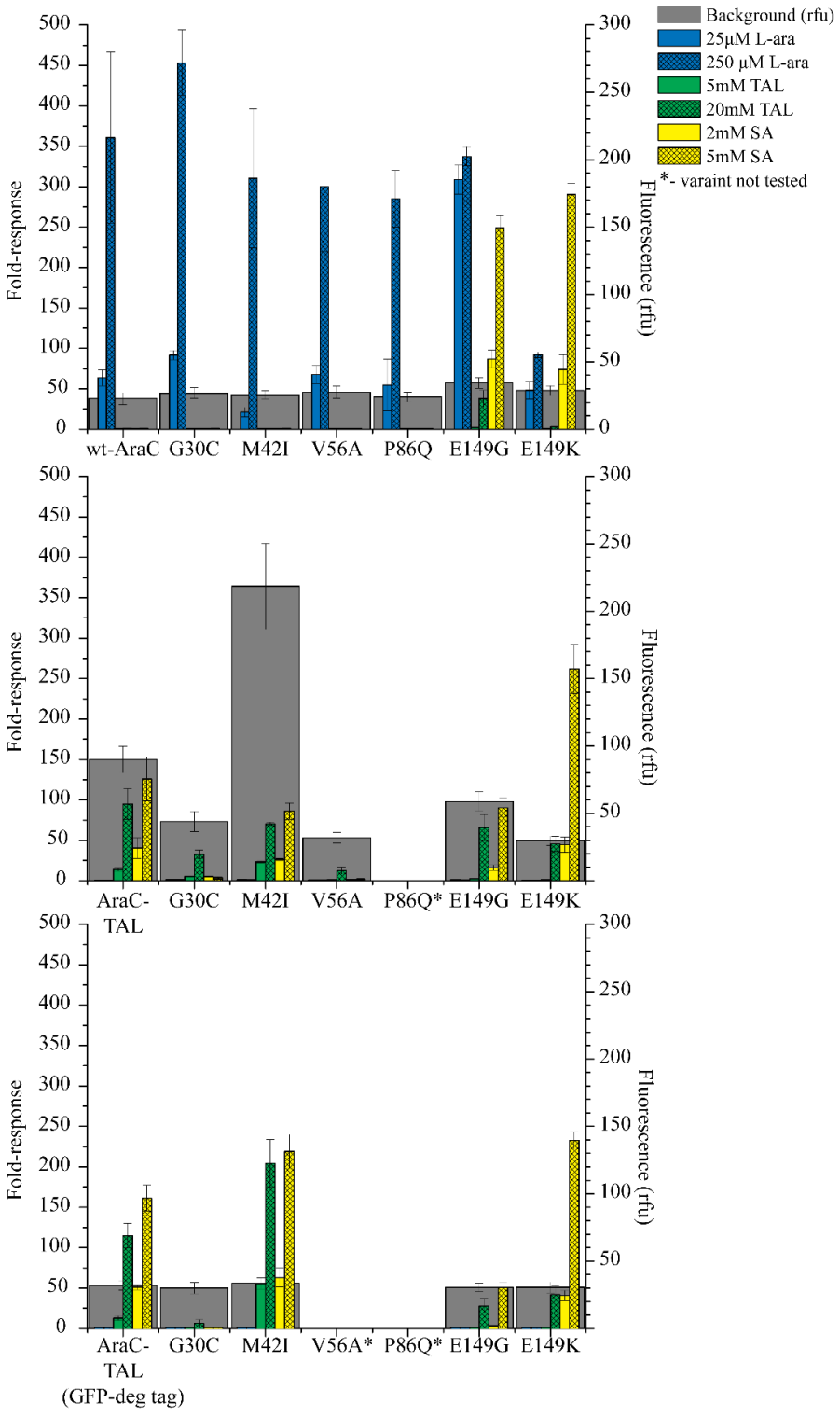


Figure IV-6 Bar graph of AraC variants with single substitutions from salicylic acid variants isolated from error-prone library. The gray bars represent the background fluorescence of each variant. The colored bars represent the fold-response of the variant to the respective compound. L-ara, L-arabinose; TAL, triacetic acid lactone; SA, salicylic acid

reduced and the response was dramatically improved. This supports our hypothesis that a low level of leakiness enhances the response of the AraC-based biosensor. The remaining substitutions either decrease the response or do not chance the response dramatically. Therefore in future rounds of screening, we will also target residues M42 and E149 in our combinatorial libraries because they show here to have dramatic effects on the biosensor response.

This enhanced ligand-induced combination screening protocol offers a versatile platform for engineering novel synthetic molecular biosensors. Throughout the design of this method, we discovered various strategies to decrease the leakiness of uninduced clones by incorporating glycerol into the culture media and tagging GFPuv with a degradation tag. The simplified single plasmid expression system alleviated growth inhibitions due to toxic growth conditions. Due to the increasing demand for high-throughput screening techniques, this should be applicable to discovery of novel TRP-based biosensors for target compounds and their subsequent use in metabolic engineering.

V Materials and Methods

This chapter is a compilation of the various experimental procedures and materials used throughout this thesis. The sections are divided according to chapter.

V.1) Chapter II

5.1.1) General Methods

Restriction enzymes, Phusion High-Fidelity DNA polymerase, and T4 DNA ligase were purchased from New England Biolaps (Ipswich, MA). Oligonucleotides were synthesized by Invitrogen (Carlsbad, CA) or Integrated DNA Technologies (Coralville, IA). DNA sequencing was performed at either the Pennsylvania State University Huck Institutes of the Life Sciences Genomics Core Facility (http://www.huck.psu.edu/facilities/genomics-up), SeqWright (Houston, TX), or Lone Star Labs (Houston, TX). All chemicals were purchased from Sigma-Aldrich (St. Louis, MO). Molecular biology techniques for DNA manipulation were performed according to standard protocols (Sambrook and Russell, 2001) and all cultures were grown in lysogeny broth (LB).

5.1.2) Plasmid and Library Construction

This work is based on the dual plasmid reporter system for AraC-controlled GFPuv expression that has been described previously(Tang et al., 2008), where AraC is expressed from plasmid pPCC423 (maintained by apramycin antibiotic resistance) by IPTG-inducible LacI and GFPuv is subsequently expressed from plasmid pPCC442 (chloramphenicol resistance) where it is controlled by P_{BAD}. Plasmid maps are shown in Supp. Figure D-2.

Two DNA libraries were created by site saturation mutagenesis at five residues of araC using overlap-extension PCR. AraC library "SLib4" was constructed previously (Tang and Cirino, 2011) using plasmid pPCC423 (Tang et al., 2008) with mutations at residue positions 8, 24, 80, 82, and 93. The second AraC library, "JLib1", was constructed in a similar fashion with

mutations at positions 24, 38, 80, 82, and 93 and cloned into plasmid pFG1. First, plasmid pFG1 was created from pPCC423 by inserting additional restriction enzyme sites upstream of P_{tac} in order to facilitate future cloning projects using alternative promoters and regulatory protein genes. Briefly, primer 423mcs4-for containing the additional restriction sites was paired with primer 423mcs-rev to amplify the promoter region through a standard PCR reaction. Primer sequences are listed in Supp. Table B-4. The resulting PCR product was digested with NcoI and NdeI and ligated into pPCC423 digested with the same enzymes to yield plasmid pFG1, which was confirmed by DNA sequencing analysis.

JLib1 was created as follows. Three parallel PCR reactions were performed to amplify three araC segments (A, B, and C) using the following three sets of primers: 423lib-for-NcoI and araC-T24-rev; araC-38-for-2 and araC-H80-Y82-rev; and araC-H93-for and araC-rev-4, respectively. Fragment A contained site-saturation mutations at residue position 24; fragment B contained mutations of positions 38, 80, and 82; and fragment C contained mutations for position 93. PCR products were gel purified and equimolar aliquots of adjacent DNA fragments were combined (A+B; B+C) and PCR-assembled without primers. Finally, outer primers 423lib-for-NcoI and araC-rev-4 were added to each assembly reactions and the products were PCR-amplified. The full-length products were digested with NdeI and HindIII, gel purified, and ligated into pFG1 that had been digested and gel purified with the same enzymes. Ligation products were dialyzed on Millipore membranes and transformed into *E.coli* strain MC1061. An aliquot of the outgrowth was diluted and streaked onto agar plates containing apr to determine the total number of transformants while the remaining outgrowth (~20 mL) was used to inoculate 480ml LB containing antibiotic and grown overnight at 37°C. Approximately 4.8×10^7 unique transformants were recovered, ensuring complete coverage of the entire library. The plasmid library was then prepared using 100ml of the overnight culture and ten randomly picked clones from the library were sequenced to reveal the expected random mutations at the targeted nucleotide positions. In addition, two of the ten clones contained a single, but different, point

insertion between residue positions 82 and 93, which were presumably introduced by either the primers or the polymerase during the PCR process.

The JLib1 and Slib4 araC gene libraries contained in plasmids pFG1 and pPCC423, respectively, were transformed into strain HF19 (Tang et al., 2008) harboring the reporter plasmid pPCC442. After 1hr of incubation at 37°C while shaking at 250 rpm, an aliquot of the outgrowth (4 mL) was diluted and streaked onto agar plates containing apr+cmr to determine the number of transformants (6.1×10^9 for JLib1; 8.1×10^8 for SLib4) and again ensure coverage of the complete library. The remainder of the outgrowth was used to inoculate 96ml LB+apr+cmr in a 500 mL glass Erlenmeyer flask and grown for 3hrs. Subsequently, 10ml of the 3hr culture were used to inoculate 500ml media with antibiotics in a 2 L glass Erlenmeyer flask and were grown overnight. Finally, the 500ml culture was mixed with 80% glycerol to achieve a final glycerol concentration of 20% and frozen in 1.0ml aliquots in liquid nitrogen before storage at -80C. These glycerol stocks are referred to as J51 (JLib1) and J10 (SLib4). For use as control, glycerol stocks of HF19 cells containing plasmids pFG1 and pPCC442 were prepared similarly and are referred to as J50 (wild-type; WT).

5.1.3) Pyramid screening using Fluorescence-Activated Cell Sorting (FACS)

Cells were prepared for screening by inoculating 9.0 mL of LB containing apr+cmr and 0.1mM IPTG with 1.0 ml of glycerol stock containing either library (or WT for control) in a 125 mL glass erlenmeyer flask and incubated overnight at 37C while shaking at 250 rpm. Where indicated, cultures were supplemented with effector ligands to the following final concentrations: L-arabinose (10 mM), D-arabinose (100 mM), p-coumaric acid (2 mM), and mevalonate (30 mM). The following morning, cells were diluted 1:200 in PBS and analyzed on a FACSJazz (BD Biosciences) to obtain flow cytometry histograms. The appropriate cultures were then subjected to FACS.

5.1.4) High-throughput sequencing of library populations

Samples for next generation sequencing were prepared as follows. The LBD sequence was amplified from plasmid DNA preps of the populations using primers with, in order form 5' to 3', an Illumina adaptor sequence, barcoding sequence, sequencing primer adaptor, and a homologous region to the LBD DNA sequence. Each set of primers (15 total) were each designed with a unique barcode so multiple samples could be run simultaneously. The PCR were setup as follows. The reaction conditions were setup as describe by the instructions of Phusion polymerase. Each reaction contained 1 ng/µL plasmid DNA, 1x HF Phusion buffer, 10 µM dNTPs, 0.5 µM of each the forward and reverse primer (AraC_Forward_Univ and AraC-reverse-BC1-15 primers in Suppl. Table B-4), 3% DMSO, and 0.04 U/µL Phusion in 300 µL total volume. The total reaction was split into three separate reaction tubes and the following PCR protocol was used: 98°C for 30 s; 98°C for 10 s, 68°C for 15 s, 72°C for 15 s, cycled 30 times; 72°C for 10 min; cooled to 4°C. The three reactions were mixed and concentrated using a QIAGEN PCR Purification kit (QIAGEN, Cat. No. 28106), eluted with 60 µL of QIAGEN elution buffer. Samples were placed in a centrifuge-vacuum chamber at 55°C for 1 hr to remove any residual ethanol from the purification kit. Samples were then gel purified from a 2% agarose gel. A QIAquick Gel Extraction kit (QIAGEN, Cat. No. 28706) was used to extract the DNA from the gel. Samples were eluted with 60 µL of QIAGEN elution buffer and subsequently 20 µL was run on a 2% agarose gel and a NanoDrop Lite to determine the total mass of each sample. Samples generally yielded >2 µg of purified DNA.

5.1.5) Deep-Well Plate Clone Screening

Library endpoint plasmid DNA was isolated and transformed into electroporation competent HF19 cells. Clones were isolated from LB agar plates and streaked onto new LB agar plates. Quadruplicate 500 µL LB supplemented with apramycin starter cultures in 2 mL 96-well (DW) plates were inoculated from each isolated clone. The starter culture was incubated for 6 hrs

at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. Quadruplicate 500 µL BM supplemented with apramycin and 100 µM IPTG subcultures in 96-well (DW) plate containing the appropriate target ligand were inoculated by 50-fold dilutions of the respective starter culture (OD_{595} ~0.2). The subcultures were incubated for 6 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. The cultures were washed once with 1 mL of phosphate buffered saline and measured the OD_{595} on a BMG Labtech NOVOstar spectrophotometer and the fluorescence on a Molecular Devices SpectraMax Gemini EM.

5.1.6) Deep-Well Plate dose responses

Isolated clones from the deep-well plate clone screening were digested and cloned back into the parent vector using the same method described above for cloning the libraries into the pFG29 vector. The re-cloned mutants were transformed into electroporation competent HF19 cells. Clones were isolated from LB agar plates supplemented with apramycin and quadruplicate 500 µL LB supplemented with apramycin starter cultures in 2 mL 96-well (DW) plates were inoculated from each isolated clone. The starter culture was incubated for 6 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. Quadruplicate 500 µL BM supplemented with apramycin and 100 µM IPTG subcultures in deep-well plate containing a range of target ligand were inoculated by 50-fold dilutions of the respective starter culture. The subcultures were incubated for 16 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. The cultures were washed once with 1 mL of phosphate buffered saline and measured the OD_{595} on a BMG Labtech NOVOstar spectrophotometer and the fluorescence on a Molecular Devices SpectraMax Gemini EM.

V.2) Chapter III

5.2.1) General

Escherichia coli strains used in the this study were MC1061 (F- Δ(ara-leu)7697 [araD139]B/r Δ(codB-lacI)3 galK16 galE15 λ- e14- mcrA0 relA1 rpsL150(strR) spoT1 mcrB1 hsdR2(r-m+)) for plasmid propagation and HF19, previously described by Tang and coworkers, for biosensor expression. Restriction enzymes, Phusion High-Fidelity DNA polymerase (Cat. No. M0530L), and T4 DNA ligase (Cat. No. M0202L) were purchased from New England Biolabs (Ipswich, MA). Oligonucleotides were synthesized by Integrated DNA Technologies (Coralville, IA). Sanger DNA sequencing was performed by SeqWright (Houston, TX). All chemicals were purchased from Sigma-Aldrich (St. Louis, MO). Molecular biology techniques for DNA manipulation were performed according to standard protocols (Sambrook and Russell, 2001) and all cultures were grown in lysogeny broth (LB) or modified LB. For biosensor analysis, LB with 50 mM TES pH 7.2 and 1% glycerol was used, referred to as biosensor media (BM). Unless otherwise stated, all cultures were grown at 37°C and 250 rpm in a New Brunswick Excella E25 incubator. Flasks were all non-baffled. Deep-well plates were purchased from Corning Life Sciences (Cat. No. 3960) and had square wells with conical bottoms for optimal aeration and pelleting. All Gibson assembly was performed using the appropriate primers and a Gibson Assembly Master Mix Kit from NEB (Cat. No. E2611) and NEBuilder provided by New England Biolabs.

The parent plasmid vectors used in this study contained either a modified RSF1030 origin (pPCC442) with high copy (contains ~200 copies per cell) (Phillips et al., 2000) or the pBR322 origin lacking the *rop* gene on the plasmid (pPCC423, pFG1, pFG29) which is medium copy (30-60 copies per cell). Triacetic acid lactone (4-hydroxy-6-methyl-2-pyrone) was purchased from Sigma-Aldrich (Cat. No. H43415). The powder was directly dissolved in the appropriate medium to a final concentration of 50 mM and buffered to pH 7.0 with 10 M NaOH. The solution was

sterile filtered with a 0.2 um syringe filter and stored at 4°C for up to one month. L-arabinose was purchased from Sigma-Aldrich (Cat. No. A3256) and was prepared by dissolving the powder into biology grade water to a final concentration of 1 M. The stock was sterile filtered using a 0.2 um syringe filter and stored at room temperature.

5.2.2) Substituted residue analysis

Cloning- All AraC-TAL back-crossing mutants were cloned using Gibson Assembly. The Gibson Assembly fragments were amplified from pFG29-TAL, and the primers carry the corresponding mutations. All PCR reactions used New England Biolabs Phusion® polymerase and the recommended PCR conditions for Phusion® polymerase was followed. Annealing temperatures were obtained from NEB Tm Calculator (http://tmcalculator.neb.com/). The PCR fragments were purified using ZymocleanTM Gel Recovery Kit. The vector was obtained by double digestion (BstAPI and AflII) of pFG29-TAL, and recovered with ZymocleanTM Gel Recovery Kit. Gibson Assembly was performed using New England Biolabs Gibson Assembly Master Mix, and the recommended protocol from New England Biolabs website was followed.

Assay- HF19 competent cells were transformed individually with the 32 *araC*/*araC-TAL* mutants through electroporation and subsequently plated on LB-agar plates supplemented with 50 µg/mL apramycin. Fresh colonies were inoculated into 500 µL cultures in 96 deep well plates with LB supplemented with 50 µg/mL of apramycin. The cultures were grown at 37°C 900 rpm up to OD_{595} of 5-6 in a Heidolph Titramaz 1000/Inkubator 1000. The cultures were diluted to OD_{595} of 0.2 in 200 µL BM supplemented with 50 µg/mL of apramycin and 100 µM IPTG. The subcultures were incubated for 6 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. The cultures were washed once with 200 µL of phosphate buffered saline and OD_{595} was measured on a BMG Labtech NOVOstar spectrophotometer and the fluorescence on a Molecular Devices SpectraMax Gemini EM.

5.2.3) Plasmid and library construction

Initial work was based on the dual plasmid reporter system for AraC-controlled GFPuv expression that was described previously , where AraC is expressed from plasmid pPCC423 (maintained by apramycin antibiotic resistance) controlled by IPTG-inducible LacI. GFPuv was subsequently expressed from plasmid pPCC442 (chloramphenicol resistance), where it is controlled by P_{BAD}. Plasmid maps are shown in Supp. Figure D-2. Subsequently, plasmid pFG1 for AraC expression was constructed from pPCC423 and pFG29 was constructed using the pFG1 vector (described below). Construction and expression of the mutant library was carried out as previously described.

pFG1 was further cloned to incorporate Pbad-*gfpuv*. This was accomplished using the Gibson method. Primers pFG29-gib-for and pFG29-gib-rev were designed to amplify Pbad-*gfpuv* from pPCC442. Primer pFG29-gib-rev incorporated a terminator sequence (AAAAAAAAAAGCCCGCACTGTCAGGTGCGGGCTTTTTTCTGTGTTT). Pbad-*gfpuv* was amplified using Phusion polymerase. The resulting PCR product was gel purified. The pFG1 vector was cleaved with the FspI restriction enzyme. These fragments were mixed according to the Gibson assembly. The resulting plasmid was named pFG29, containing Ptac-*araC* and Pbad-*gfpuv*.

All AraC-TAL clones were cloned into plasmid pFG29 from the pPCC423 vector for analysis after being isolated from the SLib4 library. This was done by PCR amplification of *araC* variants using primers pFG29-araC-GS and pPCC1305_araCTAL-rvs. The products and pFG29 vector were subjected to sequential digest by AflII and BstapI. The purified products were ligated using T4 DNA ligase and transformed into electrocompetent MC1061 cells. Sequencing of the final clones confirmed the correct sequences.

5.2.4) Library screening using fluorescence-activated cell sorting (FACS)

Cells were prepared for screening by first transforming 10 ng of isolated plasmid DNA into high OD electrocompetent HF19 cells harboring pPCC442. The naïve library was transformed into freshly prepared electrocompetent cells. The 1 mL transformation outgrowths were diluted by 10 in LB containing the appropriate antibiotic(s). The culture was then grown to an OD of 2 and diluted to an OD of 0.2 in LB containing the appropriate antibiotic(s) in order to dilute out the cells not harboring any plasmid. The subcultures were grown to an OD of 6 and diluted final time to an OD of 0.2 in BM containing the appropriate antibiotic(s) and 100 µM IPTG. Where indicated, cultures were supplemented with effector ligands to the following final concentrations: L-arabinose (0.1 mM), TAL (5 mM). After 6 hrs of growth, a sample of cells was washed once with PBS. The cell OD and bulk fluorescence were measured on a BMG Labtech NOVOstar spectrophotometer and a Molecular Devices GeminiEX fluorescence plate reader, respectively. The washed cells were diluted 1:100 in PBS and analyzed on a FACSJazz (BD Biosciences). The appropriate cultures were then subjected to FACS. Clones were isolated by either collecting the top 1% of the population based on the 488/520 fluorescence histogram (positive sort) or collecting the bottom population relative to the bottom 99% of the 488/520 fluorescence histogram of wt-AraC not induced (negative sort). After the sort finished, the samples were treated one of two ways: (1) samples from a negative sort were diluted 1:1 with 2xYT medium (e.g. 20 mL of sorted sample mixed with 20 mL of 2xYT), and half concentrations of antibiotics. Cultures were then incubated at 37°C 250 rpm until they reached an OD of 2. They were diluted to OD of 0.2 in LB containing the appropriate antibiotic(s). From here, samples were treated as described above for the subsequent round of sorting. (2) Samples from a positive sort were transferred to a centrifuge tube and treated according to (Ramesh et al., 2015). Briefly, the cells were pelleted by centrifugation at 17,900 xg for 10 min. The medium was discarded and the plasmid DNA was harvested using a modified protocol from a Zymo Plasmid Miniprep Kit. The

cell pellet (even if not visible) was suspended in 200 µL of PBS. The lysis buffers were adjusted accordingly and a column from the Zymo Clean and Concentration Kit was used to purify the plasmid DNA. Each sample was eluted from the column with 10 µL of Zymo elution buffer. The isolated plasmid DNA was then transformed into electrocompetent HF19 cells harboring pPCC422. The outgrowths were diluted and treated as described above for the subsequent round of sorting. After each sort, plasmid DNA was isolated from an aliquot of the subcultures prior to dilution in BM for future analysis and high-throughput sequencing.

5.2.5) Deep-well plate clone screening

Library endpoint plasmid DNA was isolated and transformed into electroporation competent HF19 cells harboring the pPCC442 reporter plasmid. Clones (24 total) were isolated from LB agar plates and streaked onto new LB agar plates. Quadruplicate 500 µL starter cultures in 2 mL 96-well (DW) plates were inoculated from each isolated clone. The starter culture was incubated for 6 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. Quadruplicate 500 µL subcultures in 96-well (DW) plate. Each sample was treated 3 ways (1) no ligand, (2) 100 µM L-ara, or (3) 5 mM TAL. The cultures were inoculated to an OD of 0.2 from the respective starter culture. The subcultures were incubated for 6 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. The cultures were washed once with 1 mL of phosphate buffered saline and OD_{595} was measured on a BMG Labtech NOVOstar spectrophotometer and the fluorescence on a Molecular Devices SpectraMax Gemini EM.

5.2.6) Deep-well plate dose responses

Isolated clones from the deep-well plate clone screening were digested and cloned into pFG29 as described above. The re-cloned mutants were transformed into electrocompetent HF19 cells. Clones were isolated from LB agar plates and quadruplicate 500 µL starter cultures in 2 mL 96-well (DW) plates were inoculated from each isolated clone. The starter culture was

incubated for 6 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. Quadruplicate 500 µL subcultures in 96-well (DW) plate containing a range of concentrations of target ligand were inoculated by 50-fold dilutions of the respective starter culture. The subcultures were incubated for 6 hrs (OD_{595} 10) at 37°C 900. The cultures were washed once with 1 mL of phosphate buffered saline and measured the OD_{595} on a BMG Labtech NOVOstar spectrophotometer and the fluorescence on a Molecular Devices SpectraMax Gemini EM.

V.3) Chapter IV

5.3.1) General

Restriction enzymes, Phusion High-Fidelity DNA polymerase, and T4 DNA ligase were purchased from New England Biolabs (Ipswich, MA). Oligonucleotides were synthesized by Integrated DNA Technologies (Coralville, IA). DNA sequencing was performed at SeqWright (Houston, TX) or Genewiz (South Plainfield, NJ). All chemicals were purchased from Sigma-Aldrich (St. Louis, MO). Molecular biology techniques for DNA manipulation were performed according to standard protocols (Sambrook and Russell, 2001) and all cultures were grown in lysogeny broth (LB). Antibiotics and IPTG were prepared as a 1000x stock solution in purified water and sterile filtered with EMD Millipore Millex-GP syringe driven filters (EMD Millipore, Cat. No. SLGP033RS).

Target compounds were all dissolved in the respective medium to 10 times the desired concentration in the culture. Compounds were titrated to a pH of 7 with NaOH and sterile filtered with EMD Millipore Millex-GP syringe driven filters. All solutions were stored at 4°C. Phloroglucinol, vanillin, and gallic acid were all stored in light resistant tubes. Target compounds and glycerol concentrations were determined by high performance liquid chromatography (HPLC) analysis using an Aminex HPX-87H ion-exclusion column (Bio-Rad Laboratories,

Hercules, CA, Cat. No. 125-0140) and detected with either an RID (Shimadzu, Cat. No. RID-10A) or UV-vis (Shimadzu, Cat. No. SPD-20A).

5.3.2) Plasmid and Library Construction

This work was based on the dual plasmid reporter system for AraC-controlled GFPuv expression that has been described previously (Tang et al., 2008), where AraC is expressed from plasmid pPCC423 (maintained by apramycin antibiotic resistance) by IPTG-inducible LacI and GFPuv is subsequently expressed from plasmid pPCC442 (chloramphenicol resistance) where it is controlled by P_{BAD}. Plasmid maps are shown in Figure D-2.

Two DNA libraries were created by site saturation mutagenesis at five residues of *araC* using overlap-extension PCR. Sites were saturated with NNS sites (N=A, T, G, or C; S=G or C). AraC library "SLib4" was constructed previously (Tang and Cirino, 2011) using plasmid pPCC423 with mutations at residue positions Pro-8,Thr-24, His-80, Tyr-82, and His-93. The second AraC library, "JLib1", was constructed in a similar fashion with mutations at positions Thr-24, Arg-38, His-80, Tyr-82, and His-93 and cloned into plasmid pFG1. First, plasmid pFG1 was created from pPCC423 by inserting additional restriction enzyme sites upstream of P_{tac} in order to facilitate future cloning projects using alternative promoters and regulatory protein genes. Primer 423mcs4-for containing the additional restriction sites was paired with primer 423mcs-rev to amplify the promoter region through a standard PCR reaction. Primer sequences are listed in Supp. Table D-1. The resulting PCR product was digested with NcoI and NdeI and ligated into pPCC423 digested with the same enzymes to yield plasmid pFG1. Sequence analysis of 10 randomly selected clones confirmed the proper incorporation of NNS sites at the desired sites.

The single plasmid system, pFG29, was cloned as follows. Primers pFG29-gib-for and pFG29-gib-rev were designed for Gibson assembly and used to amplify P_{BAD}-*gfpuv* from

pPCC442 (Phusion polymerase (NEB, Cat. No. M0530S); PCR protocol: 98°C for 30 s; 98°C for 10 s, 67.3°C for 30 s, 72°C for 45 s, 30 cycles; 72°C for 10 min). The reverse primer included a terminator sequence (antisense sequence-AAAAAAAAAAGCCCGCACTGTCAGGTGCGGGCTTTTTTC TGTGTTT) from Lynn and coworkers (Lynn et al., 1988). The amplicons were gel purified and cloned into pFG1 vector digested with FspI (NEB, Cat. No. R0135S) using the Gibson assembly master mix (NEB, Cat. No. E2611L). The reaction mixtures were cleaned and concentrated using a Zymo DNA Clean and Concentrator kit (Zymo Research Cat. No. D4006) and eluted with 10 µL of elution buffer from the kit. Each reaction (2 µL) was transformed into freshly prepared MC1061 electroporation competent cells (prepared based on protocol from Varadarajan and coworkers (Varadarajan et al., 2009)). Plasmid DNA was isolated from resulting transformants and the sequence was confirmed using Sanger sequencing with primers pFG28-for-seq and pFG28-rev-seq.

JLib1 was created as follows. Three parallel PCR reactions were performed to amplify three araC segments (A, B, and C) using the following three sets of primers: 423lib-for-NcoI and araC-T24-rev; araC-38-for-2 and araC-H80-Y82-rev; and araC-H93-for and araC-rev-4, respectively. Fragment A contained site-saturation mutations at residue position 24; fragment B contained mutations of positions 38, 80, and 82; and fragment C contained mutations for position 93. PCR products were gel purified and equimolar aliquots of adjacent DNA fragments were combined (A+B; B+C) and PCR-assembled without primers. Finally, outer primers 423lib-for-NcoI and araC-rev-4 were added to each assembly reactions and the products were PCR-amplified. The full-length products were digested with NdeI and HindIII, gel purified, and ligated into pPCC423 that had been digested and gel purified with the same enzymes (25°C for 1 hr, 65°C for 10 min). Ligation products were dialyzed on EMD Millipore MF-Membrane Filters (EMD Millipore, Cat. No. VSWP01300) and transformed into *E.coli* strain MC1061. An aliquot

of the outgrowth was diluted and streaked onto agar plates containing apr to determine the total number of transformants while the remaining outgrowth (~20 mL) was used to inoculate 480 mL LB containing antibiotic and grown overnight at 37C. Approximately 4.8×10^7 unique transformants were recovered, ensuring adequate coverage of the entire library. The plasmid library was then prepared using 100 mL of the overnight culture and ten randomly picked clones from the library were sequenced to reveal the expected random mutations at the targeted nucleotide positions. In addition, two of the ten clones contained a single, but different, point insertion between residue positions 82 and 93, which were presumably introduced by either the primers or the polymerase during the PCR process.

A third library, "CLib2" was created as follows. Random mutations were introduced into the ligand binding domain (nucleotides 11-537) of wild-type AraC using the GeneMorphII Random Mutagenesis kit from Agilent Technologies. The PCR reaction was carried out as directed using primer AraC-for5 and AraC-rev5, with the addition of 0.5 mM $MnCl_2$ in each reaction. The PCR products were ligated into the parent vector pFG1 as described above for Jlib1 with a total 1.0×10^8 transformants. Sequence analysis of 10 clones showed a mutation rate of 1.6%.

Library SLib4 was originally cloned into vector pPCC423, so we first cloned this library into pFG1 in order to have the same promoter controlling all three libraries. Primers AraC-for-5 and AraC-rev-5 were used to amplify the library from pPCC423 (Phusion polymerase; PCR protocol: 98°C for 30 s; 98°C for 10 s, 54°C for 45 s, 72°C for 60 s, 30 cycles; 72°C for 10 min). The resulting amplicon was digested with DpnI (NEB, Cat. No. R0176L) and subsequently digested with NdeI and HindIII restriction enzymes. The pFG1 vector was digested with the same restriction enzymes, NdeI and HindIII. The vector was gel purified and digested amplicon was ligated into the parent vector pFG1 as described above for Jlib1 with a total 1.3×10^8 transformants.

The JLib1, SLib4a, and CLib2 araC gene libraries contained in plasmids were subsequently cloned into the pFG29 vector, "JLib1s", "SLib4s", and 'CLib2s", respectively. Each library and pFG29 were digested with restrictions enzymes NdeI (NEB, Cat. No. R0111L) and HindIII (NEB, Cat. No. R0104L). The products were gel purified and ligated into the pFG29 vector with T4 DNA Ligase (NEB, Cat. No. M0202L). The ligations were dialyzed using EMD Millipore MF-Membrane Filters and transformed into freshly prepared MC1061 electrocompetent cells. The total transformants for each library was 1.6 x 10^8, 6.0 x 10^8, and 2.7 x 10^8 transformants, respectively. Libraries were verified by Sanger sequencing.

5.3.3) Culturing methods

For determining the catabolite and dual versus single plasmid dose responses, the deep-well plate assay described below was used. For the growth curves of the AraC-TAL biosensor in the presence of glycerol, pPCC1202 (10 ng) was transformed into HF19 electrocomp cells harboring pPCC442. The transformation outgrowths were diluted by 10 in LB supplemented with 50 µg/mL apramycin and 25 µg/mL chloramphenicol (LB+A+C). The culture was then grown to an OD of 2 and diluted to an OD_{595} 0.2 in LB supplemented with apramycin in order to dilute out the cells not harboring any plasmid. The subcultures were grown to an OD_{595} 6 and diluted a final time to an OD_{595} of 0.2 in 125 mL Erlenmeyer flasks biosensor media ("BM"; LB,1% v/v glycerol, 50 mM TES buffer, pH 7.0 with NaOH) supplemented with 50 µg/mL apramycin, 25 µg/mL chloramphenicol, 100 µM IPTG (BM+A+C+100I), with or without 5 mM TAL, and the respective concentration of glycerol (0-137 mM glycerol prepared in purified water and filtered with EMD Millipore Millex-GP syringe driven filters). Aliquots (200 µL) were taken out for each time point and washed once with phosphate buffered saline pH 7.4 (PBS). The samples were measured for OD_{595} on a BMG Labtech NOVOstar spectrophotometer and the fluorescence (Ex: 400 nm; Em: 510 nm) on a Molecular Devices SpectraMax Gemini EM.

5.3.4) Library screening

Samples were prepared for screening by transforming 10 ng of the plasmid DNA (greater than 10 ng yielded high frequency of multiple vector transformants) into HF19 high OD electroporation competent cells. The transformation outgrowths were diluted by 10 in LB containing the appropriate 50 μg/mL apramycin. The culture was then grown to an OD_{595} 2 and diluted to an OD_{595} 0.2 in LB+A in order to dilute out the cells not harboring any plasmid. The subcultures were grown to an OD_{595} 6 and diluted final time to an OD_{595} of 0.2 in BM+A+100I and the appropriate concentration of the target compound in 125 mL Erlenmeyer flask. Cells were harvested after 6 hrs of incubation at 37°C 250 rpm (OD_{595} ~10) and washed once with an equal volume of PBS. The cell OD and bulk fluorescence were measured on a BMG Labtech NOVOstar spectrophotometer and a Molecular Devices GeminiEX fluorescence plate reader, respectively. The washed cells were diluted 1:100 in PBS and analyzed on a FACSJazz (BD Biosciences). The appropriate cultures were then subjected to FACS. Clones were isolated by either collecting the top 1% of the population based on the 488/520 fluorescence histogram (positive sort) or collecting the bottom population relative to the bottom 99% of the GFPuv fluorescence histogram of wild-type AraC not induced (negative sort). After the sort finished, the samples were treated one of two ways: (1) samples from a negative sort were diluted 1:1 with 2xYT medium (e.g. 20 mL of sorted sample mixed with 20 mL of 2xYT), and half concentrations of antibiotics. Cultures were then incubated at 37°C 250 rpm until they reached an OD of 2. They were diluted to OD_{595} 0.2 in LB supplemented with apramycin. From here, samples were treated as described above for the subsequent round of sorting. (2) Samples from a positive sort were transferred to a centrifuge tube and treated according to (Ramesh et al., 2015). Briefly, the cells were pelleted by centrifugation at 17,900 xg for 10 min. The medium was discarded and the plasmid DNA was harvested using a modified protocol from a Zymo Plasmid Miniprep Kit. The cell pellet (even if not visible) was suspended in 200 μL of PBS. The lysis buffers were adjusted

accordingly and a column from the Zymo Clean and Concentration Kit was used to purify the plasmid DNA. Each sample was eluted from the column with 10 µL of Zymo elution buffer. The isolated plasmid DNA was then transformed into electroporation competent HF19 cells harboring pPCC422. The outgrowths were diluted and treated as described above for the subsequent round of sorting. After each sort, plasmid DNA was isolated from an aliquot of the subcultures prior to dilution in BM for future analysis and high-throughput sequencing. Selections were carried out by transforming 10 ng of the respective population into SQ12 high OD electroporation competent cells. The outgrowths were either diluted as described above for FACS or directly plated on LB plates supplemented with 1% glycerol, 50 µg/mL apramycin, 100-300 µg/mL ampicillin, and 100 µM IPTG as well as the respective compound. For the former protocol, cells were plated prior to being diluted in BM. All selection plates were grown at 37°C until colonies were visible and still spatially separated. The DNA was isolated from the selection plates by scraping the plates with 2-5 mL of purified water. The cells were pelleted and the plasmid DNA was extracted using a QIAprep Spin Miniprep kit (QIAGEN, Cat. No. 27106). The plasmid DNA was then transformed into the appropriate strain for the subsequent round of screening.

5.3.5) Deep-Well Plate Clone Screening

Library endpoint plasmid DNA was isolated and transformed into electroporation competent HF19 cells. Clones were isolated from LB agar plates and streaked onto new LB agar plates. Quadruplicate 500 µL LB supplemented with apramycin starter cultures in 2 mL 96-well (DW) plates were inoculated from each isolated clone. The starter culture was incubated for 6 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. Quadruplicate 500 µL BM supplemented with apramycin and 100 µM IPTG subcultures in deep-well plates containing the appropriate target ligand were inoculated by 50-fold dilutions of the respective starter culture (OD_{595} ~0.2). The subcultures were incubated for 6 hrs at 37°C 900 rpm in a Heidolph Titramaz

1000/Inkubator 1000. The cultures were washed once with 1 mL of phosphate buffered saline and measured the OD_{595} on a BMG Labtech NOVOstar spectrophotometer and the fluorescence on a Molecular Devices SpectraMax Gemini EM.

5.3.6) Deep-Well Plate dose responses

Isolated clones from the deep-well plate clone screening were digested and cloned back into the parent vector using the same method described above for cloning the libraries into the pFG29 vector. The re-cloned mutants were transformed into electroporation competent HF19 cells. Clones were isolated from LB agar plates supplemented with apramycin and quadruplicate 500 µL LB supplemented with apramycin starter cultures in 2 mL deep-well plates were inoculated from each isolated clone. The starter culture was incubated for 6 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. Quadruplicate 500 µL BM supplemented with apramycin and 100 µM IPTG subcultures in 96-well (DW) plate containing a range of target ligand were inoculated by 50-fold dilutions of the respective starter culture. The subcultures were incubated for 16 hrs at 37°C 900 rpm in a Heidolph Titramaz 1000/Inkubator 1000. The cultures were washed once with 1 mL of phosphate buffered saline and measured the OD595 on a BMG Labtech NOVOstar spectrophotometer and the fluorescence on a Molecular Devices SpectraMax Gemini EM.

REFERENCES

Aharoni A, Amitai G, Bernath K, Magdassi S, Tawfik DS. 2005a. High-throughput screening of enzyme libraries: Thiolactonases evolved by fluorescence-activated sorting of single cells in emulsion compartments. *Chemistry & Biology* 12(12):1281-1289.

Aharoni A, Griffiths AD, Tawfik DS. 2005b. High-throughput screens and selections of enzyme-encoding genes. *Current Opinion in Chemical Biology* 9(2):210-216.

Alper H, Fischer C, Nevoigt E, Stephanopoulos G. 2005. Tuning genetic control through promoter engineering. *Proceedings of the National Academy of Sciences of the United States of America* 102(36):12678-12683.

Andersen JB, Sternberg C, Poulsen LK, Bjorn SP, Givskov M, Molin S. 1998. New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Applied and Environmental Microbiology* 64(6):2240-2246.

Arnold FH. 1996. Directed evolution: Creating biocatalysts for the future. *Chemical Engineering Science* 51(23):5091-5102.

Asai T, Tsukada K, Ise S, Shirata N, Hashimoto M, Fujii I, Gomi K, Nakagawara K, Kodama EN, Oshima Y. 2015. Use of a biosynthetic intermediate to explore the chemical diversity of pseudo-natural fungal polyketides. *Nat Chem* advance online publication.

Baeshen NA, Baeshen MN, Sheikh A, Bora RS, Ahmed MMM, Ramadan HAI, Saini KS, Redwan EM. 2014. Cell factories for insulin production. *Microbial Cell Factories* 13(1):141.

Bailey JE. 1991. Toward a science of metabolic engineering. *Science* 252(5013):1668-1675.

Barghini P, Di Gioia D, Fava F, Ruzzi M. 2007. Vanillin production using metabolically engineered Escherichia coli under non-growing conditions. *Microbial Cell Factories* 6(1):13-13.

Beggah S, Vogne C, Zenaro E, Meer JRvd. 2008. Mutant HbpR transcription activator isolation for 2-chlorobiphenyl via green fluorescent protein-based flow cytometry and cell sorting. *Microbial biotechnology* 1(1):68-78.

Bentley WE, Mirjalili N, Andersen DC, Davis RH, Kompala DS. 1990. Plasmid-encoded protein - the principal factor in the metabolic burden associated with recombinant bacteria. *Biotechnology and Bioengineering* 35(7):668-681.

Bettenbrock K, Sauter T, Jahreis K, Kremling A, Lengeler JW, Gilles E-D. 2007. Correlation between growth rates, EIIA(Crr) phosphorylation, and intracellular cyclic AMP levels in Escherichia coli K-12. *Journal of Bacteriology* 189(19):6891-6900.

Bi HK, Zhu L, Wang HH, Cronan JE. 2014. Inefficient Translation Renders the Enterococcus faecalis fabK Enoyl-Acyl Carrier Protein Reductase Phenotypically Cryptic. *Journal of Bacteriology* 196(1):170-179.

Bienick MS, Young KW, Klesmith JR, Detwiler EE, Tomek KJ, Whitehead TA. 2014. The Interrelationship between Promoter Strength, Gene Expression, and Growth Rate. *PLOS One* 9(10).

Binz TM, Wenzel SC, Schnell H-J, Bechthold A, Müller R. 2008. Heterologous Expression And Genetic Engineering of the Phenalinolactone Biosynthetic Gene Cluster by Using Red/ET Recombineering. *Chembiochem* 9(3):447-454.

Birnbaum S, Bailey JE. 1991. Plasmid presence changes the relative levels of many host-cell proteins and ribosome components in recombinant Escherichia coli. *Biotechnology and Bioengineering* 37(8):736-745.

Boersma YL, Droge MJ, van der Sloot AM, Pijning T, Cool RH, Dijkstra BW, Quax WJ. 2008. A novel genetic selection system for improved enantioselectivity of Bacillus subtilis lipase A. *Chembiochem* 9(7):1110-1115.

Bonella S, Raimondo D, Milanetti E, Tramontano A, Ciccotti G. 2014. Mapping the hydropathy of amino acids based on their local solvation structure. *The Journal of Physical Chemistry.B* 118(24):6604.

Bornscheuer UT, Altenbuchner J, Meyer HH. 1998. Directed evolution of an esterase for the stereoselective resolution of a key intermediate in the synthesis of epothilones. *Biotechnology and Bioengineering* 58(5):554-559.

Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K. 2012. Engineering the third wave of biocatalysis. *Nature* 485(7397):185-194.

Bosley AD, Ostermeier M. 2005. Mathematical expressions useful in the construction, description and evaluation of protein libraries. *Biomolecular Engineering* 22(1-3):57-61.

Bustos SA, Schleif RF. 1993. Functional domains of the AraC protein. *Proceedings of the National Academy of Sciences of the United States of America* 90(12):5638-5642.

Causey TB, Zhou S, Shanmugam KT, Ingram LO. 2003. Engineering the metabolism of Escherichia coli W3110 for the conversion of sugar to redox-neutral and oxidized products: Homoacetate production. *Proceedings of the National Academy of Sciences of the United States of America* 100(3):825-832.

Cebolla A, Sousa C, deLorenzo V. 1997. Effector specificity mutants of the transcriptional activator NahR of naphthalene degrading Pseudomonas define protein sites involved in binding of aromatic inducers. *Journal of Biological Chemistry* 272(7):3986-3992.

Changeux JP, Edelstein SJ. 2005. Allosteric mechanisms of signal transduction. *Science* 308(5727):1424-1428.

Chen W, Zhang S, Jiang PX, Yao J, He YZ, Chen LC, Gui XW, Dong ZY, Tang SY. 2015. Design of an ectoine-responsive AraC mutant and its application in metabolic engineering of ectoine biosynthesis. *Metabolic Engineering* 30:149-155.

Chia M, Schwartz TJ, Shanks BH, Dumesic JA. 2012. Triacetic acid lactone as a potential biorenewable platform chemical. *Green Chemistry* 14(7):1850-1853.

Chin JW, Khankal R, Monroe CA, Maranas CD, Cirino PC. 2009. Analysis of NADPH supply during xylitol production by engineered Escherichia coli. *Biotechnol Bioeng* 102(1):209-20.

Choi YJ, Lee SY. 2013. Microbial production of short-chain alkanes. *Nature* 502(7472):571.

Chusacultanachai S, Glenn KA, Rodriguez AO, Read EK, Gardner JF, Katzenellenbogen BS, Shapiro DJ. 1999. Analysis of Estrogen Response Element Binding by Genetically Selected Steroid Receptor DNA Binding Domain Mutants Exhibiting Altered Specificity and Enhanced Affinity. *Journal of Biological Chemistry* 274(33):23591-23598.

Cirino PC, Mayer KM, Umeno D. 2003. Generating Mutant Libraries Using Error-Prone PCR. In: Arnold FH, Georgiou G, editors. Directed Evolution Library Creation. Humana Press. p 3-9.

Clackson T, Wells JA. 1995. A hot-spot of binding-energy in a hormone-receptor interface. *Science* 267(5196):383-386.

Cobb RE, Chao R, Zhao HM. 2013. Directed Evolution: Past, Present, and Future. *Aiche Journal* 59(5):1432-1440.

Coco WM, Levinson WE, Crist MJ, Hektor HJ, Darzins A, Pienkos PT, Squires CH, Monticello DJ. 2001. DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nature Biotechnology* 19(4):354-359.

Cohen SN, Boyer HW, Helling RB. 1973. Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proceedings of the National Academy of Sciences of the United States of America* 70(11):3240-3244.

Condon N, Klemick H, Wolverton A. 2015. Impacts of ethanol policy on corn prices: A review and meta-analysis of recent evidence. *Food Policy* 51(0):63-73.

Cox RJ, Gibson JS, Mayo Martín MB. 2002. Aspartyl Phosphonates and Phosphoramidates: The First Synthetic Inhibitors of Bacterial Aspartate-Semialdehyde Dehydrogenase. *Chembiochem* 3(9):874-886.

Cunningham BC, Wells JA. 1989. High-resolution epitope mapping of high-receptor interactions by alanine-scanning mutagenesis. *Science* 244(4908):1081-1085.

Damjanovic A, Miller BT, Schleif R. 2013. Understanding the basis of a class of paradoxical mutations in AraC through simulations. *Proteins: Structure, Function, and Bioinformatics* 81(3):490-498.

Daruwalla KR, Paxton AT, Henderson PJ. 1981. Energization of the transport systems for arabinose and comparison with galactose transport in Escherichia coli. *Biochemical Journal* 200(3):611-627.

Datsenko KA, Wanner BL. 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A* 97(12):6640-5.

Daugherty PS, Chen G, Iverson BL, Georgiou G. 2000. Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proceedings of the National Academy of Sciences of the United States of America* 97(5):2029-2034.

de las Heras A, Carreno CA, de Lorenzo V. 2008. Stable implantation of orthogonal sensor circuits in Gram-negative bacteria for environmental release. *Environmental Microbiology* 10(12):3305-3316.

de las Heras A, Carreño CA, Martínez-García E, De Lorenzo V. 2010. Engineering input/output nodes in prokaryotic regulatory circuits. *Fems Microbiology Reviews* 34(5):842-865.

de las Heras A, de Lorenzo V. 2011a. Cooperative amino acid changes shift the response of the sigma 54-dependent regulator XylR from natural m-xylene towards xenobiotic 2,4-dinitrotoluene. *Molecular Microbiology* 79(5):1248-1259.

de las Heras A, de Lorenzo V. 2011b. In situ detection of aromatic compounds with biosensor Pseudomonas putida cells preserved and delivered to soil in water-soluble gelatin capsules. *Analytical and Bioanalytical Chemistry* 400(4):1093-1104.

Denard CA, Ren HQ, Zhao HM. 2015. Improving and repurposing biocatalysts via directed evolution. *Current Opinion in Chemical Biology* 25:55-64.

Dietrich JA, McKee AE, Keasling JD. 2010. High-throughput metabolic engineering: advances in small-molecule screening and selection. Palo Alto: Annual Reviews. 563-590 p.

Dirla S, Chien JY-H, Schleif R. 2009. Constitutive Mutations in the Escherichia coli AraC Protein. *Journal of Bacteriology* 191(8):2668.

Dixon RA, Steele CL. 1999. Flavonoids and isoflavonoids - a gold mine for metabolic engineering. *Trends in Plant Science* 4(10):394-400.

Domach MM. 2015. Perspectives and Prospects for Whole Cell Catalysis. *Catalysis Letters* 145(1):346-359.

Drummond DA, Iverson BL, Georgiou G, Arnold FH. 2005. Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *Journal of Molecular Biology* 350(4):806-816.

Eggeling L, Bott M, Marienhagen J. 2015. Novel screening methods — biosensors. *Current Opinion in Biotechnology* 35:30-36.

Eggert US. 2013. The why and how of phenotypic small-molecule screens. *Nature Chemical Biology* 9(4):206-209.

Eustance RJ, Bustos SA, Schleif RF. 1994. Locating and lengthening the interdomain linker in AraC protein. *Journal of Molecular Biology* 242(4):330-338.

Fache M, Boutevin B, Caillol S. 2015. Vanillin, a key-intermediate of biobased polymers. *European Polymer Journal* 68(0):488-502.

Feldhaus MJ, Siegel RW, Opresko LK, Coleman JR, Feldhaus JMW, Yeung YA, Cochran JR, Heinzelman P, Colby D, Swers J and others. 2003. Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. *Nature Biotechnology* 21(2):163-170.

Fiet SvS, Beilen JBv, Witholt B. 2006. Selection of Biocatalysts for Chemical Synthesis. *Proceedings of the National Academy of Sciences of the United States of America* 103(6):1693-1698.

Foster PL. 2007. Stress-Induced Mutagenesis in Bacteria. *Critical reviews in biochemistry and molecular biology* 42(5):373-397.

Fowler CC, Brown ED, Li YF. 2008. A FACS-based approach to engineering artificial riboswitches. *Chembiochem* 9(12):1906-1911.

Fu J, Bian X, Hu S, Wang H, Huang F, Seibert PM, Plaza A, Xia L, Müller R, Stewart AF and others. 2012. Full-length RecE enhances linear-linear homologous recombination and facilitates direct cloning for bioprospecting. *Nature Biotechnology* 30(5):440.

Furuya T, Miura M, Kino K. 2014. A Coenzyme-Independent Decarboxylase/Oxygenase Cascade for the Efficient Synthesis of Vanillin. *Chembiochem* 15(15):2248-2254.

Furuya T, Miura M, Kuroiwa M, Kino K. 2015. High-yield production of vanillin from ferulic acid by a coenzyme-independent decarboxylase/oxygenase two-stage process. *New biotechnology* 32(3):335.

Galvao TC, Mencia M, de Lorenzo V. 2007. Emergence of novel functions in transcriptional regulators by regression to stem protein types. *Molecular Microbiology* 65(4):907-919.

Gao X, Wang P, Tang Y. 2010. Engineered polyketide biosynthesis and biocatalysis in Escherichia coli. *Applied Microbiology and Biotechnology* 88(6):1233-1242.

Garmendia J, Devos D, Valencia A, de Lorenzo V. 2001. A la carte transcriptional regulators: unlocking responses of the prokaryotic enhancer-binding protein XyIR to non-natural effectors. *Molecular Microbiology* 42(1):47-59.

Gill RT, Valdes JJ, Bentley WE. 2000. A Comparative Study of Global Stress Gene Regulation in Response to Overexpression of Recombinant Proteins in Escherichia coli. *Metabolic Engineering* 2(3):178-189.

Golynskiy MV, Koay MS, Vinkenborg JL, Merkx M. 2011. Engineering Protein Switches: Sensors, Regulators, and Spare Parts for Biology and Biotechnology. *Chembiochem* 12(3):353-361.

Gomez-Escribano JP, Bibb MJ. 2011. Engineering Streptomyces coelicolor for heterologous expression of secondary metabolite gene clusters. *Microbial Biotechnology* 4(2):207-215.

Goodey NM, Benkovic SJ. 2008. Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol* 4(8):474-482.

Gounaris Y. 2010. Biotechnology for the production of essential oils, flavours and volatile isolates. A review. *Flavour and Fragrance Journal* 25(5):367-386.

Gredell JA, Frei CS, Cirino PC. 2012. Protein and RNA engineering to customize microbial molecular reporting. *Biotechnology Journal* 7(4):477-499.

Guntas G, Mansell TJ, Kim JR, Ostermeier M. 2005. Directed evolution of protein switches and their application to the creation of ligand-binding proteins. *Proc Natl Acad Sci U S A* 102(32):11224-9.

Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. *Trends in biotechnology* 22(7):346-353.

Guzman LM, Belin D, Carson MJ, Beckwith J. 1995. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *Journal of Bacteriology* 177(14):4121-4130.

Haldimann A, Wanner BL. 2001. Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. *Journal of Bacteriology* 183(21):6384.

Hamamatsu N, Nomiya Y, Aita T, Nakajima M, Husimi Y, Shibanaka Y. 2006. Directed evolution by accumulating tailored mutations: Thermostabilization of lactate oxidase with less trade-off with catalytic activity. *Protein Engineering Design & Selection* 19(11):483-489.

Hansen LH, Sorensen SJ. 2001. The use of whole-cell biosensors to detect and quantify compounds or conditions affecting biological systems. *Microbial Ecology* 42(4):483-494.

Hara KY, Araki M, Okai N, Wakai S, Hasunuma T, Kondo A. 2014. Development of bio-based fine chemical production through synthetic bioengineering. *Microbial Cell Factories* 13(1):173-173.

Hertweck C. 2015. Decoding and reprogramming complex polyketide assembly lines: prospects for synthetic biology. *Trends in Biochemical Sciences* 40(4):189-199.

Jares-Erijman EA, Jovin TM. 2003. FRET imaging. *Nature Biotechnology* 21(11):1387-1395.

Keasling JD. 2010. Manufacturing Molecules Through Metabolic Engineering. *Science* 330(6009):1355-1358.

Khatri W, Hendrix R, Niehaus T, Chappell J, Curtis WR. 2014. Hydrocarbon Production in High Density Botryococcus braunii Race B Continuous Culture. *Biotechnology and Bioengineering* 111(3):493-503.

Khlebnikov A, Datsenko KA, Skaug T, Wanner BL, Keasling JD. 2001. Homogeneous expression of the P-BAD promoter in Escherichia coli by constitutive expression of the low-affinity high-capacity AraE transporter. *Microbiology-Sgm* 147:3241-3247.

Kille S, Acevedo-Rocha CG, Parra LP, Zhang ZG, Opperman DJ, Reetz MT, Acevedo JP. 2013. Reducing Codon Redundancy and Screening Effort of Combinatorial Protein Libraries Created by Saturation Mutagenesis. *Acs Synthetic Biology* 2(2):83-92.

Kolodrubetz D, Schleif R. 1981. L-arabinose transport systems in Escherichia coli K-12. *Journal of Bacteriology* 148(2):472-479.

Krebsfanger N, Schierholz K, Bornscheuer UT. 1998a. Enantioselectivity of a recombinant esterase from Pseudomonas fluorescens towards alcohols and carboxylic acids. *Journal of Biotechnology* 60(1-2):105-111.

Krebsfanger N, Zocher F, Altenbuchner J, Bornscheuer UT. 1998b. Characterization and enantioselectivity of a recombinant esterase from Pseudomonas fluorescens. *Enzyme and Microbial Technology* 22(7):641-646.

Kunichika K, Hashimoto Y, Imoto T. 2002. Robustness of hen lysozyme monitored by random mutations. *Protein Engineering* 15(10):805-809.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* 157(1):105-132.

Lan EI, Liao JC. 2011. Metabolic engineering of cyanobacteria for 1-butanol production from carbon dioxide. *Metabolic Engineering* 13(4):353-363.

Larionov V, Kouprina N. 2008. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast Saccharomyces cerevisiae. *Nature Protocols* 3(3):371-377.

Leonard E, Yan Y, Fowler ZL, Li Z, Lim CG, Lim KH, Koffas MAG. 2008. Strain improvement of recombinant Escherichia coli for efficient production of plant flavonoids. *Molecular Pharmaceutics* 5(2):257-265.

Lerner SA, Wu TT, Lin ECC. 1964. Evolution of a Catabolic Pathway in Bacteria. *Science* 146(3649):1313-1315.

Levintha.C. 1968. Are there pathways for protein folding. *Journal De Chimie Physique Et De Physico-Chimie Biologique* 65(1):44-&.

Li CK, Wen AY, Shen BC, Lu J, Huang Y, Chang YC. 2011. FastCloning: a highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. *Bmc Biotechnology* 11.

Link AJ, Jeong KJ, Georgiou G. 2007. Beyond toothpicks: new methods for isolating mutant bacteria. *Nature Reviews Microbiology* 5(9):680-688.

Liu LF, Li YF, Liotta D, Lutz S. 2009. Directed evolution of an orthogonal nucleoside analog kinase via fluorescence-activated cell sorting. *Nucleic Acids Research* 37(13):4472-4481.

Liu Q, Wu KY, Cheng YB, Lu L, Xiao ET, Zhang YC, Deng ZX, Liu TG. 2015. Engineering an iterative polyketide pathway in Escherichia coli results in single-form alkene and alkane overproduction. *Metabolic Engineering* 28:82-90.

Lobell RB, Schleif RF. 1990. DNA Looping and Unlooping by AraC Protein. *Science* 250(4980):528-532.

Luo YZ, Cobb RE, Zhao HM. 2014. Recent advances in natural product discovery. *Current Opinion in Biotechnology* 30:230-237.

Lutz S, Ostermeier M, Moore GL, Maranas CD, Benkovic SJ. 2001. Creating Multiple-Crossover DNA Libraries Independent of Sequence Identity. *Proceedings of the National Academy of Sciences of the United States of America* 98(20):11248-11253.

Lynn SP, Kasper LM, Gardner JF. 1988. Contributions of RNA secondary structure and length of the thymidine tract to transcription termination at the thr operon attenuator. *Journal of Biological Chemistry* 263(1):472.

Martin K, Huo L, Schleif RF. 1986. The DNA loop model for *ara* repression - AraC protein occupies the proposed loop sites in vivo and repression-negative mutations lie in these same sites. *Proceedings of the National Academy of Sciences of the United States of America* 83(11):3654-3658.

Martínez-Gómez K, Flores N, Castañeda HM, Martínez-Batallar G, Hernández-Chávez G, Ramírez OT, Gosset G, Encarnación S, Bolivar F. 2012. New insights into Escherichia coli metabolism: carbon scavenging, acetate metabolism and carbon recycling responses during growth on glycerol. *Microbial Cell Factories* 11(1):46-46.

Matsui D, Okazaki S, Matsuda M, Asano Y. 2015. Enhancement of stability of L-tryptophan dehydrogenase from Nostoc punctiforme ATCC29133 and its application to L-tryptophan assay. *Journal of Biotechnology* 196:27-32.

Michener JK, Smolke CD. 2012. High-throughput enzyme evolution in Saccharomyces cerevisiae using a synthetic RNA switch. *Metabolic Engineering* 14(4):306-316.

Michener JK, Thodey K, Liang JC, Smolke CD. 2012. Applications of genetically-encoded biosensors for the construction and control of biosynthetic pathways. *Metabolic Engineering* 14(3):212-222.

Nazor J, Schwaneberg U. 2006. Laboratory evolution of P450BM-3 for mediated electron transfer. *Chembiochem* 7(4):638-644.

Neidhardt FC, American Society for M. 1987. Escherichia coli and Salmonella typhimurium: cellular and molecular biology. Washington, D.C: American Society for Microbiology.

Ogden S, Haggerty D, Stoner CM, Kolodrubetz D, Schleif R. 1980. The Escherichia coli L-arabinose operon: binding sites of the regulatory proteins and a mechanism of positive and negative regulation. *Proceedings of the National Academy of Sciences of the United States of America* 77(6):3346-3350.

Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. 2007. Crystal structure of an ancient protein: Evolution by conformational epistasis. *Science* 317(5844):1544-1548.

Packer MS, Liu DR. 2015. Methods for the directed evolution of proteins. *Nat Rev Genet* 16(7):379-394.

Pareja E, Pareja-Tobes P, Manrique M, Pareja-Tobes E, Bonal J, Tobes R. 2006. ExtraTrain: a database of Extragenic regions and Transcriptional information in prokaryotic organisms. *Bmc Microbiology* 6.

Petrovič U. 2015. Next-generation biofuels: a new challenge for yeast. *Yeast* 32(9):583-593.

Phillips GJ, Park SK, Huber D. 2000. High copy number plasmids compatible with commonly used cloning vectors. *Biotechniques* 28(3):400-+.

Ramesh B, Frei CS, Cirino PC, Varadarajan N. 2015. A method for direct plasmid recovery after FACS screening. *Biotechniques* (Submitted).

Reed B, Blazeck J, Alper H. 2012. Evolution of an alkane-inducible biosensor for increased responsiveness to short-chain alkanes. *Journal of Biotechnology* 158(3):75-79.

Reed WL, Schleif RF. 1999. Hemiplegic mutations in AraC protein. *Journal of Molecular Biology* 294(2):417-425.

Reetz MT, Kahakeaw D, Lohmer R. 2008. Addressing the numbers problem in directed evolution. *Chembiochem : a European journal of chemical biology* 9(11):1797-1804.

Rodgers ME, Schleif R. 2009. Solution structure of the DNA binding domain of AraC protein. *Proteins-Structure Function and Bioinformatics* 77(1):202-208.

Rodgers ME, Schleif R. 2012. Heterodimers Reveal That Two Arabinose Molecules Are Required for the Normal Arabinose Response of AraC. *Biochemistry* 51(41):8085-8091.

Roessner CA, Scott AI. 1996. Genetically engineered synthesis of natural products: From alkaloids to corrins. *Annual Review of Microbiology* 50:467-490.

Romero PA, Arnold FH. 2009. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology* 10(12):866-876.

Ross JJ, Gryczynski U, Schleif R. 2003. Mutational analysis of residue roles in AraC function. *J Mol Biol* 328(1):85-93.

Salehi Jouzani G, Taherzadeh MJ. 2015. Advances in consolidated bioprocessing systems for bioethanol and butanol production from biomass: a comprehensive review. *Biofuel Research Journal* 2(1):152-195.

Salis HM, Mirsky EA, Voigt CA. 2009. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* 27(10):946-U112.

Sambrook J, Russell DW. 2001. Molecular cloning: a laboratory manual. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press.

Santoro SW, Wang L, King DS, Herberich B, Schultz PG. 2002. An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. *Nature Biotechnology* 20(10):1044-1048.

Saviola B, Seabold R, Schleif RF. 1998. Arm-domain interactions in AraC. *J Mol Biol* 278(3):539-48.

Schallmey M, Frunzke J, Eggeling L, Marienhagen J. 2014. Looking for the pick of the bunch: high-throughput screening of producing microorganisms with biosensors. *Current Opinion in Biotechnology* 26:148-154.

Schell MA, Poser EF. 1989. Demonstration, characterization, and mutational analysis of NahR protein-binding to nah and sal promoters. *Journal of Bacteriology* 171(2):837-846.

Schleif R. 1969. An l-arabinose binding protein and arabinose permeation in Escherichia coli. *Journal of Molecular Biology* 46(1):185-196.

Schleif R. 2010. AraC protein, regulation of the l-arabinose operon in Escherichia coli, and the light switch mechanism of AraC action. *Fems Microbiology Reviews* 34(5):779-796.

Schmidt-Dannert C, Arnold FH. 1999. Directed evolution of industrial enzymes. *Trends in biotechnology* 17(4):135-6.

Schwaneberg U, Appel D, Schmitt J, Schmid RD. 2000. P450 in biotechnology: zinc driven omega-hydroxylation of p-nitrophenoxydodecanoic acid using P450BM-3 F87A as a catalyst. *Journal of Biotechnology* 84(3):249-257.

Schwaneberg U, Otey C, Cirino PC, Farinas E, Arnold FH. 2001. Cost-Effective Whole-Cell Assay for Laboratory Evolution of Hydroxylases in Escherichia coli. *Journal of Biomolecular Screening* 6(2):111-117.

Schwaneberg U, Schmidt-Dannert C, Schmitt J, Schmid RD. 1999. A continuous spectrophotometric assay for P450 BM-3, a fatty acid hydroxylating enzyme, and its mutant F87A. *Analytical Biochemistry* 269(2):359-366.

Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. 2010. Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science* 330(6007):1099-1102.

Seabold RR, Schleif RF. 1998. Apo-AraC actively seeks to loop. *Journal of Molecular Biology* 278(3):529-538.

Seitz T, Thoma R, Schoch GA, Stihle M, Benz J, D'Arcy B, Wiget A, Ruf A, Hennig M, Sterner R. 2010. Enhancing the Stability and Solubility of the Glucocorticoid Receptor Ligand-Binding Domain by High-Throughput Library Screening. *Journal of Molecular Biology* 403(4):562-577.

Shaner NC, Patterson GH, Davidson MW. 2007. Advances in fluorescent protein technology. *Journal of Cell Science* 120(24):4247-4260.

Shaner NC, Steinbach PA, Tsien RY. 2005. A guide to choosing fluorescent proteins. *Nature Methods* 2(12):905-909.

Shapiro HM. 2003. Practical flow cytometry. Hoboken, N.J: Wiley-Liss.

Shen B. 2003. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology* 7(2):285-295.

Shih SCC, Goyal G, Kim PW, Koutsoubelis N, Keasling JD, Adams PD, Hillson NJ, Singh AK. 2015. A Versatile Microfluidic Device for Automating Synthetic Biology. *Acs Synthetic Biology*.

Siegele DA, Hu JC. 1997. Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proceedings of the National Academy of Sciences of the United States of America* 94(15):8168-8172.

Skarfstad E, O'Neill E, Garmendia J, Shingler V. 2000. Identification of an effector specificity subregion within the aromatic-responsive regulators DmpR and XylR by DNA shuffling. *Journal of Bacteriology* 182(11):3008-3016.

Soisson SM, MacDougallShackleton B, Schleif R, Wolberger C. 1997. Structural basis for ligand-regulated oligomerization of AraC. *Science* 276(5311):421-425.

Stanton BC, Nielsen AAK, Tamsir A, Clancy K, Peterson T, Voigt CA. 2014. Genomic mining of prokaryotic repressors for orthogonal logic gates. *Nature Chemical Biology* 10(2):99-105.

Stemmer WPC. 1994a. DNA shuffling by random fragmentation and reassembly- In vitro recombination for molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* 91(22):10747-10751.

Stemmer WPC. 1994b. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* 370(6488):389-391.

Stephanopoulos G, Vallino JJ. 1991. Network rigidity and metabolic engineering in metabolite overproduction. *Science* 252(5013):1675-1681.

Sullivan B, Walton AZ, Stewart JD. 2013. Library construction and evaluation for site saturation mutagenesis. *Enzyme and Microbial Technology* 53(1):70.

Sun YS, Day RN, Periasamy A. 2011. Investigating protein-protein interactions in living cells using fluorescence lifetime imaging microscopy. *Nature Protocols* 6(9):1324-1340.

Takase K, Taguchi S, Doi Y. 2003. Enhanced synthesis of poly(3-hydroxybutyrate) in recombinant Escherichia coli by means of error-prone PCR mutagenesis, saturation mutagenesis, and in vitro recombination of the type II polyhydroxyalkanoate synthase gene. *Journal of Biochemistry* 133(1):139-145.

Tang S-Y, Cirino PC. 2011. Design and application of a mevalonate-responsive regulatory protein. *Angewandte Chemie International Edition* 50(5):1084-1086.

Tang S-Y, Qian S, Akinterinwa O, Frei CS, Gredell JA, Cirino PC. 2013. Screening for enhanced triacetic acid lactone production by recombinant Escherichia coli expressing a designed triacetic acid lactone reporter. *Journal of the American Chemical Society* 135(27):10099-10103.

Tang SY, Cirino PC. 2010. Elucidating residue roles in engineered variants of AraC regulatory protein. *Protein Science* 19(2):291-298.

Tang SY, Fazelinia H, Cirino PC. 2008. AraC regulatory protein mutants with altered effector specificity. *Journal of the American Chemical Society* 130(15):5267-5271.

Telford WG, Hawley T, Subach F, Verkhusha V, Hawley RG. 2012. Flow cytometry of fluorescent proteins. *Methods* 57(3):318-330.

Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL and others. 2013. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* 501(7466):212-216.

Tokuriki N, Tawfik DS. 2009. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* 19(5):596-604.

Umeno D, Tobias AV, Arnold FH. 2005. Diversifying carotenoid biosynthetic pathways by directed evolution. *Microbiol Mol Biol Rev* 69(1):51-78.

Urvoas A, Valerio-Lepiniec M, Minard P. 2012. Artificial proteins from combinatorial approaches. *Trends in biotechnology* 30(10):512-520.

Valle-Rodriguez JO, Shi SB, Siewers V, Nielsen J. 2014. Metabolic engineering of Saccharomyces cerevisiae for production of fatty acid ethyl esters, an advanced biofuel, by eliminating non-essential fatty acid utilization pathways. *Applied Energy* 115:226-232.

Van Dilla MA. 1985. Flow cytometry: instrumentation and data analysis. London; Orlando, FL: Academic Press.

Varadarajan N, Georgiou G, Cantor JR, Iverson BL. 2009. Construction and flow cytometric screening of targeted enzyme libraries. *Nature Protocols* 4(6):893-901.

Waegeman H, Soetaert W. 2011. Increasing recombinant protein production in Escherichia coli through metabolic and genetic engineering. *Journal of Industrial Microbiology & Biotechnology* 38(12):1891-1910.

Walsh G. 2014. Biopharmaceutical benchmarks 2014. *Nature Biotechnology* 32(10):992-1000.

Walton NJ, Narbad A, Faulds CB, Williamson G. 2000. Novel approaches to the biosynthesis of vanillin. *Current Opinion in Biotechnology* 11(5):490-496.

Wang M, Si T, Zhao H. 2012. Biocatalyst Development by Directed Evolution. *Bioresource Technology* 115C:117-125.

Weber T, Charusanti P, Musiol-Kroll EM, Jiang X, Tong Y, Kim HU, Lee SY. 2015. Metabolic engineering of antibiotic factories: new tools for antibiotic production in actinomycetes. *Trends in biotechnology* 33(1):15-26.

Wegener G, Brandt M, Duda L, Hofmann J, Klesczewski B, Koch D, Kumpf R-J, Orzesek H, Pirkl H-G, Six C and others. 2001. Trends in industrial catalysis in the polyurethane industry. *Applied Catalysis A: General* 221(1-2):303-335.

Weiss GA, Watanabe CK, Zhong A, Goddard A, Sidhu SS. 2000. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proceedings of the National Academy of Sciences of the United States of America* 97(16):8950-8954.

Weldon JE, Rodgers ME, Larkin C, Schleif RF. 2007. Structure and properties of a truely apo form of AraC dimerization domain. *Proteins-Structure Function and Bioinformatics* 66(3):646-654.

Wenda S, Illner S, Mell A, Kragl U. 2011. Industrial biotechnology-the future of green chemistry? *Green Chemistry* 13(11):3007-3047.

Wilcox G, Meuris P. 1976. Stabilization and size of AraC protein. *Molecular & General Genetics* 145(1):97-100.

Woolston BM, Edgar S, Stephanopoulos G. 2013. Metabolic Engineering: Past and Future. *Annual Review of Chemical and Biomolecular Engineering* 4(Journal Article):259-288.

Wu H, San KY. 2014. Efficient Odd Straight Medium Chain Free Fatty Acid Production by Metabolically Engineered Escherichia Coli. *Biotechnology and Bioengineering* 111(11):2209-2219.

Xie XK, Wong WW, Tang Y. 2007. Improving simvastatin bioconversion in Escherichia coli by deletion of bioH. *Metabolic Engineering* 9(4):379-386.

Yanischperron C, Vieira J, Messing J. 1985. Improved M13 phage cloning vectors and host strains - nucleotide-sequences of the M13mp18 and pUC19 vectors. *Gene* 33(1):103-119.

Yoon SH, Li C, Kim JE, Lee SH, Yoon JY, Choi MS, Seo WT, Yang JK, Kim JY, Kim SW. 2005. Production of vanillin by metabolically engineered Escherichia coli. *Biotechnology Letters* 27(22):1829-1832.

Yoshida M, Tsuru S, Hirata N, Seno S, Matsuda H, Ying B-W, Yomo T. 2014. Directed evolution of cell size in Escherichia coli. *BMC Evolutionary Biology* 14:257.

Yuzawa S, Kim W, Katz L, Keasling JD. 2012. Heterologous production of polyketides by modular type I polyketide synthases in Escherichia coli. *Current Opinion in Biotechnology* 23(5):727-735.

Zaccolo M, Gherardi E. 1999. The effect of high-frequency random mutagenesis on in vitro protein evolution: A study on TEM-1 beta-lactamase. *Journal of Molecular Biology* 285(2):775-783.

Zhang FZ, Keasling J. 2011. Biosensors and their applications in microbial metabolic engineering. *Trends in Microbiology* 19(7):323-329.

Zhang FZ, Rodriguez S, Keasling JD. 2011. Metabolic engineering of microbial pathways for advanced biofuels production. *Current Opinion in Biotechnology* 22(6):775-783.

Zhang W, Li Y, Tang Y. 2008. Engineered biosynthesis of bacterial aromatic polyketides in Escherichia coli. *Proc Natl Acad Sci U S A* 105(52):20683-8.

Zhang YM, Buchholz F, Muyrers JPP, Stewart AF. 1998. A new logic for DNA engineering using recombination in Escherichia coli. *Nature Genetics* 20(2):123-128.

Zhao HM, Giver L, Shao ZX, Affholter JA, Arnold FH. 1998. Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nature Biotechnology* 16(3):258-261.

Zhou H, Xie X, Tang Y. 2008. Engineering natural products using combinatorial biosynthesis and biocatalysis. *Curr Opin Biotechnol* 19(6):590-6.

Zhou P, Wang YM, Gao R, Tong J, Yang ZY. 2015. Transferring NiFe hydrogenase gene from Rhodopeseudomonas palustris into E. coli BL21(DE3) for improving hydrogen production. *International Journal of Hydrogen Energy* 40(12):4329-4336.

Appendix A Chapter I Supplementary Information

Supp. Table A-1 Table of biopharmaceuticals produced form *E. coli* approved by the Food and Drug Administration as of 2014.

Recombinant hormones			
Insulin			
Product	Company (location)	Therapeutic indication	Date approved
Afrezza (rh insulin, produced in *E. coli*)	MannKind (Danbury, CT, USA)	Diabetes mellitus	2014 (US)
Insulin Human Winthrop (rhInsulin produced in *E. coli*)	Sanofi (Frankfurt, Germany)	Diabetes mellitus	2007 (EU)
Apidra (insulin glulisine), rapid acting insulin analog, produced in *E. coli*	Sanofi (Frankfurt, Germany)	Diabetes mellitus	2004 (EU and US)
Lantus (insulin glargine), long-acting rh insulin analog, produced in *E. coli*	Sanofi (Frankfurt, Germany)	Diabetes mellitus	2000 (EU and US)
Liprolog (Insulin lispro), insulin analog, produced in *E. coli*	Eli Lilly (Houten, the Netherlands)	Diabetes mellitus	2001 (EU)
Insuman (rh insulin), produced in *E. coli*	Sanofi (Frankfurt, Germany)	Diabetes mellitus	1997 (EU)
Humalog (insulin lispro, rh insulin analog), produced in *E. coli*	Eli Lilly (Houten, the Netherlands)	Diabetes mellitus	1996 (EU and US)
Humulin (rh insulin), produced in *E. coli*	Eli Lilly	Diabetes mellitus	1982 (US)
Human Growth Hormone			
Accretropin (somatropin) rhGH produced in *E. coli*	Emergent Biosolutions (Rockville, MD, USA)/Cangene (Winnipeg, MB, Canada)	Growth failure or short stature associated with Turner syndrome in pediatric patients	2008 (US)
Omnitrope (somatropin) biosimilar (in EU) r hGH produced in *E. coli*	Sandoz (Kundl, Austria)/ Novartis (Princeton, NJ, USA)	Certain forms of growth disturbance in children and adults	2006 (EU and US)
Somavert (pegvisomant) PEGylated r hGH analog (antagonist) produced in *E. coli*	Pfizer (Sandwich, UK)/Nektar Therapeutics (San Francisco)	Acromegaly	2003 (US), 2002 (EU)
Genotropin (somatropin), r hGH produced in *E. coli*	Pfizer	hGH deficiency in children	1995 (US)
Norditropin (somatropin), r hGH, produced in *E. coli*	Novo Nordisk	Growth failure in children due to inadequate growth hormone secretion	1995 (US)
Tev-tropin/Bio-tropin (somatropin) (r hGH) produced in *E. coli*	Teva Pharmaceuticals USA (North Wales, PA, USA)	hGH deficiency in children	1995 (US)
Nutropin (somatropin), r hGH produced in *E. coli*	Roche/Genentech	hGH deficiency in children	1994 (US)
Humatrope (somatropin) r hGH produced in *E. coli*	Eli Lilly	hGH deficiency in children	1987 (US)
Other Hormones			
Myalept (metreleptin), rh leptin analog, produced in *E. coli*	AstraZeneca (London)/Amylin	Some forms of lipodystrophy	2014 (US)

Gattex (in US)/Revestive (in EU); (teduglutide), rh GLP-2 analog, produced in *E. coli*	NPS Pharma (Dublin)	Short bowel syndrome	2012 (US and EU)
Preotact, rh parathyroid hormone, produced in *E. coli*	NPS Pharma (Dublin)	Osteoporosis	2006 (EU)
Fortical (r salmon calcitonin), produced in *E. coli*	Upsher-Smith Laboratories (Minneapolis, MN, USA)/Unigene (Fairfield, NJ, USA)	Postmenopausal osteoporosis	2005 (US)
Forsteo(EU)/Forteo (US) (teriparatide), r shortened human parathyroid hormone produced in *E. coli*	Eli Lilly (Houten, the Netherlands)	Established osteoporosis in some postmenopausal women	2003 (EU), 2002 (US)
Natrecor (nesiritide), rh natriuretic peptide produced in *E. coli*	Johnson & Johnson/Scios (Titusville, NJ, USA)	Acutely decompensated congestive heart failure	2001 (US)
Glucagon (glucagon, recombinant), rhGlucagon, produced in *E. coli*	Eli Lilly	Hypoglycemia	1998 (US)
Recombinant growth factors			
Colony-stimulating factors			
Grastofil (biosimilar filgrastim), rh G-CSF produced in *E. coli*	Apotex (Leiden, the Netherlands)	Neutropenia	2013 (EU)
Lonquex (lipegfilgrastim), PEGylated rh G-CSF produced in *E. coli*	Teva Pharmaceuticals (Utrecht, the Netherlands)	Neutropenia	2013 (EU)
Granix (tbo-filgrastim) (rh G-CSF produced in *E. coli*) (Note: this is identical to the product 'Tevagrastim', approved as a biosimilar in EU in 2008; see Tevagrastim entry below)	Teva (Frazer, PA, USA)/Cephalon (Malvern, PA, USA)	Neutropenia	2012 (US)
Nivestim (biosimilar filgrastim, rhG-CSFproduced in *E. coli*)	Hospira (Lemington Spa, UK)	Neutropenia	2010 (EU)
Filgrastim hexal biosimilar filgrastim, rh G-CSF produced in *E. coli*)	Hexal (Holzkirchen, Germany)	Neutropenia	2009 (EU)
Zarzio (biosimilar filgrastim, rh G-CSF produced in *E. coli*)	Sandoz (Kundl, Austria)	Neutropenia	2009 (EU)
Biograstim (biosimilar filgrastim, rh G-CSF produced in *E. coli*)	ABZ pharma (Ulm, Germany)	Neutropenia	2008 (EU)
Ratiograstim (biosimilar filgrastim; rh G-CSF produced in *E. coli*)	Ratiopharm (Ulm, Germany)	Neutropenia	2008 (EU)
Tevagrastim (biosimilar filgrastim, rh G-CSF produced in *E. coli*)	Teva (Radebeul, Germany)	Neutropenia	2008 (EU)
Leukine (sargramostim), rh GM-CSF, differs from the native human protein by one amino acid, R23?L; produced in *E. coli*	Sanofi/Berlex Laboratories	Autologous bone marrow transplantation	1991 (US) Withdrawn 2008 and reformulated without EDTA since 2008
Neupogen (filgrastim), rh G-CSF differs from human protein by containing an additional N-terminal methionine; produced in *E. coli*	Amgen	Chemotherapy-induced neutropenia	1991 (US)

Other Growth Factors			
Increlex (mecaserim), rh IGF-1 produced in *E.coli*	Ispen Pharma (Boulogne-Billancourt, France) (formerly Tercica, Brisbane, CA, USA)	Growth failure in children with IGF-1 deficiency or GH gene deletion (long-term treatment)	2007 (EU), 2005 (US)
Kepivance (palifermin), a rh KGF produced in *E. coli*	Swedish Orphan Biovitrum (Stockholm, Sweden) (acquired from Amgen since last listed)	Severe oral mucositis in selected patients with hematologic cancers	2005 (EU), 2004 (US)
Recombinant interferons, interleukins and tumor			
Interferon-α			
PEGintron/ribetol combo pack (peginterferon-α), PEGylated rh IFNα-2b produced in *E. coli* and ribavirin	Schering Plough (Kenilworth, NJ, USA)	Chronic hepatitis C	2008 (US)
Pegasys (PEGinterferon α-2a), produced in *E. coli*	Roche/Genentech (Welwyn Garden City, UK)	Hepatitis C	2002 (EU and US)
PegIntron (PEG rIFN-α-2b), produced in *E. coli*	Merck Sharp & Dohem (MSD, Hoddesdon, UK)	Chronic hepatitis C	2001 (US), 2000 (EU)
ViraferonPeg (PEG rIFN-α-2b), produced in *E. coli*	MSD (Hoddesdon, UK)	Chronic hepatitis C	2000 (EU)
Intron A (also known as Alfatronol) (rIFN-α-2b), produced in *E. coli*	MSD (Hoddesdon, UK)	Cancer, genital warts, hepatitis	2000 (EU), 1986 (US)
Rebetron (combination of ribavirin and rh IFN-α2b) produced in *E. coli*	Schering Plough	Chronic hepatitis C	1999 (US)
Infergen (interferon alficon-1), r IFN-α, synthetic type I IFN produced in *E. coli*	InterMune/Amgen	Chronic hepatitis C	1997 (US), 1999 (EU), Withdrawn (EU) 2006
Interferons β & γ			
Extavia (interferon beta-1b), rh IFN-β1b produced in *E. coli*	Novartis	Multiple sclerosis	2009 (US), 2008 (EU)
Betaferon (interferon-β-1b), r IFN-β1b, differs from human protein by C17?S, produced in *E. coli*	Bayer Pharma (Berlin)	Multiple sclerosis	1995 (EU)
Betaseron (rIFN-β1b), differs from human protein by C17?S, produced in *E. coli*	Bayer/Berlex Labs (Richmond, CA, USA)/Chiron (Emeryville, CA, USA)	Relapsing/remitting multiple sclerosis	1993 (US)
Actimmune (rh IFN-γ1b, produced in *E. coli*)	Vidara Therapeutics (Dublin)	Chronic granulomatous disease	1990 (US)
Others			
Kineret (anakinra), rh IL-1 receptor antagonist produced in *E. coli*	Swedish Orphan Biovitrum/Amgen	Rheumatoid arthritis	2001 (US)
Beromun (tasonermin), rh TNF-α, produced in *E. coli*	Boehringer Ingelheim (Ingelheim, Germany)	Adjunct to surgery for subsequent tumor removal, to prevent or delay amputation	1999 (EU)
Neumega (oprelvekin), rh IL-11, lacks N-terminal proline of native molecule produced in *E. coli*	Pfizer/Genetics Institute	Prevention of chemotherapy induced thrombocytopenia	1997 (US)

Proleukin (aldesleukin), rh IL-2, differs from human molecule in absence of an N-terminal alanine and contains C125?S substitution, produced in *E. coli*	Prometheus Therapeutics and Diagnostics (San Diego)/Chiron	Renal cell carcinoma	1992 (US)
Recombinant vaccines			
Bexsero (meningococcal group B vaccine, rDNA component, absorbed). Multicomponent subunit vaccine, produced in *E. coli.*	Novartis (Siena, Italy)	Immunization against invasive meningococcal disease	2013 (EU)
Monoclonal antibody (mAb)-based products			
Cimzia (certolizumab pegol), anti-TNFα humanized and PEGylated antibody Fab´ fragment produced in E. coli	UCB Pharma (Brussels, Belgium)	Crohn's disease, rheumatoid arthritis	2009 (EU), 2008 (US)
Lucentis (ranibizumab), humanized IgG fragment that binds and inactivates VEGF-A, produced in *E. coli*	Roche/Genentech	Neovascular (wet) age-related macular degeneration	2007 (EU), 2006 (US)
Other recombinant products			
Recombinant enzymes			
Krystexxa (pegloticase), r urate oxidase, PEGylated post synthesis, produced in *E. coli*	Savient (Dublin)/Crealta Pharmaceuticals (Lake Forest, IL, USA)	Gout	2013 (EU), 2010 (US)
Voraxaze (glucarpidase) r carboxypeptidase, produced in *E. coli*	BTG International	Treatment of toxic plasma methotrexate concentrations in patients with delayed methotrexate clearance due to impaired renal function	2012 (US)
Fusion Proteins			
Nplate (romiplostim), a dimeric fusion protein with each monomer consisting of two thrombopoietin receptor binding domains and the Fc region of hIgG-1, produced in *E. coli*	Amgen (Breda, the Netherlands)	Thrombocytopenia	2009 (EU), 2008 (US)
Ontak (denileukin diftitox), r IL-2–diphtheria toxin fusion protein that targets cells displaying a surface IL-2 receptor, produced in *E. coli*	Eisai (Tokyo)/Ligand Pharmaceuticals (San Diego)	Cutaneous T-cell lymphoma	1999 (US)

Appendix B Chapter II Supplementary Information

Supp. Table B-1 List of p-values for pyramid and endpoint (all data from all pyramids for each endpoint combined) comparisons. All p-Values were calculated using the Mann-Whitney U-Test (rank sum based test) because all sets of data were not normally distributed. Values in red indicate p-Value<0.05. Values labeled as zero are less than 0.001.

Pyramid p-Values (Leakiness)

	Pyr1	Pyr2	Pyr3	Pyr5
Pyr1		0.00	0.39	0.00
Pyr2			0.00	0.34
Pyr3				0.00
Pyr5				

	Pyr4	Pyr6
Pyr4		0.00
Pyr6		

	Pyr25	Pyr26
Pyr25		0.00
Pyr26		

Pyramid p-Values (Fold-response)

	Pyr1	Pyr2	Pyr3	Pyr5
Pyr1		0.00	0.01	0.00
Pyr2			0.00	0.00
Pyr3				0.00
Pyr5				

	Pyr4	Pyr6
Pyr4		0.00
Pyr6		

	Pyr25	Pyr26
Pyr25		0.00
Pyr26		

Combined Endpoint p-Values (Fold-response)

	EP1	EP2	EP3	EP4	EP5	EP6	EP7	EP8
EP1		0.83	0.00	0.04	0.00	0.39	0.01	0.34
EP2			0.00	0.04	0.00	0.31	0.01	0.20
EP3				0.00	0.16	0.00	0.03	0.00
EP4					0.02	0.27	0.36	0.29
EP5						0.00	0.39	0.01
EP6							0.07	0.92
EP7								0.17
EP8								

Combined Endpoint p-Values (Leakiness)

	EP1	EP2	EP3	EP4	EP5	EP6	EP7	EP8
EP1		0.30	0.15	0.03	0.13	0.00	0.01	0.09
EP2			0.00	0.00	0.00	0.00	0.00	0.00
EP3				0.21	0.39	0.00	0.13	0.80
EP4					0.79	0.05	0.63	0.16
EP5						0.02	0.42	0.48

Supp. Table B-2 Table outlining the sequences of the top 12 clones for D-ara, mevalonate, and p-coumaric acid

nonploar / polar — basic / acidic — Not mutated position

	#	Clone	AA 8	24	38	80	82	93	Truncated (41 AA) 36 (107_T)	Mutation 83 (248_G)	Truncated (105 AA) 86 (258_G)
		WT	P	T	R	H	Y	H			
		MutA1	R		D	L	Q				
		MutB1	G	W	P	A					
Jlib1	1 J108_3D	**MutDA1***		P	I	P	V	T			
D-ara	2 J108_8D	**MutDA2***		N	P	N	V	P			
Neg	3 J108_6D	MutDA3		STOP	M	T	H	W			
Pyr 1	4 J107_7D	MutDA4		L	V	K	G	V			
Slib4	13 J111_9D	**MutDA5***	S	P		D	P	S			
D-ara	14 J110_11D	**MutDA6***	E	P		R	P	A			
Neg	15 J110_18D	MutDA7	E	P		R	P	A			
Pyr 2	16 J110_2D	MutDA8	E	P		R	P	A			
Jlib1	25 J137_18D	MutDA9		R	W	Y	A	S	deleted		
D-ara	26 J138_18D	MutDA10		R	W	Y	A	S	deleted		
Pos	27 J138_17D	MutDA11		R	W	Y	A	S	deleted		
Pyr 3	28 J158_8D	**MutDA12***		T	T	P	L	D			
Slib4	49 J199_20D	**MutDA13***	G	D		S	V	L			
D-ara	50 J206_5D	MutDA14	L	P		R	A	N			
Pos	51 J187_24D	MutDA15	V	D		R	V	W			
Pyr 5	52 J199_10D	MutDA16	V	L		M	A	R			
Jlib1	37 J153_9D	**MutCA1***		E	R	G	F	G			
p-CA	38 J153_21D	**MutCA2***		M	E	T	L	R			
Pos	39 J153_8D	MutCA3		E	R	G	F	G			
Pyr 4	40 J153_12D	MutCA4		E	R	L	T	C			
Pyr 4	J152-4	MutCA9		W	A	K	P	S			
Slib4	61 J193_11D	**MutCA5***	N	K		R	Q	L			
p-CA	62 J193_21D	MutCA6	D	S		K	L	I			
Pos	63 J193_4D	MutCA7	N	Y		A	H	A			
Pyr 6	64 J204_1D	MutCA8	D	S		K	L	I			
Jlib1	73 J268_6D	**MutMev1***		I	H	L	L	D			
Mev	74 J281_3D	**MutMev2***		Q	G	N	W	L			
Pos	75 J270_24D	MutMev3		P	K	S	S	E		G→A (Gly→Cis)	
Pyr 25	76 J268_1D	MutMev4		P	R	P	Y	G			
Slib4	85 J258_13D	MutMev5	E	P		R	P	A			
Mev	86 J290_14D	**MutMev6***	Q	L		A	I	A			Insert (C)
Pos	87 J287_24D	MutMev7	S	P		D	P	S			
Pyr 26	88 J287_12D	MutMev8	S	P		D	P	S			

Supp. Table B-3 Table outlining the amino acid (AA) and codon sequences of top variants isolated for response to indicated compounds.

nonploar / polar / basic / acidic / Not mutated position

Endpoint	Clone	Compound	Clone	24	38	80	82	93	AA 24	AA 38	AA 80	AA 82	AA 93	Truncated (69 AA) nt 82	Truncated (41 AA) nt 107	Truncated (105 AA) nt 256
			WT	ACG	CGA	CAT	TAC	CAC	T	R	H	Y	H			
J235	**14D**	**t-cinnamic acid**	**MutCin1**	TCG	AGC	CCG	GGC	CTC	S	S	P	G	L		Deletion→ C	
J235	**18D**	**t-cinnamic acid**	**MutCin2**	CCC	GCC	TAC	ACC	TGG	P	A	Y	T	W			
J321	**21D**	**Nicotinic Acid**	**MutNic1**	TTG	GGG	TCC	TCG	CGG	L	G	S	S	R			
J280	**23D**	**Propionic Acid**	**MutProp1**	AGG	CCC	CCG	ATG	CTG	R	P	P	M	L			
J280	**20D**	**Propionic Acid**	**MutProp2**	ACC	CGG	GTC	CTC	AAG	T	R	V	L	K			
J307	**15D**	**Butyric Acid**	**MutBut1**	CGC	ATG	GGC	ACG	TTC	R	M	G	T	F			
J307	**4D**	**Butyric Acid**	**MutBut2**	ACC	TGC	AAC	GCC	TGC	T	C	N	A	C			
J309	**12D**	**Vanillin**	**MutVan1**	TTC	CCG	AGC	AGG	GGG	F	P	S	R	G			Insert→C
J323	**21D**	**Vanillin**	**MutVan2**	GGG	TGC	TTG	GTC	CCG	G	C	L	V	P	Insert→ G		
J309	**17D**	**Vanillin**	**MutVan3**	GAC	TAG	ATG	AAC	AGG	D	STOP	M	N	R			
J279	**9D**	**Theophylline**	**MutTheo1**	ACC	CGG	GTC	CTC	AAG	T	R	V	L	K			
J263	**13D**	**Theophylline**	**MutTheo2**	CTC	CGC	TGC	GCC	GAG	L	R	C	A	E			
J340	**11D**	**Ferulic Acid**	**MutFer1**	TCC	CTG	TTC	CAG	CGC	S	L	F	Q	R			
J342	**23D**	**Levulinic Acid**	**MutLev1**	CTC	ACC	CAG	GCC	TTG	L	T	Q	A	L			
J357	**1D**	**Succinic Acid**	**MutSuc1**	CAG	GGC	GCC	CCC	GCG	Q	G	A	P	A			
J357	**7D**	**Succinic Acid**	**MutSuc2**	AGC	GCC	ACG	AAC	TGG	S	A	T	N	W			
J348	**6D**	**Phloroglucinol**	**MutPhloro1**	GTC	TGG	ACC	GAC	AGC	V	W	T	D	S			
J348	**24D**	**Phloroglucinol**	**MutPhloro2**	AGC	TGC	ACC	TCC	ATC	S	C	T	S	I			

Supp. Table B-4 Table of primers used in Chapter III experiments

Primer name	*Primer sequence (5'-3')*
423mcs4-for	GCTAGGCCATGGGAATTCGCTAGCGCGGCCGCGAGCTGTTGACAATTAATCA
423mcs-rev	GTGATACCATTCGCGAGCCT
pFG29-gib-for	CGCCATTCAGGCTGCAACGACGGCCAGTGAGCG
pFG29-gib-rev	CCTTCCCAACAGTTGCAAAAAAAAAAGCCCGCACTGTCAGGTGCGGGCTTTTTTCTGTGTTT GCGCAATTAACCCTCACTAAAGG
423lib-for-NcoI	GGCGCTATCATGCCATACCG
AraC-T24-rev	TAACCGTTGGCCTCAATCGGSNNTAAACCCGC
AraC-38-for-2	CCGATTGAGGCCAACGGTTATCTCGATTTTTTTATCGACNNSCCGCTGGGA
AraC-H80-Y82-rev	CGAGCCTCCGGATGACGACCSNNGTGSNNAATCTCTCC
araC-H93-for	GGTCGTCATCCGGAGGCTCGCGAATGGTATNNSCAGTGGGTT
AraC-rev-4	ATTGCTGTCTGCCAGGTGATC
AraC-for-5	CAGGAGATATCATATGGCTGAAG
AraC-rev-5	ATGTACTGACAAGCTTCGCG
AraC_Forward_Univ	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNN NCACACAGGAGATATCATATGGCTGAAGCGC
AraC_reverse_BC1	CAAGCAGAAGACGGCATACGAGATTTACCGACGAGTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC2	CAAGCAGAAGACGGCATACGAGATATTGGACACGCTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC3	CAAGCAGAAGACGGCATACGAGATTCGCATGGATACGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC4	CAAGCAGAAGACGGCATACGAGATAGCGAACCTGTTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC5	CAAGCAGAAGACGGCATACGAGATAGCTTCGACAGTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC6	CAAGCAGAAGACGGCATACGAGATGTCAGCCGTTAAGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC7	CAAGCAGAAGACGGCATACGAGATTCCAGATAGCGTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC8	CAAGCAGAAGACGGCATACGAGATGAGAGTCCACTTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC9	CAAGCAGAAGACGGCATACGAGATGCTCACAATGTGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC10	CAAGCAGAAGACGGCATACGAGATTTGACGACATCGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC11	CAAGCAGAAGACGGCATACGAGATCTTAGAACGTGCGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC12	CAAGCAGAAGACGGCATACGAGATCGGTTCACATAGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC13	CAAGCAGAAGACGGCATACGAGATCGATAGGCCTTAGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC14	CAAGCAGAAGACGGCATACGAGATGCTATATCCAGGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC15	CAAGCAGAAGACGGCATACGAGATGTCTTCAGCAAGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTAGCCAGTTAAGCCATTCATGCCAGTAGGCG

Appendix C Chapter III Supplementary Information

C.1) Overview of AraC-TAL clones isolation

QLib3, expressed from pPCC423 vector (P_{tac}, pBR322ΔROP) and transformed into HF19 cells harboring pPCC442 (P_{BAD}-*gfpuv*, RSF1030 origin), was initially negatively screened using FACS against cells with greater fluorescence than the uninduced AraC-TAL using 530/40 channel after excitation with a blue laser (488 nm, 80 mW) on a BD FACSJazz flow cytometer. The sorted population was subjected to an additional four rounds of sorting, alternating positive (induced with 5 mM TAL, top 1%) and negative (induced with 100 µM L-arabinose, % relative to AraC-TAL uninduced) sorting, with 2 consecutive positive sorts concluding the sort scheme. Also to explore the sort path fitness, an alternative sort scheme was carried out, incorporating four rounds of sorting, where the first 2 post-negative sorts were both positive, followed by a negative sort and a final round of positive sorting. Clones from the resulting populations (Q99 and Q101, respectively) were spatially isolated on agar plates and 24 clones from each population were randomly selected and tested for a response to 5 mM TAL and 100 µM L-arabinose in liquid cultures in deep well plates. Clones from Q99 (3) and clones from Q101 (1) showing a positive response (>2-fold response over the background fluorescence) to 5 mM TAL presented similar responses to the original AraC-TAL, but upon sequencing, three of the four clones were unique and all were different from the original AraC-TAL (AraC-TAL1) amino acid sequence. The uninduced leakiness of each newly isolated clone was slightly higher than the original AraC-TAL, and the 5 mM TAL induced response was lower than AraC-TAL1, which lead to an overall lower fold-response.

The number of clones in the endpoint populations, Q99 and Q101, not showing a response to TAL (92%) was higher than expected and AraC-TAL was not isolated. We assumed that the diversity of the endpoint populations were greater than anticipated. Therefore in order to enrich

the positive population of responsive clones, we created a selection using strain SQ12 (HF19::P_{BAD}-*bla*). The plasmid DNA was isolated from each endpoint population, transformed into SQ12, and directly plated on LB medium supplemented with ampicillin and 5 mM TAL. The survival rate in the presence of 300 ug/mL ampicillin and 100 µM L-arabinose or 5 mM TAL, calculated by the number of colony formation units on selection plates (N_{sel}) per the number of transformants (N_{trans}), for the wild-type AraC system (31%) and the AraC-TAL system (80%), respectively, adequately represented a functional selection for the AraC-based biosensor. To test the robustness of the sorting scheme with an endpoint selection, new populations were isolated from SLib4 spiked with AraC-TAL, 1 in 10^7. The same sort paths were followed as described earlier, resulting in populations Q106 and Q109. These populations were again subjected to the endpoint screen in the deep-well plates and after enrichment through selection on 300 ug/mL ampicillin plates in *E. coli* strain SQ12 (10.2% and 2.8% survival rate, respectively), lead to 11 out of 44 screened clones showing a positive response. Selection with 100 ug/mL ampicillin was also tested with populations Q106 and Q109, which lead to 3 out 44 responsive clones. Surprisingly, of the 14 newly isolated clones from Q106 and Q109, six of them were unique and three of them were the original AraC-TAL (now labeled as AraC-TAL1). All of the newly isolated AraC-TAL variants were re-cloned into the original pPCC423 vector, as well as the single plasmid biosensor system plasmid pFG29 (P_{tac}-*araC-TAL*,P_{BAD}-*gfpuv*, pBR322ΔROP) and named AraC-TAL2-10.

C.2) General Materials and Methods

Restriction enzymes, Phusion High-Fidelity DNA polymerase (Cat. No. M0530L), Gibson Assembly Master Mix (Cat. No. E2611) and T4 DNA ligase (Cat. No. M0202L) were purchased from New England Biolabs (Ipswich, MA). Oligonucleotides were synthesized by Integrated DNA Technologies (Coralville, IA) and are reported in Supp. Table C-1. Sanger DNA

sequencing was performed by SeqWright (Houston, TX). Molecular biology techniques for DNA manipulation were performed according to standard protocols (Sambrook and Russell, 2001).

Supp. Table C-1 List of primers used in this study. The underlined sequence highlights the terminator sequence incorporated into pFG29.

Primer name	***Primer sequence (5'-3')***	***Plasmid/clone***
pFG29-gib-for	CGCCATTCAGGCTGCAACGACGGCCAGTGAGCG	pFG29
pFG29-gib-rev	CCTTCCCAACAGTTGCAAAAAAAAAAAGCCCGCACTGTCAGGTGCGGGCTTTTTTCTGTGTTTGCGCAATTAACCCTCACTAAAGG	pFG29
pET45-araC-LBD-for	ATGACGACGACAAGAGTCCCATGGCTGAAGCGCAAAATGAT	pPCC1212
pET45-araC-LBD-rev	AGCTCCCAATTGGGATACCCTCACGACTCGTTAATCGCTTCCATG	pPCC1212
pFG29_araC_GS_fwd_1	ATAAGAGACACCGGCATACT	All MutXXXXX mutants, re-cloned araC-TAL mutants
pCC1305_araCTAL_rvs	ATGCGTTGGTCCTCGCGC	All MutXXXXX mutants, re-cloned araC-TAL mutants
00001_1_WT_rvs	ATACCATTCGCGAGCCTC	00001
00001_2_TAL_fwd	GAGGCTCGCGAATGGTAT	00001
10000_1_WT_rvs	TACTCGTTTAACGCCCAT	10000
10000_2_WT_fwd	GCCACCAGATGGGCGTTA	10000
00010_1_WT_rvs	CAAGTGATGAATCTCTCCTGGC	00010
00010_2_WT_fwd	GCCAGGAGAGATTCATCACTTGGGTCGTCATCC	00010
00100_1_WT_rvs	GCCAATCTCTCCTGGCGGGAACAGCA	00100, 11100
00100_2_WT_fwd	TGCTGTTCCCGCCAGGAGAGATTGGCCACTTGG	00100, 11100
00111_1_WT_rvs	ATCTCTCCTGGCGGGAACAG	00111, 01011, 10100, 01001, 01010, 10110, 11001, 11010
00111_2_TAL_fwd	CTGTTCCCGCCAGGAGAGAT	00111
01000_1_WT_rvs	TTAAACCCGCCACCAGATG	01000
01000_2_TAL_fwd	CCATCTGGTGGCGGGTTTAA	01000

01000_3_WT_fwd	GCTGTTCCCGCCAGGAGAGATT	01000
01000_2_TAL_rvs	AATCTCTCCTGGCGGGAACAGC	01000
01011_1_WT_rvs	CCACCAGATGGGCGTTAAAC	01011, 01100, 01101, 10001, 01110, 10001, 10010, 10011, 10100, 10101, 01001, 01010, 10110
01011_2_TAL_fwd	GTTTAACGCCCATCTGGTGG	01011, 01100, 01101, 10001, 01110, 10001, 10010, 10011, 10100, 10101, 01001, 01010, 10110
01011_3_TAL_fwd	CTGTTCCCGCCAGGAGAGATTCATCACTTGGG	01011
01100_3_WT_fwd	CTGTTCCCGCCAGGAGAGATTGGCCACTACGGT	01100
01100_2_TAL_rvs	GCCAATCTCTCCTGGCGGGAACAG	01100
01101_3_TAL_fwd	GCCAGGAGAGATTGGCCACTACGGTCGTCAT	01101
01101_2_TAL_rvs	GTAGTGGCCAATCTCTCCTGGC	01101
10001_3_TAL_fwd	CGGAGGCTCGCGAATGGTAT	10001, 01110
10001_2_WT_rvs	ATACCATTCGCGAGCCTCCG	10001, 01110
01111_1_WT_rvs	TAAACCCGCCACCAGATGGGC	01111
01111_2_TAL_fwd	GCCCATCTGGTGGCGGGTTTA	01111
10001_3_TAL_fwd	CGGAGGCTCGCGAATGGTAT	10001
10001_2_WT_rvs	ATACCATTCGCGAGCCTCCG	10001
10010_3_WT_fwd	CCGCCAGGAGAGATTCATCACTTGGGTCGTCA	10010
10010_2_WT_rvs	TGATGAATCTCTCCTGGCGG	10010
10011_3_TAL_fwd	GCCAGGAGAGATTCATCACTTGGGTCGTCA	10011, 00011
10011_2_WT_rvs	AGTGATGAATCTCTCCTGGC	10011, 00011
10100_3_WT_fwd	CTGTTCCCGCCAGGAGAGATTGGCCACTACGGT	10100
10101_3_TAL_fwd	TCCCGCCAGGAGAGATTGGCCACTACGGTCGTCAT	10101
10101_2_WT_rvs	GCCAATCTCTCCTGGCGGGAACA	10101
10111_1_TAL_	TTAAACCCGCCACCAGATG	10111

rvs		
10111_2_WT_fwd	CCATCTGGTGGCGGGTTTAA	10111
10111_3_TAL_fwd	GCTGTTCCCGCCAGGAGAGATT	10111
10111_2_WT_rvs	AATCTCTCCTGGCGGGAACAGC	10111
11011_1_TAL_rvs	ATGAATCTCTCCTGGCGGGAACAGCA	11011
11011_2_TAL_fwd	TGCTGTTCCCGCCAGGAGAGATTCATCACTTGGGTC	11011
11101_1_TAL_rvs	GTAGTGGCCAATCTCTCCTGGC	11101
11101_2_TAL_fwd	GCCAGGAGAGATTGGCCACTACGGTCGTCATCC	11101
11110_1_TAL_rvs	ATACCATTCGCGAGCCTC	11110
11110_2_WT_fwd	GAGGCTCGCGAATGGTAT	11110
00101_1_WT_rvs	GCCAATCTCTCCTGGCGGGAAC	00101
00101_2_TAL_fwd	CCCGCCAGGAGAGATTGGCCACTACGGTCGTCA	00101
00110_1_WT_rvs	CAAGTGGCCAATCTCTCCTGGCGGGAACA	00110
00110_2_WT_fwd	TGTTCCCGCCAGGAGAGATTGGCCACTTGGGTCGTCA	00110, 10110
01001_3_TAL_fwd	CTGTTCCCGCCAGGAGAGATTCATCACTACGGTCGTCA	01001, 11001
01010_3_WT_fwd	CTGTTCCCGCCAGGAGAGATTCATCACTTGGGTCGTCA	01010
11010_2_WT_fwd	CTGTTCCCGCCAGGAGAGATTCATCACTTGGGTCGTCA	11010
Adding_ScaI_Fragment_1_fwd	CCCCAGCAGGCGAAAATCCTGTTTG	pPCC1321
araCTAL_5_AvrII_rvs	ATTTTGCGCTTCAGCCATCCTAGGTATCTCCTGTG	pPCC1321
AraCTAL_5	ATGGCTGAAGCGCAAAATGATCC	pPCC1321
pCC1305_araCTAL_rvs	ATGCGTTGGTCCTCGCGC	pPCC1321
pCC1321_PciI_PBAD	TTTTGCTGGCCTTTTGCTCAACTTTTCATACTCCCGCC	pPCC1322
pCC1321_AgeI_PBAD	GCTTTTAATAAGCGGGGTTA	pPCC1322

C.3) Integration of a P_{BAD}-*bla* gene into HF19 for ampicillin selection

SQ12 strain was created by integrating a fragment of DNA containing P_{BAD}-*bla* (conferring ampicillin resistance regulated by the AraC cognate promoter P_{BAD}) into the genome of HF19 using CRIM method (Haldimann and Wanner, 2001). The P_{BAD}-*gfpuv* was amplified using primers pPCC1215-gib-for and pPCC1215-gib-rev, and then ligated with into NcoI digested CRIM plasmid pPCC20 (Chin et al., 2009) by Gibson Assembly, resulting in pPCC1215. The *bla* gene was amplified from pET45b_SmaI using primers pPCC1217-gib-for and pPCC1217-fib-rev. The *bla* gene was cloned with pPCC1215 vector digested with NdeI and SpeI using Gibson Assembly, resulting in pPCC1217. Plasmid pPCC1217 was subsequently integrated into the chromosome of HF19 at the HK022 site resulting in SQ11. Apramycin resistant colonies were selected and the integration was verified by PCR. Removal of FRT-flanked apramycin resistance cassette was achieved as described (Causey et al., 2003), resulting in strain SQ12.

C.4) Cloning of plasmids for screening and AraC library

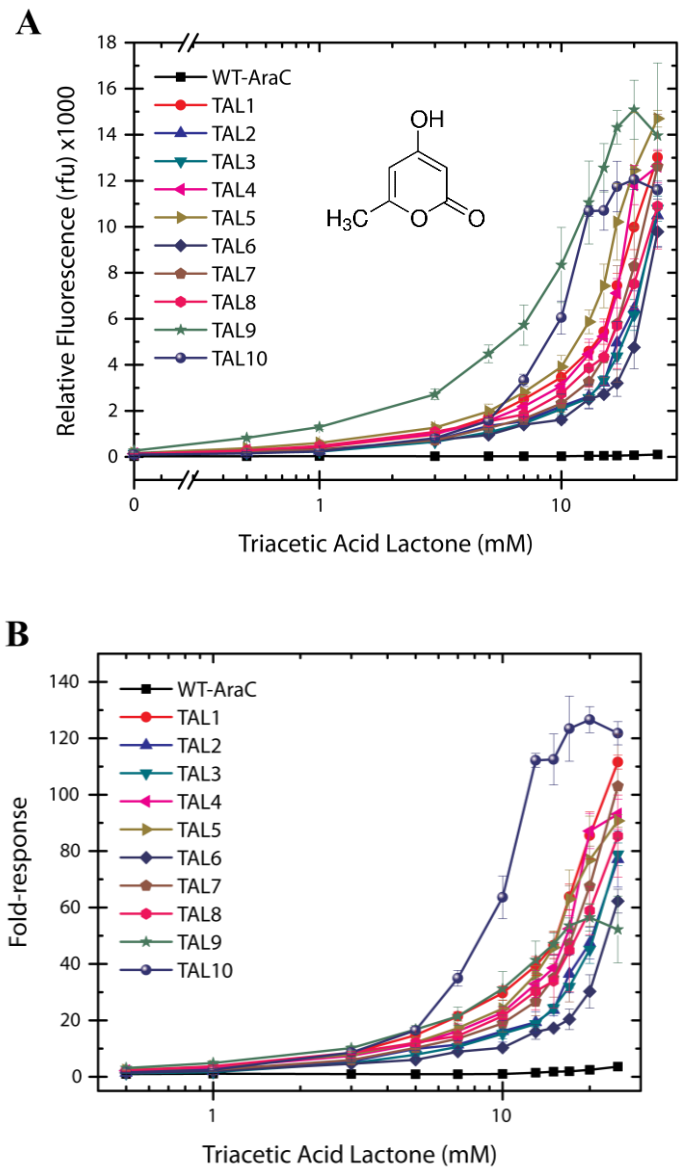
Initial work was based on the dual plasmid reporter system for AraC-controlled GFPuv expression that was described previously (Tang et al., 2008), where AraC and the AraC combinatorial library (SLib4) are expressed from plasmid pPCC423 (maintained by apramycin antibiotic resistance) controlled by IPTG-inducible LacI. GFPuv is subsequently expressed from plasmid pPCC442 (chloramphenicol resistance), where it is controlled by P_{BAD}. Subsequently, plasmid pFG1 for AraC expression was constructed from pPCC423 and pFG29 was constructed using the pFG1 vector (described below). Construction and expression of the mutant library was carried out as previously described (Tang et al., 2008).

pFG1 was cloned to incorporate P_{BAD}-*gfpuv*. This was accomplished using the Gibson method. Primers pFG29-gib-for and pFG29-gib-rev were designed to amplify P_{BAD}-*gfpuv* from pPCC442. Primer pFG29-gib-rev incorporated a terminator sequence (AAAAAAAAAAGCCCGCACTGTCAGGTGCGGGCTTTTTTCTGTGTTT). P_{BAD}-*gfpuv* was amplified using Phusion polymerase. The resulting PCR product was gel purified. The pFG1 vector was cleaved with the FspI restriction enzyme. These fragments were mixed and assembled according to the Gibson Assembly protocol. The resulting plasmid was named pFG29, containing P_{tac}-*araC* and P_{BAD}-*gfpuv*.

All AraC-TAL clones were cloned into plasmid pFG29 from the pPCC423 vector for analysis after isolation from the SLib4 library. This was done by PCR amplification of *araC* variants using primers pFG29-araC-GS and pPCC1305_araCTAL-rvs. The products and pFG29 vector were subjected to sequential digest by AflII and BstapI. The purified products were ligated using T4 DNA ligase and transformed into electroporation competent MC1061 cells. Sequencing of the final clones confirmed the correct sequences.

C.5) Dose-dependent responses of AraC-TAL variants

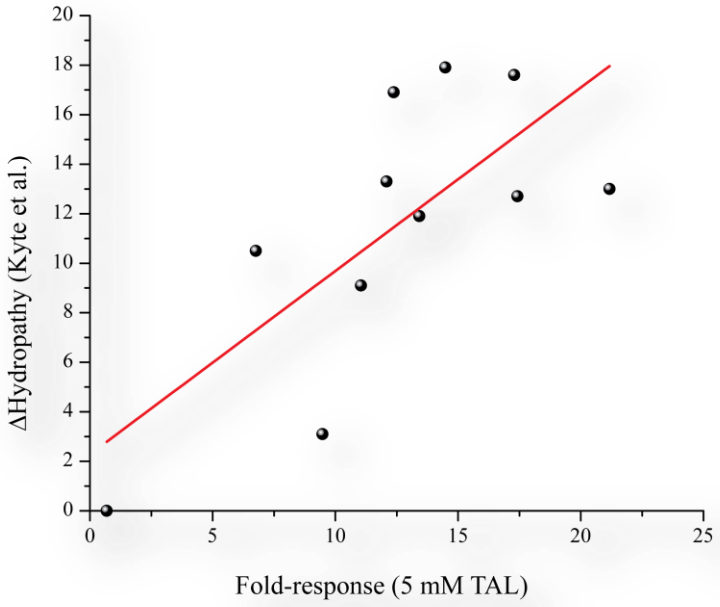
To better characterize the newly isolated AraC-TAL variants, we determined the dose-dependent response of each variant to TAL (1-25 mM) using the protocol outlined in the Materials and Methods section. Each variant shows a dose-dependent response to TAL (Supp. Figure C-1). However due to the toxicity of high concentrations of TAL (>25 mM), the full dynamic range of response could not be determined for most variants.



Supp. Figure C-1 TAL-dependent dose response of AraC-TAL variants. Data is reported as the average and standard deviation of three independent experiments in relative fluorescence units (A) and fold-response (B). The fold-response was calculated by dividing the bulk fluorescence in the presence of TAL by the bulk background fluorescence in the absence of TAL.

C.6) AraC-TAL fold-response depends on residue hydrophobicity and charge

The hydrophobicity of each amino acid substitution was determined from residue sidechain hydrophobicity values provided Kyte and coworkers (Kyte and Doolittle, 1982). Plotting the total change in hydrophobicity for each variant versus the respective response of the variant shows a positive correlation between an increase in amino acid substitution hydrophobicity and response (Supp. Figure C-2). The change in hydropathy (ΔHydropathy) of the ligand binding domain (LBD) was calculated by summing the corresponding hydropathy values for the substituted residues of each variant and subtracting the wt-AraC value (Supp. Table C-2). In addition to the hydropathy, the net charge was also calculated. In all variants, the net charge was positive (at a neutral pH), but the variants with the greatest charge were the least responsive.



Supp. Figure C-2 Scatter plot representing the trend of hydrophobicity from amino acid substitutions and fold-response to 5 mM TAL. The hydrophobicity data was calculated from the hydrophobicity indices calculated by Kyte and coworkers. The red line in the scatter plot represents the simple linear regression model fitted to the respective data. There is a positive correlation showing that the fold-response to 5 mM TAL increases as the hydrophobicity increases.

Supp. Table C-2 Charge and hydrophobicity of amino acid substitutions in the AraC-TAL clones. (A) Net change in charge (Δz) of the LBD shows there is a net positive charge for all AraC-TAL variants. The net change in hydrophobicity was calculated according to (B) Kyte and coworkers. Each AraC-TAL variant showed positive net charge and a positive net hydropathy (more hydrophobic) in the LBD.

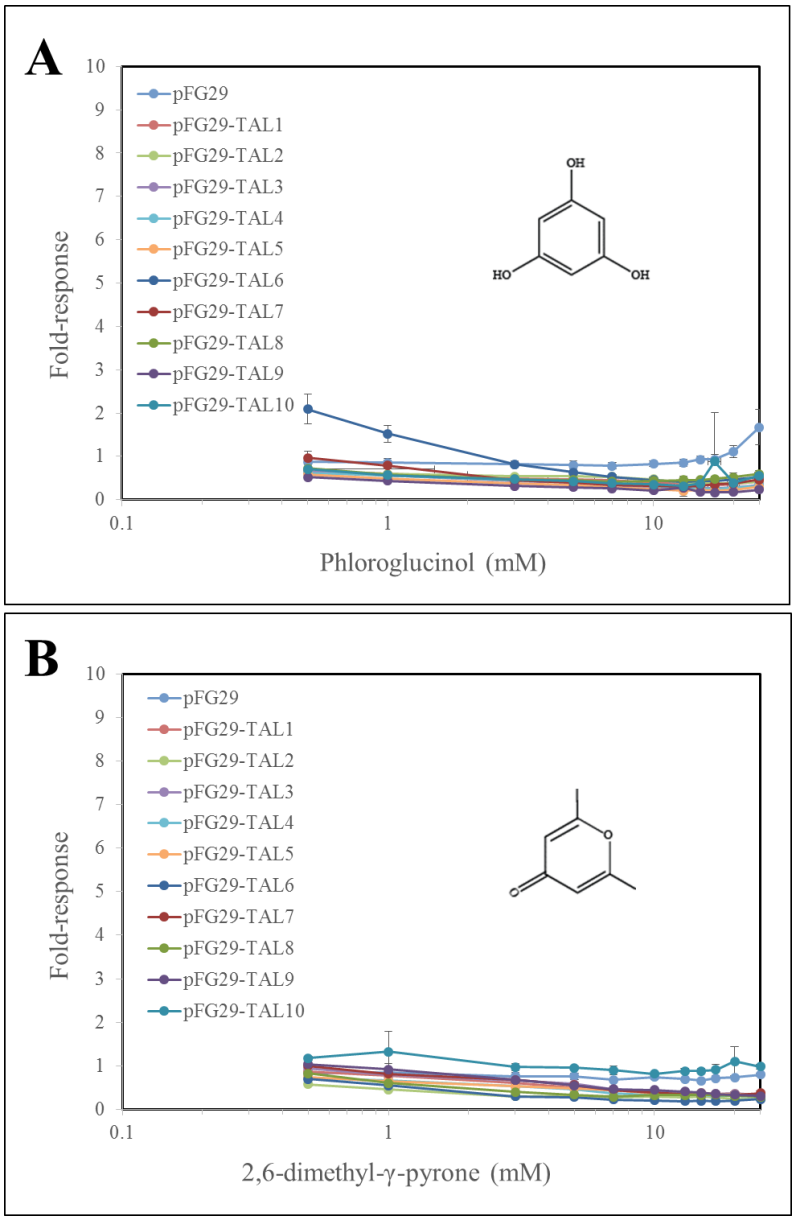
A	Residue					z (pH 7)						
Clone	8	24	80	82	93	8	24	80	82	93	z_{sum}	Δz
WT-AraC	P	T	H	Y	H	0	0	0.09	0	0.09	0.18	0.00
AraC-TAL1	V	I	G	L	R	0	0	0	0	1	1	0.82
AraC-TAL2	G	H	H	K	L	0	0.09	0.09	1	0	1.18	1.00
AraC-TAL3	S	I	G	I	R	0	0	0	0	1	1	0.82
AraC-TAL4	S	L	G	L	R	0	0	0	0	1	1	0.82
AraC-TAL5	I	L	G	I	R	0	0	0	0	1	1	0.82
AraC-TAL6	G	L	H	K	V	0	0	0.09	1	0	1.09	0.91
AraC-TAL7	V	L	G	L	R	0	0	0	0	1	1	0.82
AraC-TAL8	G	L	H	K	F	0	0	0.09	1	0	1.09	0.91
AraC-TAL9	T	I	G	L	R	0	0	0	0	1	1	0.82
AraC-TAL10	G	L	G	I	R	0	0	0	0	1	1	0.82

B	Residue					Hydropathy (Kyte et al. 1982)						
Clone	8	24	80	82	93	8	24	80	82	93	*HI*	*ΔHI*
WT-AraC	P	T	H	Y	H	-1.6	-0.7	-3.2	-1.3	-3.2	-10.0	-
AraC-TAL1	V	I	G	L	R	4.2	4.5	-0.4	3.8	-4.5	7.6	17.6
AraC-TAL2	G	H	H	K	L	-0.4	-3.2	-3.2	-3.9	3.8	-6.9	3.1
AraC-TAL3	S	I	G	I	R	-0.8	4.5	-0.4	4.5	-4.5	3.3	13.3
AraC-TAL4	S	L	G	L	R	-0.8	3.8	-0.4	3.8	-4.5	1.9	11.9
AraC-TAL5	I	L	G	I	R	4.5	3.8	-0.4	4.5	-4.5	7.9	17.9
AraC-TAL6	G	L	H	K	V	-0.4	3.8	-3.2	-3.9	4.2	0.5	10.5
AraC-TAL7	V	L	G	L	R	4.2	3.8	-0.4	3.8	-4.5	6.9	16.9
AraC-TAL8	G	L	H	K	F	-0.4	3.8	-3.2	-3.9	2.8	-0.9	9.1
AraC-TAL9	T	I	G	L	R	-0.7	4.5	-0.4	3.8	-4.5	2.7	12.7
AraC-TAL10	G	L	G	I	R	-0.4	3.8	-0.4	4.5	-4.5	3.0	13.0

C.7) AraC-TAL variants show specificity towards TAL

The specificity of the AraC-TAL variants was tested in the presence of two compounds similar to TAL, phloroglucinol and 2,6-dimethyl-γ-pyrone (Supp. Figure C-3). The dose response was setup as described in the Materials and Methods. Phloroglucinol and 2,6-dimethyl-

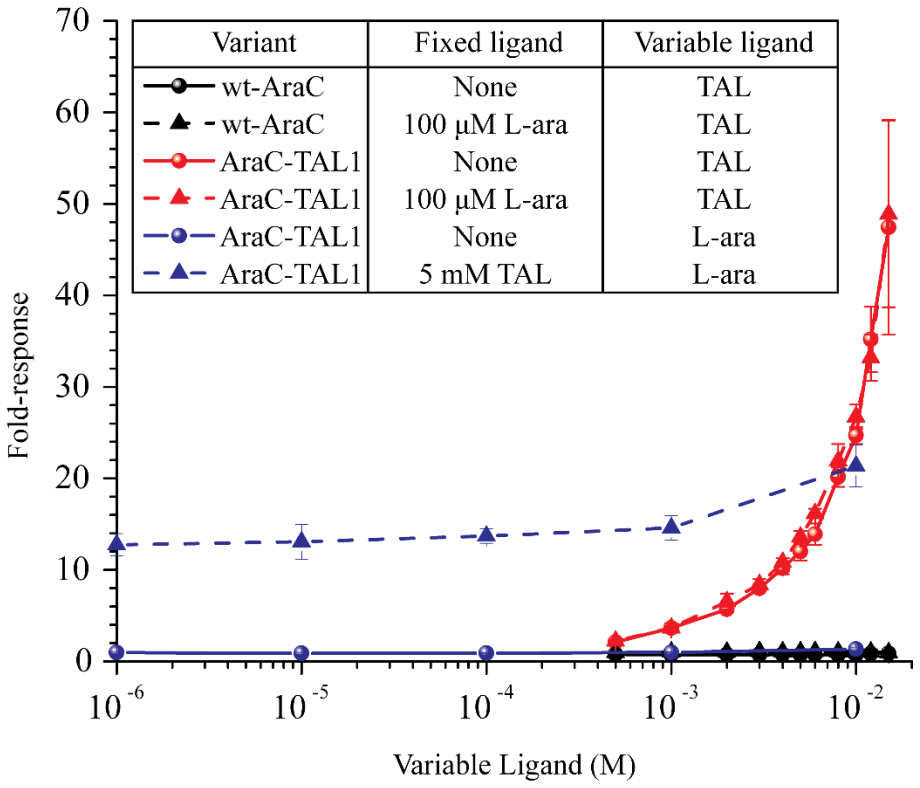
γ-pyrone were prepared fresh and dissolved directly in the media to 50 mM. Error bars were incorporated and represent the standard deviation of four replicate cultures. Dilutions were made from this stock solution. None of the variants show a high response to either compound. AraC-TAL6 shows a slight response to phloroglucinol at low concentrations.



Supp. Figure C-3 Dose dependent responses of AraC-TAL variants to (A) phloroglucinol and (B) 2,6-dimethyl-γ-pyrone.

C.8) L-arabinose is not an inhibitor of TAL response

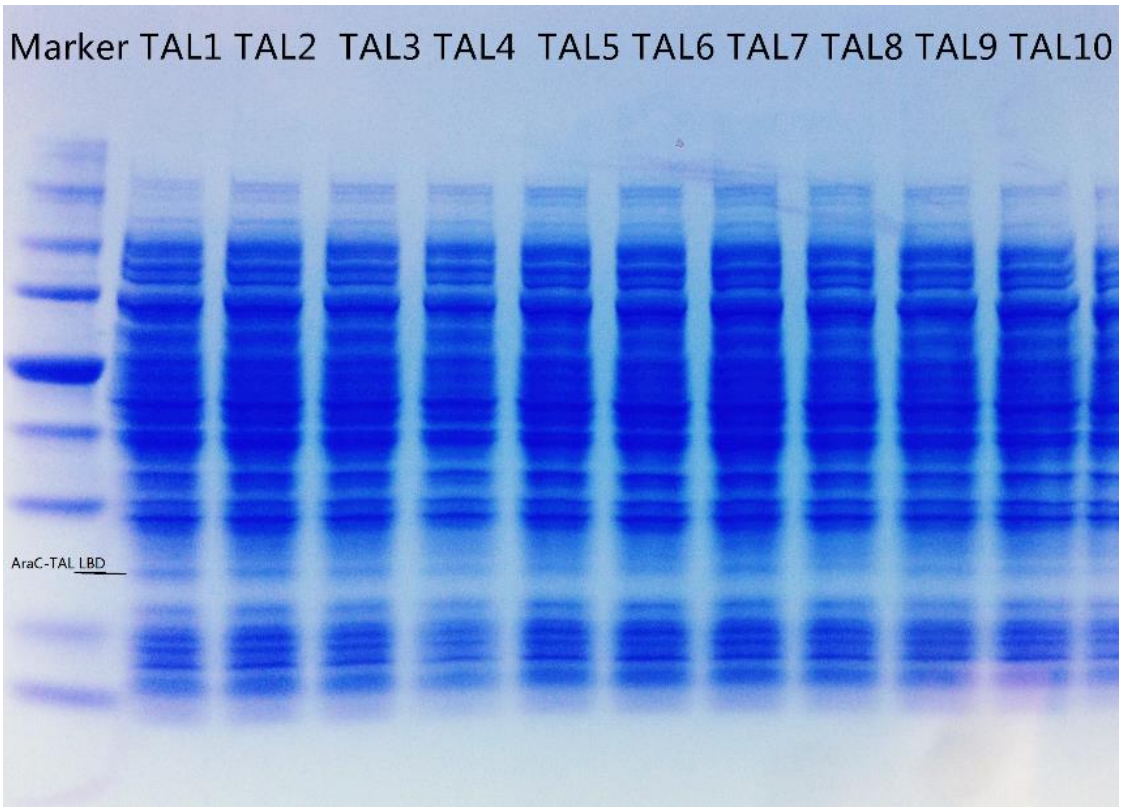
The response of AraC-TAL1 in the presence of L-ara was investigated to determine the extent of L-ara binding. The competition assay was setup following the protocol for deep-well plate dose responses described in the Materials and Methods section. HF19 cells harboring either pFG29 or pFG29-TAL1 were screened for response to TAL (0.5 – 25 mM) in the presence and absence of 1 mM L-ara in the media. Also, the response of AraC-TAL1 to 5 mM TAL in the presence of varying concentrations of L-ara (0.001-10 mM L-ara) was determined. As can be seen in Supp. Figure C-4, the response of AraC-TAL1 was not affected by the presence of L-ara. This supports our conclusions that the polar L-ara molecules have reduced bind affinity in the LBD.



Supp. Figure C-4 Competition assay to determine the effect of the presence of L-ara on AraC-TAL1 response to TAL.

C.9) Protein gel analysis of soluble LBD of AraC-TAL clones

Due to the low solubility of the AraC-TAL1 LBD, we examined the remaining AraC-TAL variants to see if they were more soluble than AraC-TAL1. AraC-TAL variants were cloned into the pET45b vector and expressed as described in the Materials and Methods section of this manuscript. The induced cells were lysed by boiling in lysis buffer for 10 min. The soluble fraction of the lysed cells was loaded onto a SDS-PAGE gel. As can be seen in Supp. Figure C-5, none of the other clones were significantly more soluble that AraC-TAL1. Most of the variants were less soluble as determined by the intensity of the band corresponding to the AraC LBD.

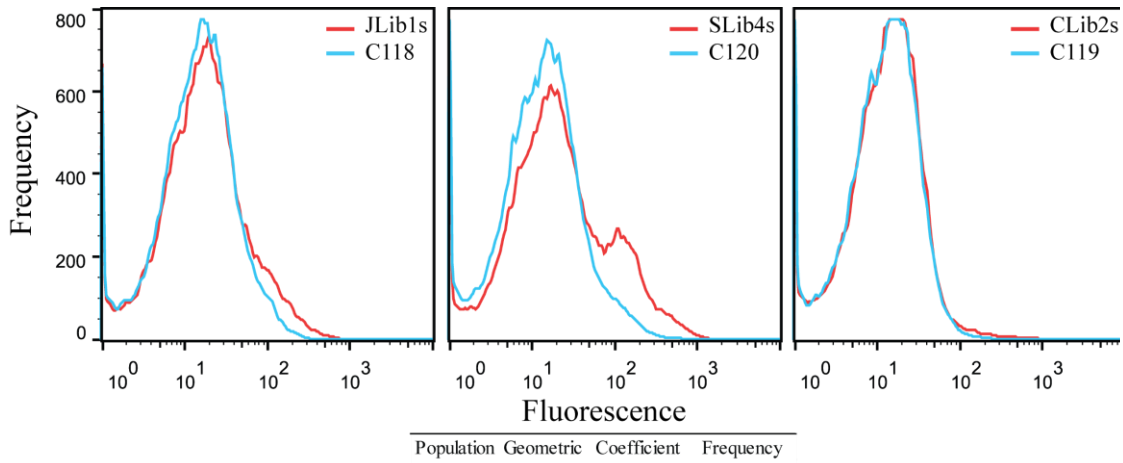


Supp. Figure C-5 SDS-PAGE of the ligand binding domains (LBDs) of all AraC-TAL clones. Cells induced to express the AraC-TAL LBD were lysed by boiling the cells. Cell debris were pelleted briefly by centrifugation. The supernatant of each sample was run on a polyacrylamide gel. There was no significant increase in solubility for the LBDs of AraC-TAL2-10 when compared to AraC-TAL1 (indicated by the black line on the gel image).

Appendix D Chapter IV Supplementary Information

D.1) Negative sorting of naïve libraries reduces leaky clone frequency

The naïve libraries JLib1s, SLib4s, and CLib2s (all expressed from the single plasmid system which is indicated by "s") were subjected to a first round negative sort in the absence of any ligand. We did this to reduce the frequency of leaky clones present in the population so they did not get enriched in the immediate subsequent round of selection. Supp. Figure D-1 shows effect of the negative sort on the library populations. The negative sort had the most effect on SLib4s, lowering the geometric mean of the uninduced population from 19.8 rfu to 9.3 rfu.

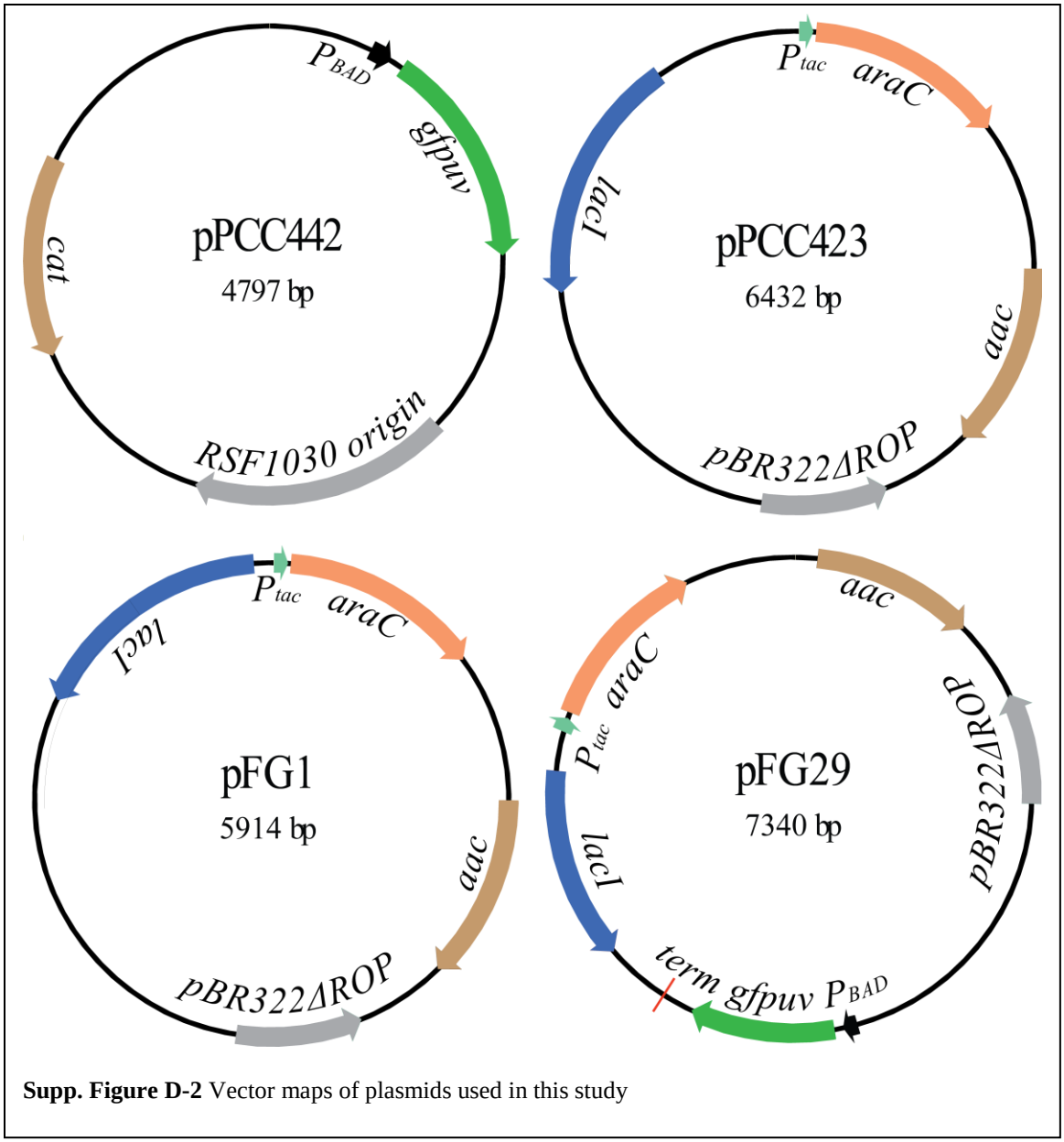


Population Name	Geometric Mean	Coefficient of Variance	Frequency of WT-AraC
WT-AraC	14.4	84.0	99.9
JLib1s	16.3	190.0	94.4
C118	12.5	144.0	98.5
SLib4s	19.8	244.0	88.5
C120	9.3	236.0	98.5
CLib2s	12.1	397.0	98.6
C119	11.0	265.0	99.7
CLib5s	10.2	187.0	98.9

Supp. Figure D-1 Histograms of the fluorescence response of library populations and their resulting geometric means. The red line corresponds to the naïve library and the blue line is the respective population after the first round of negative FACS screening.

Supp. Table D-1 List of primers used in chapter 5

Primer name	*Primer sequence (5'-3')*
423mcs4-for	GCTAGGCCATGGGAATTCGCTAGCGCGGCCGCGAGCTGTTGACAATTAATCA
423mcs-rev	GTGATACCATTCGCGAGCCT
pFG29-gib-for	CGCCATTCAGGCTGCAACGACGGCCAGTGAGCG
pFG29-gib-rev	CCTTCCCAACAGTTGCAAAAAAAAAAGCCCGCACTGTCAGGTGCGGGCTTTTTTCTGTGTTT GCGCAATTAACCCTCACTAAAGG
423lib-for-NcoI	GGCGCTATCATGCCATACCG
AraC-T24-rev	TAACCGTTGGCCTCAATCGGSNNTAAACCCGC
AraC-38-for-2	CCGATTGAGGCCAACGGTTATCTCGATTTTTTTATCGACNNSCCGCTGGGA
AraC-H80-Y82-rev	CGAGCCTCCGGATGACGACCSNNGTGSNNAATCTCTCC
araC-H93-for	GGTCGTCATCCGGAGGCTCGCGAATGGTATNNSCAGTGGGTT
AraC-rev-4	ATTGCTGTCTGCCAGGTGATC
AraC-for-5	CAGGAGATATCATATGGCTGAAG
AraC-rev-5	ATGTACTGACAAGCTTCGCG
AraC_Forward_Univ	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNN NCACACAGGAGATATCATATGGCTGAAGCGC
AraC_reverse_BC1	CAAGCAGAAGACGGCATACGAGATTTACCGACGAGTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC2	CAAGCAGAAGACGGCATACGAGATATTGGACACGCTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC3	CAAGCAGAAGACGGCATACGAGATTCGCATGGATACGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC4	CAAGCAGAAGACGGCATACGAGATAGCGAACCTGTTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC5	CAAGCAGAAGACGGCATACGAGATAGCTTCGACAGTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC6	CAAGCAGAAGACGGCATACGAGATGTCAGCCGTTAAGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC7	CAAGCAGAAGACGGCATACGAGATTCCAGATAGCGTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_reverse_BC8	CAAGCAGAAGACGGCATACGAGATGAGAGTCCACTTGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC9	CAAGCAGAAGACGGCATACGAGATGCTCACAATGTGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC10	CAAGCAGAAGACGGCATACGAGATTTGACGACATCGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC11	CAAGCAGAAGACGGCATACGAGATCTTAGAACGTGCGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC12	CAAGCAGAAGACGGCATACGAGATCGGTTCACATAGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCAGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC13	CAAGCAGAAGACGGCATACGAGATCGATAGGCCTTAGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC14	CAAGCAGAAGACGGCATACGAGATGCTATATCCAGGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTGCCAGTTAAGCCATTCATGCCAGTAGGCG
AraC_Reverse_BC15	CAAGCAGAAGACGGCATACGAGATGTCTTCAGCAAGGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTAGCCAGTTAAGCCATTCATGCCAGTAGGCG



Supp. Figure D-2 Vector maps of plasmids used in this study

Appendix E Preliminary results from high-throughput sequencing of screened populations

E.1) General Materials and Methods

Sorted populations from Chapters II, III, and IV were subjected to high-throughput sequencing in collaboration with the Joint Genome Institute. We sought to explore the level of enrichment from the various rounds of FACS and selections. Naïve and screened populations were amplified using Phusion polymerase with primers were designed to amplify the first 100 codons of the *araC* gene, which includes all targeted codons for saturation mutagenesis. Both the forward and reverse primers contain an Illumina adaptor sequence (29 nt), a barcode sequence (12 nt), a sequencing primer adapter (17-34 nt), and an *araC* homologous region (30-33 nt). The primer sequences are outlined in Supp. Table E-I. The reaction was setup with: 1x Phusion HF buffer, 0.25 mM dNTPs, 0.5 µM of each primer, 3% dimethyl sulfoxide, 0.04 U/µL Phusion polymerase, 1 ng/µL template DNA (purified plasmid from screened population). The reaction was run under the following conditions: 98°C for 30 seconds; 30 cycles of 98°C for 10 seconds, 68°C for 15 seconds, and 72°C for 15 seconds; 72°C for 10 minutes; 4°C for an unlimited time. Each amplified fragment was gel purified. A minimum of 2 µg purified fragment was needed for sample preparation for running on the Illumina MiSeq (Illumina, San Diego, CA) next generation sequencer. Sample preparation was done according the protocol (Preparing Libraries for Sequencing on the MiSeq) provided on the Illumina website.

E.2) Results

The frequencies of isolated AraC variants in sequenced populations are represented in Supp. Figure E-1 as a heat map. In general, variants were enriched over several rounds of screening for the compound they show a response to. The data was extracted from amino acid sequences with respect to the target residues (e.g., “PTRHYH” is the representative sequence for

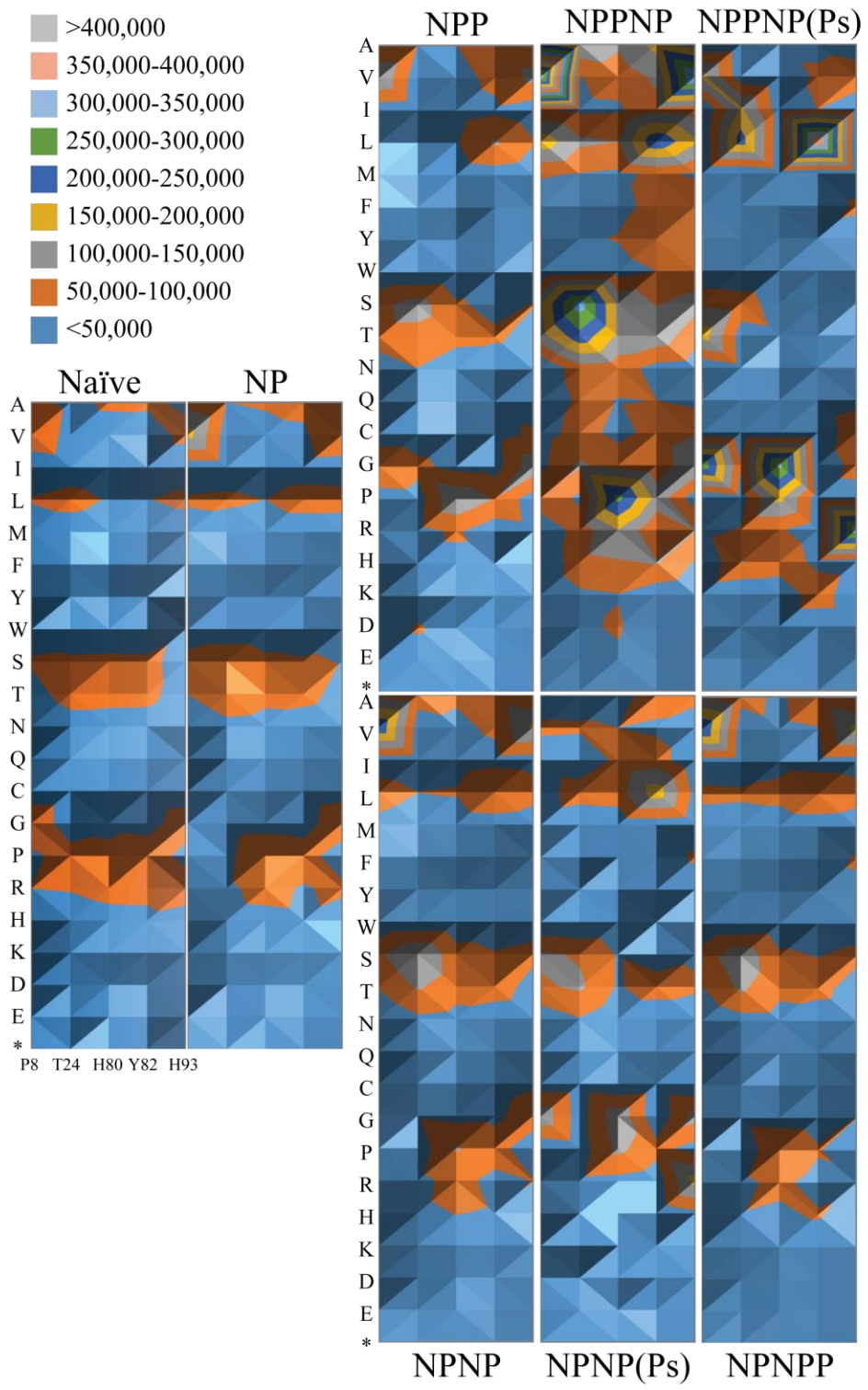
Supp. Table E-1 Table of primers for high-throughput sequencing on an Illumina MiSeq next generation sequencer. Primers with staggered spaces (0, 1-, 2-, 3- nt.) between Illumina adapter and specific primer. The first section (CAAGCAGAAGACGGCATACGAGAT) is the Illumina adaptor. The second section for example TTACCGACGAGT are the barcodes. The third section is the sequencing primer adapter. The last section of the primers are the homologous regions for amplification of the *araC* fragment.

Primer name	*Length (nt)*	*Illumina adaptor*	*Barcode*	*Sequencing primer adaptor*
AraC_Forward_Univ	93	5'- AATGATACGGCGACCACCGAGATCTACAC	TCTTTCCCTACA	CGACGCTCTTCCGATCT
AraC_Reverse_BC1	98	5'- CAAGCAGAAGACGGCATACGAGAT	TTACCGACGAGT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC2	99	5'- CAAGCAGAAGACGGCATACGAGAT	ATTGGACACGCT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC3	100	5'- CAAGCAGAAGACGGCATACGAGAT	TCGCATGGATAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC4	101	5'- CAAGCAGAAGACGGCATACGAGAT	AGCGAACCTGTT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC5	98	5'- CAAGCAGAAGACGGCATACGAGAT	AGCTTCGACAGT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC6	99	5'- CAAGCAGAAGACGGCATACGAGAT	GTCAGCCGTTAA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC7	100	5'- CAAGCAGAAGACGGCATACGAGAT	TCCAGATAGCGT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC8	101	5'- CAAGCAGAAGACGGCATACGAGAT	GAGAGTCCACTT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC9	98	5'- CAAGCAGAAGACGGCATACGAGAT	GCTCACAATGTG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC10	99	5'- CAAGCAGAAGACGGCATACGAGAT	TTGACGACATCG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC11	100	5'- CAAGCAGAAGACGGCATACGAGAT	CTTAGAACGTGC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC12	101	5'- CAAGCAGAAGACGGCATACGAGAT	CGGTTCACATAG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC13	98	5'- CAAGCAGAAGACGGCATACGAGAT	CGATAGGCCTTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC14	99	5'- CAAGCAGAAGACGGCATACGAGAT	GCTATATCCAGG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AraC_Reverse_BC15	100	5'- CAAGCAGAAGACGGCATACGAGAT	GTCTTCAGCAAG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

wt-AraC). Here we aim to determine if there is an amino acid substitution pattern seen in early rounds of screening that may help predict potential responsive variants. If we can find a pattern early in the screening, we may be able to use this information to predict which amino acid substitutions will lead to functioning variants. A closer look at the amino acid frequencies per residue position over several rounds of screening is shown in Supp. Figure E-2 for the TAL sorting that was discussed in Chapter III. The sorted populations with two early positive sorts show a much higher enrichment than the other screening path. The "NPPNP" population shows a high frequency of T24S substitutions, but after a single round of positive selection, the T24S frequency dramatically dropped and was taken over by T24L and T24I. All of the AraC-TAL variants discussed in Chapter III have either the T24L or T24I substitution. Also, the majority of the AraC-TAL variants have a H80G substitutions, but that substitution did not show any significant frequency in the populations until after a round of selection. However, the Y82L substitution is common among the AraC-TAL variants and shows enrichment in every round (except for the "NPNPP" population) and is seen as early as the first round of sorting in the "NP" population. The original AraC-TAL variant has a P8V substitution, which again shows enrichment in almost every round of screening. Finally, the H93R substitution also shows some enrichment throughout the rounds of screening. Though some the amino acid substitutions that are common in the AraC-TAL variants are not enriched early in the screening path, we should be able to predict some of the important amino acid substitutions after only a couple of rounds of screening.

The second aim of this study was to determine the efficiency of our screening strategies. Each round of FACS screening from the pyramid sorting described in Chapter II was positive sorted by collecting the top 1% of most fluorescent cells in the presence of the target compound. The average total different variants after at least three rounds of positive sorting was approximately 2×10^{5}. This was much higher than expected, especially because we were only

Supp. Figure E-1 Heat map of high-throughput sequencing results. Green indicates high frequency and red indicates low frequency in the population.

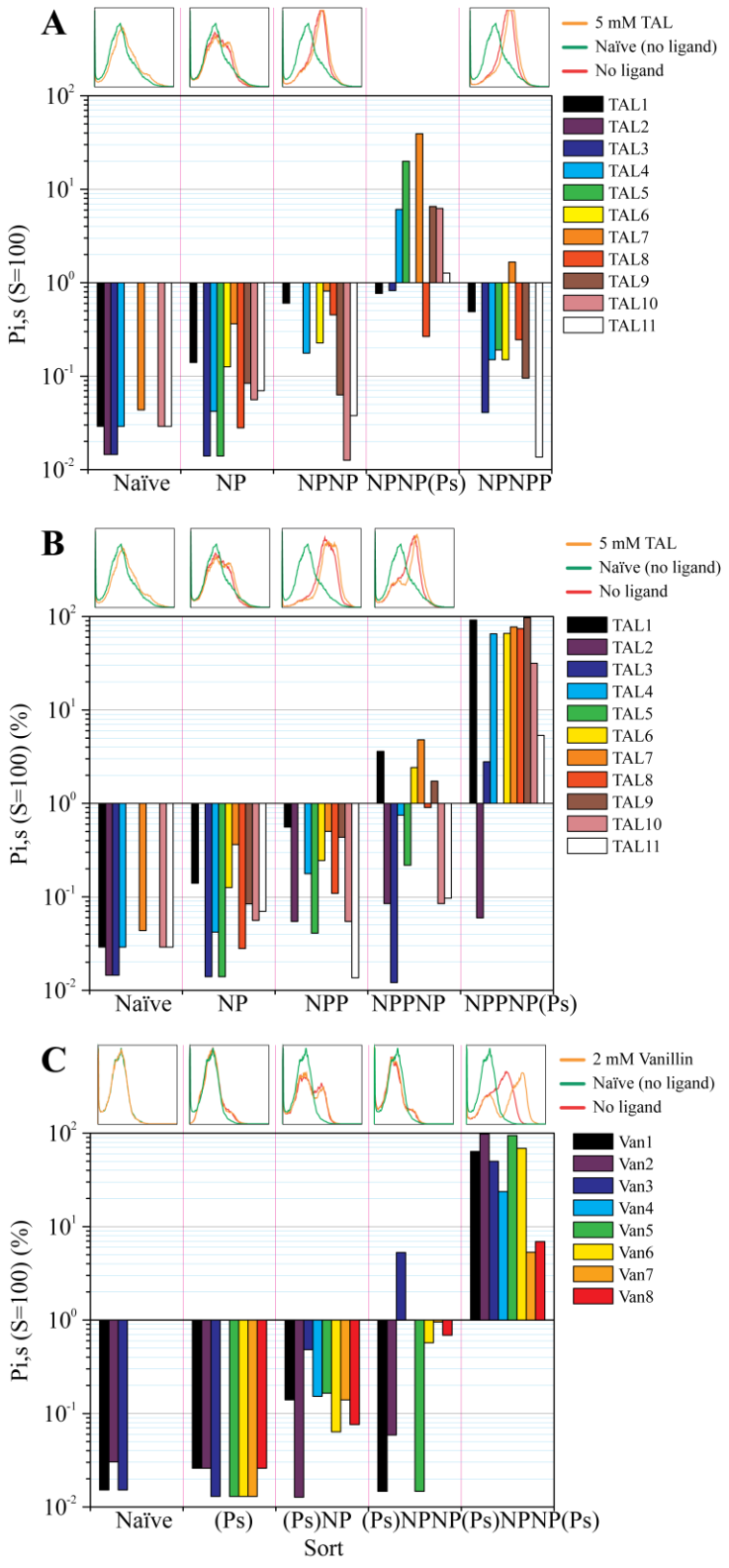


Supp. Figure E-2 Contour plots of SLib4 screened populations comparing amino acid substitutions and residue position. The colors represent the frequencies that the respective amino acid (y-axis) is present at the respective residue position (x-axis).

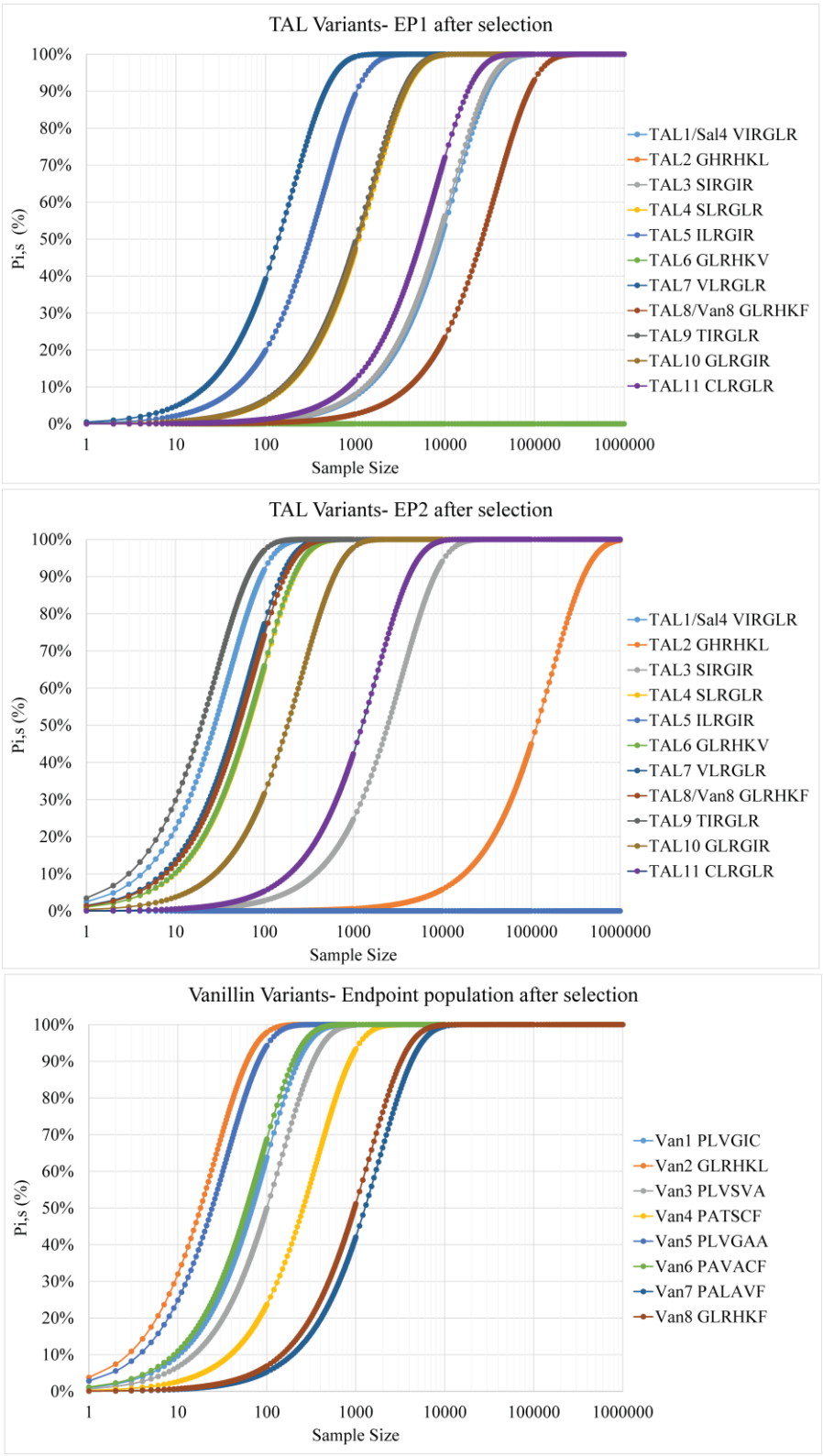
collecting 5 x 10^5 cells each round after the first positive sort. However when the frequency of the variants is taken into account and we only include variants with counts greater than 10, the average number of different variants drops greater than one orders of magnitude (1.5 x 10^4), meaning the majority of the population diversity is represented by variants with very low frequencies. These results suggest that we are achieving good enrichment after at least three positive sorts, but how does this effect our chances of isolating a functional variant in an endpoint population? Our endpoint assay was developed to minimize the number of false positive hits, which reduced the throughput of the assay. When we initially developed the assay, we assumed that the sorter was much more efficient than the high-throughput sequencing shows, so we were not as concerned about the throughput at this point in the screening. As is seen in Supp. Figure E-3 and Supp. Figure E-4, the probability of selecting the AraC-TAL and AraC-Van variants from their respective endpoints is in most cases better than 10% (most of the endpoint assays mentioned in previous chapters screened 24 different clones). The probabilities were calculated according to the equation

$$P_{i,S} = P_i\,[1 - (1 - F_i)^S], \qquad \textbf{E-1}$$

where P_i is the probability of a particular sequence i is in the library, F_i is the frequency that a particular sequence i is present in the library, and S is the sample size. Most AraC-TAL variants do not have a high probability of being isolated from the endpoint assay for both endpoint populations, EP1 and EP2. For example, the EP1 population shows $P_{i,S}$(S=24) less than 10% for all variants except for AraC-TAL7, which was not isolated from the endpoint assay of EP1. Instead, AraC-TAL5 was isolated from this population but only has a $P_{i,S}$(S=24) of 5%. AraC-TAL5 has a much lower frequency in EP2 and was not detected in the high-throughput sequencing results of EP2. The EP2 population before the final round of selection shows 36 counts and AraC-TAL10 shows 14 counts. The AraC-TAL10 variant was enriched to approximately 3200 counts after the final round of selection. Both the background and the fold-



Supp. Figure E-3 Bar graphs of probabilities ($P_{i,S}(S=100)$) for isolated variants relative to the screened population



Supp. Figure E-4 Probabilities of isolating variants relative to the endpoint assay sample size

response of AraC-TAL5 is about the average for all of the AraC-TAL variants, yet this variant was not enriched during the final round of screening. Other variants with $P_{i,S}(S=24)$ greater than all of the other isolated variants were not isolated as well. Finally, we explored the responses of the most frequent variants in an endpoint population that were not isolated from the endpoint assay, as well as the top variants after a single round of selection. We had the genes of each variant listed in Supp. Table E-2 synthesized by Invitrogen. The fragments were amplified using primers Ins-gib-for1 (GATAACAATTTCACA CAGGAGATATCATATGGC) and Ins-gib-rev1 (GGACGAAAGTAAACCCACTG). The pFG29 vector was amplified using primers pFG29-vector-for2 (GTGGGTTTACTTTCGTCCG CGCGC) and pFG29-vector-rev2 (GATATCTCCTGTGTGAAATTGTTATCCGCTCACAA TTCCACACATTATACGAGCCGATGA). The vector and gene fragment were digested with DpnI and assembled using NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, Cat. No. E2621). The variants were then tested for response to their respective target compound using the deep-well plate assay described for the endpoint assays in Chapter V. The results are reported in Supp. Table E-2. None of the top variants from the single round of selection show a response to vanillin (the target compound used during the selection), suggesting the need for further enrichment. The top variant from each population shows a response to the target compound. Some of these variants have a higher $P_{i,S}(S=24)$ than the variants that were isolated. However, none of these variants show significantly improved response compared to the variants that were isolated from the endpoint assay. These results suggest that we should screen a greater number of clones from the endpoint populations. Altogether, the high-throughput sequencing results presented here will help guide future combinatorial library screening to more efficiently isolate variants with desirable functional properties.

Supp. Table E-2 Table of responses from top most frequent variants from screened populations. Yellow boxes, variant responds to target compound; Pink boxes, variant responds to non-target compound.

Top Most Frequent Clones Not Selected

Clone Name	Population	Screened Compound	AA Rank	AA seq Freq	AA Seq	$P_{i,s}$ (S=24)	1st nt ranking	nt seq Freq	Fold response 10 mM L-ara	100 mM D-ara	5 mM TAL	5 mM Van
TAL1		5 mM TAL	1464	55	VIRGLR				1.1	0.7	12.2	1.7
MutTop1	EP1 (NPNP(Ps))	5 mM TAL	1	6954	FVRGLR	0.21	2	4352	1.1	0.7	6.4	1.3
MutTop2	EP1 (NPNP(Ps))	5 mM TAL	2	4826	GYRTLG	0.15	1	4772	0.9	0.7	0.9	1.4
MutTop3	EP1 (NPNP(Ps))	5 mM TAL	4	3393	STRPIT	0.11	4	3132	0.9	0.7	1.0	1.5
TAL1		5 mM TAL	3	20724	VIRGLR				1.1	0.7	12.2	1.7
MutTop4	EP2 (NPPNP(Ps))	5 mM TAL	2	20848	GLRGLR	0.45	35	2002	1.0	0.7	11.4	4.4
MutTop5	EP2 (NPPNP(Ps))	5 mM TAL	4	19622	TLRGLR	0.43	3	9341	1.1	0.6	10.0	2.6
MutTop6	EP2 (NPPNP(Ps))	5 mM TAL	7	10806	GHRGLR	0.27	17	2974	1.1	0.9	1.9	1.8
MutTop7	N(Ps)NPNP(Ps)	2 mM Van	1	67482	GIRHYH	0.93	1	35716	1.1	1.4	1.4	3.6
MutTop8	N(Ps)NPNP(Ps)	2 mM Van	3	24247	GLRHYH	0.60	5	13220	0.9	1.4	1.1	2.0
MutTop9	N(Ps)NPNP(Ps)	2 mM Van	5	12106	PLVHYH	0.37	13	5507	1.0	1.0	1.3	1.0

Top clones from CLib5s (Ps)-once not selected

Clone Name	Population	Screened Compound	AA Rank	AA seq Freq	AA Seq	Prob EP tested	nt ranking	nt seq Freq				
MutTop10	N(Ps)	2 mM Van	2	8825	VIRGYH	0.24	2	8738	0.8	0.7	0.9	1.2
MutTop11	N(Ps)	2 mM Van	3	4663	PTRGYH	0.14	3	4029	0.8	0.9	1.0	1.1
MutTop12	N(Ps)	2 mM Van	4	3215	VIRHYH	0.10	4	2969	1.2	1.7	1.3	1.2