

Building artificial genetic circuits to understand protein function

Louis H. Scott^{a,b}, James C. Mathews^{a,b}, Aleksandra Filipovska^{a,b,c},
Oliver Rackham^{a,d,e,*}

^aHarry Perkins Institute of Medical Research, Nedlands, WA, Australia

^bThe University of Western Australia Centre for Medical Research, Crawley, WA, Australia

^cSchool of Molecular Sciences, The University of Western Australia, Crawley, WA, Australia

^dSchool of Pharmacy and Biomedical Sciences, Curtin University, Bentley, WA, Australia

^eCurtin Health Innovation Research Institute, Curtin University, Bentley, WA, Australia

*Corresponding author: e-mail address: oliver.rackham@curtin.edu.au

Contents

1. Introduction	2
2. General principles	3
3. Methodology	5
3.1 Genetic circuits to select for protein activity	5
3.2 Mutant library construction	9
3.3 Mutant library analysis	14
4. Summary	17
Acknowledgments	17
References	17

Abstract

Intrinsic protein properties that may not be apparent by only examining three-dimensional structures can be revealed by careful analysis of mutant protein variants. Deep mutational scanning is a technique that allows the functional analysis of millions of protein variants in a single experiment. To enable this high-throughput technique, the mutant genotype of protein variants must be coupled to a selectable function. This chapter outlines how artificial genetic circuits in the yeast *Saccharomyces cerevisiae* can maintain the genotype-phenotype link, thus enabling the general application of this approach. To do this, we describe how to engineer genetic selections in yeast, methods to construct mutant libraries, and how to analyze sequencing data. We investigate the structure-function relationships of the antimicrobial resistance protein TetX to illustrate this process. In doing so, we demonstrate that deep mutational scanning is a powerful method to dissect the importance of individual residues for the inactivation of antibiotic analogues, with consequences for the rational design of new drugs to combat antimicrobial resistance.



1. Introduction

Dissecting the relationship between a protein's structure and its activity is vital for understanding how basic biological processes occur and has allowed us to improve upon or evolve new enzymatic reactions and develop new drugs (Bornscheuer et al., 2012; Lutz, 2010; Steiner & Schwab, 2012; van Montfort & Workman, 2017). Three-dimensional snapshots of a protein's structure in distinct states as shown from X-ray crystallography (Shi, 2014) or cryo-electron microscopy (Cheng, Grigorieff, Penczek, & Walz, 2015) can illustrate conspicuous interactions of residues and ligands. However, as proteins are not static structures, their important dynamic processes can be investigated using protein nuclear magnetic resonance spectroscopy (Marion, 2013). While these methods alone are required for elucidating a protein's three-dimensional structure, mutational studies are equally important for directly investigating the function of specific residues. For example, alanine scanning can establish the importance of individual amino acids to protein stability and function (Cunningham & Wells, 1989; Morrison & Weiss, 2001). Alternatively, mutagenesis to the binding pocket or around the catalytic residues of enzymes can give insight into the specifics of catalysis, or a glimpse into possible promiscuity that can be appropriated into new functions (Gupta & Varadarajan, 2018). While these methods rely on analytical validation to link mutated genotype to phenotype, advances in synthetic biology offer an alternative when convenient phenotypic readouts are not available.

Genetic circuits that follow computational logic, where experimental inputs are translated into selectable or screenable outputs, have been engineered in many model organisms (Brophy & Voigt, 2014; Filipovska & Rackham, 2008; Liu, Yuan, & Stewart, 2013; Weber & Fussenegger, 2010). Using genetic circuitry, if a specific input can be linked to the survival of a microbe, then a single cell can act as an independent computational laboratory; thus, within a single culture, tens of millions of experiments can be run and analyzed concurrently. Therefore, coupling the genotype of mutant protein variants to the survival of a microbe is a powerful technique to study mutations, obviating the need for laborious analytical screening. Deep sequencing of resulting mutant libraries can then generate large datasets to reveal intrinsic protein properties possibly too subtle to otherwise gauge, a technique known as "deep mutational scanning" (Araya & Fowler, 2011; Fowler et al., 2010; Fowler & Fields, 2014; Fowler, Stephany, & Fields, 2014). The yeast *Saccharomyces cerevisiae*

is well suited as a microscopic laboratory for this purpose as it is easy to culture, easy to genetically modify, and has many well-characterized biological parts shown to follow computational logic when combined (Blount, Weenink, & Ellis, 2012).

This chapter will describe how to design genetic circuits in yeast to probe the structure and function relationships of a protein of interest. We will explain how variants in random mutant libraries can be selected for the gain, loss, or retention of protein activity, and how deep mutational scanning can reveal what and how prevalent these mutations are. We will illustrate this process by probing the structure and function relationships of residues within TetX, an emerging antimicrobial resistance threat that causes the enzymatic inactivation of all known tetracycline antibiotics (Scott, Mathews, Flematti, Filipovska, & Rackham, 2018). Understanding the mutability of specific residues within a protein has broad consequences for understanding how it functions; in the case of an antimicrobial resistance protein, this understanding can guide the development of vitally important new drugs.



2. General principles

We propose three general steps to use genetic circuits in yeast to probe the structure and function relationships of a protein, as illustrated in Fig. 1. These steps involve engineering a genetic selection for protein activity, making a library of protein variants, and then analyzing the genotype of escape mutants.

The first crucial step is to link protein function to cell survival. This requires careful consideration, as how to engineer this association may not be directly recognizable. Fortunately, yeast strains already engineered to assay macromolecular interactions (for example, yeast two-hybrid systems) can be appropriated for part of this genetic circuitry (Fields & Song, 1989; Licitra & Liu, 1996; Van Criekeing & Beyaert, 1999; Wallis, Filipovska, & Rackham, 2018; Wallis, Richman, Filipovska, & Rackham, 2018). In these strains, when an experimental macromolecular interaction occurs, a functional transcription factor is created that activates the expression of a reporter gene. Therefore, if the activity of a protein of interest can also be linked to the expression of such a transcription factor (or generally, a “master regulator”) in one of these strains, then this activity should be selectable. A wealth of engineered yeast one-hybrid, two-hybrid and three-hybrid strains containing a multitude of different reporter genes capable of both

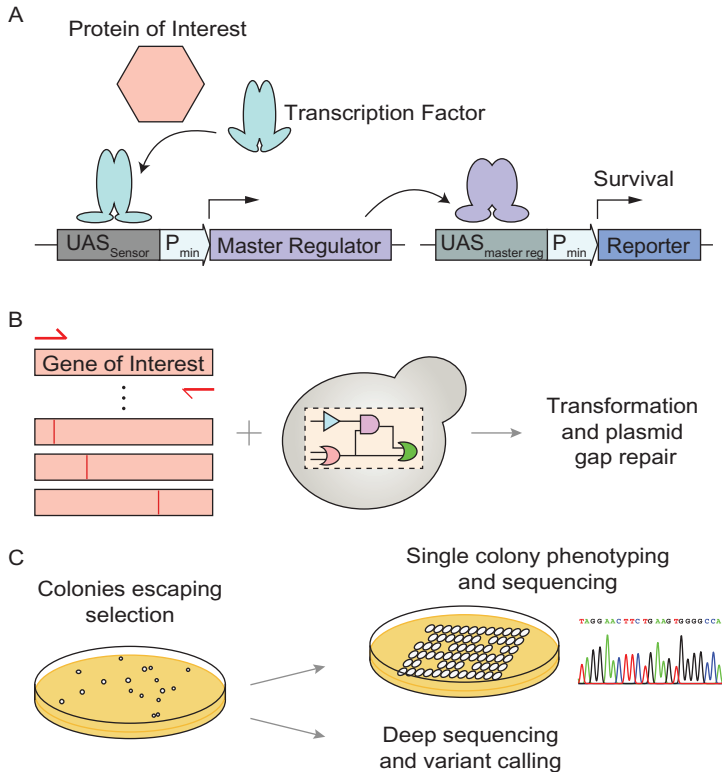


Fig. 1 The function of a protein of interest can be selected for if it activates a transcription factor that in turn activates a master regulator that turns on downstream reporter genes (A). A mutant library, generated for example by error-prone PCR, can be introduced into yeast engineered with genetic circuitry described in (A) and assembled *in vivo* with an expression vector by homologous recombination (B). Library transformants escaping selection can be analyzed as single clones by phenotypic screens in conjunction with Sanger sequencing, or all escape mutants can be analyzed in parallel with deep mutational scanning (C).

positive and negative selections is available, and this existing well-classified foundation can be used to build more complex genetic circuits to assay a specific protein function.

Second, a library of mutant proteins must be expressed in yeast. The mutational strategy chosen will be a trade-off between mutational depth versus amino acid sequence coverage and will need careful attention based on the size of the protein and on available structural or mutational information.

Third, mutants escaping selection should be analyzed. While very labor intensive, the analysis of individual escape mutants provides a direct link

between genotype and phenotype and should be done with a manageable number of variants. Importantly, deep sequencing of all selected mutants will give a general view of the protein's mutational landscape. Using deep mutational scanning as a high-throughput approach can reveal intrinsic protein properties that would be missed by conventional sampling (Fowler & Fields, 2014).



3. Methodology

We recently used deep mutational scanning in yeast to investigate the structure and function of the antimicrobial resistance enzyme TetX (Scott et al., 2018). TetX is a flavin-dependent monooxygenase that can degrade all known tetracycline antibiotics (Yang et al., 2004). However, as not all tetracyclines are degraded at the same rate, there was potential to redesign current drugs to be better inhibitors and bypass this resistance mechanism. To investigate this, we designed genetic circuits that linked the function of TetX to the survival of yeast. Random mutant libraries of TetX were then selected for variants that could degrade tetracycline analogues or not. Finally, using deep mutational scanning, we investigated the structural requisites of TetX, dependent on the tetracycline analogues it could degrade. Combined with research into the structure of TetX and related enzymes (Moore, Hughes, & Wright, 2005; Park et al., 2017; Volkers et al., 2013; Volkers, Palm, Weiss, Wright, & Hinrichs, 2011), our study provides valuable information to be leveraged to rationally redesign tetracycline antibiotics to potentially overcome enzymatic destruction.

The next sections describe a methodological framework to study structure–function relationships of a protein using genetic circuits and deep mutational scanning in yeast. We describe this methodology using our recent study on the antimicrobial resistance protein TetX as example.

3.1 Genetic circuits to select for protein activity

3.1.1 A biosensor to select for TetX function

The following sections outline how an artificial genetic circuit can select for protein function in engineered yeast. As a starting point, one must test if the protein of interest is functionally expressed. Therefore, analytical techniques that give a direct readout on protein function should be used. Expression from a strong constitutive or inducible promoter in these initial experiments should ensure some detectable protein activity. Furthermore, codon usage optimization is recommended for all genes not originating from yeast—most

gene synthesis companies offer this service. In our recent work, we found expression of the yeast codon usage optimized gene for TetX from the *ADH1* promoter on a 2-micron plasmid was sufficient to destroy the tetracycline anhydrotetracycline (ATC) in yeast cultures, as detected by high performance liquid chromatography-mass spectrometry (HPLC-MS) (Scott et al., 2018). While a similar expression configuration can be an initial foundation for circuit design, expression may need to be adjusted during the final library selection to ensure enough sensitivity to detect mutations that cause subtle changes in function. Many endogenous yeast promoters spanning orders of magnitude in activity have been characterized and can be appropriated for this purpose (Keren et al., 2013; Lee, DeLoache, Cervantes, & Dueber, 2015).

Once protein function in yeast is confirmed, the next crucial step is to link this activity to the expression of a “master regulator.” A master regulator is simply a transcription factor that can activate downstream reporter genes that enable genetic selection. Many yeast-hybrid strains have reporter genes engineered to be activated by transcription factors, such as the yeast two-hybrid strain MaV204K that uses Gal4 to activate *HIS3* and *URA3* auxotrophic markers (Ito et al., 2000). These reporter genes can be selected for on media lacking histidine or uracil, but also against using 5-fluoroorotic acid (5-FOA) as it is transformed into a toxic product by *URA3*. Therefore, the function of TetX, or generally any other protein of interest, can be selected for or against if it can induce *GAL4* expression in MaV204K.

As TetX functions by degrading tetracyclines, a tetracycline inducible gene expression system was used to express the master regulator Gal4 in MaV204K (Scott et al., 2018). We built a genetic circuit where the reverse tetracycline-controlled transactivator (rtTA) can sense tetracyclines and bind to promoter elements upstream of the *GAL4* gene, encoding the Gal4 transcription factor. If this strain is cultured with tetracyclines, Gal4 expression is turned on, which in turn activates *HIS3* and *URA3* expression. If TetX is expressed in this strain, it will degrade the tetracycline inducer before it can stimulate Gal4 expression and downstream auxotrophic marker activation. Therefore, using this system, the function of TetX can be selected for or against dependent on the culture conditions. Integral to consistent results of such a selection is to use a centromeric plasmid that replicates as a single copy to avoid possible variability in Gal4 expression confounded by copy number.

Our initial exploratory experiments demonstrated that induction of Gal4 expression by ATC caused toxicity to yeast (Scott et al., 2018).

We hypothesized this may be caused by pleiotropic effects of Gal4 over-expression (Keren et al., 2016). One way to overcome this is to modulate the translation of the master regulator using strong and weak Kozak consensus sequences or start codons. This can easily be achieved by site directed mutagenesis without having to replace promoters (Chen, Lin, Chang, Yeh, & Wang, 2008). Subtle effects on translation can be made by altering the six nucleotides preceding the start codon, for example, to cytosines, thus creating a weak Kozak consensus sequence. Alternatively, changing the start codon from ATG to the less commonly used TTG can cause a more pronounced reduction in translation. Using this technique, a yeast strain conditionally expressing Gal4 with a TTG start codon enabled yeast to survive on media lacking histidine and uracil in the presence of ATC, thus creating cells that were dependent on tetracyclines for survival. When TetX was expressed in this strain it could inactivate ATC, thus prohibiting Gal4 expression from triggering of the nutritional marker expression cascade.

3.1.2 General genetic circuit design considerations

A transcriptional regulator, that operates through the tetracycline repressor (TetR), was used to sense the presence of tetracyclines in our recent work. Nevertheless, one is not constricted to this class of antibiotics as biosensors have been designed against a myriad of other targets (Dietrich, McKee, & Keasling, 2010). Therefore, if a biosensor exists against a protein or ligand of interest, then the structure and function of a protein that modulates it can be explored using the methodology presented in this chapter. Furthermore, the concept of utilizing a transcriptional regulator as a sensor is generally applicable to many other antibiotic and small molecule altering enzymes far beyond those already detectable by a biosensor. Most biosynthetic clusters encode a self-resistance mechanism that is transcriptionally regulated by the metabolite produced. This duality concept has been exploited to discover new natural products via genome mining (Yan et al., 2018). Furthermore, there are currently over 200,000 TetR family transcriptional regulators in public sequence databases that recognize many diverse small molecules, including numerous desirable natural products that are modified by a multitude of accessory proteins (Ramos et al., 2005). The structures of about 200 of these transcriptional regulators are known, and protein engineering approaches have shown that they can be redesigned to bind new ligands (Tang & Cirino, 2011; Tang et al., 2013). Therefore, the concept of linking the activity of a protein that alters a natural product,

via a transcriptional regulator, to life-death selections in yeast should be generally applicable to most natural products of interest.

While Gal4 functions well in the context of available yeast-hybrid strains, toxicity from overexpression is a confounding factor that could be an issue (Keren et al., 2016). Many yeast two-hybrid strains have operator sites for the LexA repressor upstream of reporter genes; therefore, expressing a fusion of the LexA repressor with an activating domain could overcome potential pleotropic Gal4 effects (Fashena, Serebriiskii, & Golemis, 2000). Furthermore, expanding the repertoire of reporter strains to encompass fully orthogonal components will be highly beneficial for genetic circuit design. Strain engineering is becoming even easier with CRISPR-Cas9 systems (Jessop-Fabre et al., 2016), and future iterations should make use of the expanding collection of well-characterized genetic parts (Naseri et al., 2017).

In terms of practicalities, a biosensor should have a wide operational and dynamic range (see Fig. 2) (Dietrich et al., 2010; Rogers, Taylor, & Church, 2016). Typically, high molecular specificity should also be sought after (Lederer et al., 1996). However, the promiscuous nature of TetR used to induce tetracycline dependent Gal4 expression allowed the analysis of different inducers in our recent study, emphasizing that choice of parts during genetic circuit construction is highly dependent on the experimental question asked. Another important aspect for genetic circuit design is the ability for rapid prototyping, or failing quickly, as this will drastically decrease the

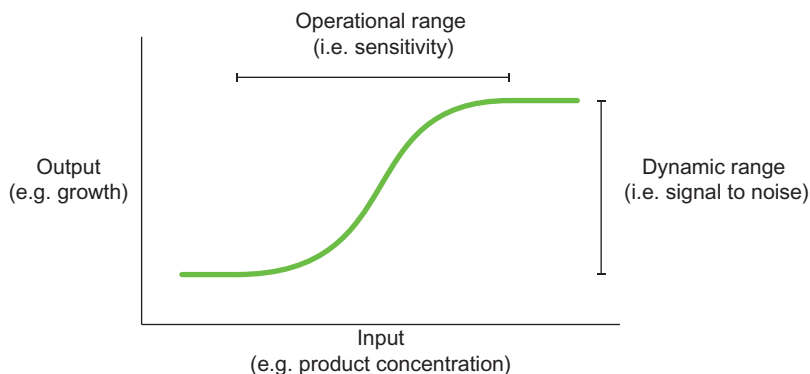


Fig. 2 Input concentrations over which the biosensor responds is the operational range. It spans the limit of detection to the saturation point. The dynamic range is the output to any given concentration of input. It can be measured in cell growth or fluorescence depending on the reporter gene(s) chosen. A large dynamic range will ensure the signal resulting from the experimental input can be detected over background noise.

design-build-test cycle required to engineer a suitable selection. To aid rapid prototyping, yeast survival assays are a quick semi-quantitative way to assess genetic circuit performance. However, thorough exploratory experiments are essential to large-scale library screening to understand how the selection mechanism will behave *in situ*. Consequently, more detailed assessment of short-listed selections should be done mimicking the final selection conditions.

3.2 Mutant library construction

Before constructing a mutant library, planning is required to ensure that it contains useful answers to your research questions. Choice of mutational strategy is a key decision when planning deep mutational scanning, with saturation or random mutagenesis being two broad classes of mutational strategy. Saturation mutagenesis is the exhaustive examination of amino acid changes within a discrete sequence region. This approach, enabled by efficient chemical synthesis of combinatorial DNA libraries (Medina-Cucurella et al., 2019) or by polymerase chain reaction (PCR) with trimer phosphoramidites (Filipovska, Razif, Nygård, & Rackham, 2011), offers comprehensive and largely unbiased exploration of the sequence space but is limited by practical constraints to targeting short protein sequences.

Random mutagenesis strategies include a variety of techniques that introduce DNA changes by chance. For example, mutations can be introduced by modified PCR (Leung, Chen, & Goeddel, 1989), by treatment with chemical or radiological agents (Bose, 2016), or by introducing target sequences to “mutator” strains of organisms with inherently low fidelity of DNA replication (Muteeb & Sen, 2010). Although libraries produced by random strategies will not be as comprehensive or controlled as saturation mutagenesis libraries, random strategies are often simpler to implement and can more easily target longer sequences of whole genes or even whole genomes. Furthermore, random mutagenesis also more closely approximates the conditions by which mutations arise in nature, and so is a powerful tool for evolving proteins with new or improved functions.

3.2.1 Error-prone polymerase chain reaction libraries

Error-prone polymerase chain reaction (EP-PCR) is a tried and tested method for introducing random mutations to full-length proteins where reaction conditions can be changed to fine tune the mutation rate. Overall mutation rate and library size are two key considerations controlling diversity within the library: if the mutation rate is too low and/or the library

is too small, amino acid changes will not be sufficiently diverse to provide useful data (especially when synonymous mutations are expected). If the rate is too high, there will be many loss of function mutants carrying multiple mutations, making it difficult to attribute function loss to any one mutation. For these reasons, aiming for a mutation rate close to two substitutions per kilobase can maximize the number of transformants carrying proteins with just one productive amino acid change. At this rate, if mutation frequency is assumed to follow a Poisson distribution, roughly 14% of clones will be not mutated, roughly 27% will carry one mutation, roughly 27% will carry two mutations, and the remainder more than two. Although there are many methods for adjusting mutation rate, the simplest to modify is number of PCR cycles with an error-prone polymerase—this parameter reportedly titrating mutation rates between 0.01% and 10% per base position (Vanhercke, Ampe, Tirry, & Denolf, 2005).

Another key consideration for EP-PCR library design is reducing mutation bias. DNA polymerase enzymes are naturally biased toward certain nucleotide substitutions, notably $AT \rightarrow GC$ and $AT \rightarrow TA$. Although a limited repertoire of base substitutions can theoretically still result in a very diverse library of amino acid substitutions, it is advisable to consider methods for reducing this bias to allow detection of rare mutant phenotypes. Common methods for controlling mutation bias are addition of Mn^{2+} up to 5mM in the reaction buffer (Lin-Goerke, Robbins, & Burczak, 1997), changed ratio of deoxynucleoside triphosphates (dNTPs) (Cadwell & Joyce, 1992), and combining multiple polymerase enzymes with different mutational preferences (Vanhercke et al., 2005). For our example TetX library, we chose to experiment with a novel method for combatting mutational bias. By using an EP-PCR template obtained by fairly low fidelity gene fragment synthesis (7.52×10^{-4} errors per base position) additional non-polymerase biased mutations can be introduced into the library (Invitrogen, 2015).

When planning an EP-PCR library size target for a whole open reading frame, one should maximize the completeness of the library, so that it contains as many protein sequences as possible that differ from the wild type by a single amino acid. The proportion of these sequences present can be thought of as the “completeness” of the library, and the maximum number of single amino acid substitutions calculated by multiplying the length of the protein sequence by the 19 alternative amino acids possible at each position. Since the start codon ordinarily needs to be preserved, the amino acid sequence length can be reduced by one for this calculation. For example, at a rate

of two mutations per kilobase for the *TETX* gene, an ideal EP-PCR library containing codons for all 7353 possible single amino acid substitutions would require the impractical 2×10^{11} yeast transformants to completely sample. Knowing that large-scale yeast transformations typically yield closer to 10^7 transformants, the practical library completeness would be closer to 57% as calculated by the online tool PEDEL-AA (Firth & Patrick, 2008). It should also be considered at this point that library size targets are limited by the requirements of downstream sequencing experiments. When using next-generation sequencing technologies, it is advisable for each variant to be covered by at least 100 sequencing reads for accurate quantitative analysis (Fowler & Fields, 2014).

Selecting a method for incorporation of genes of interest into the host also requires some thought. One advantage of using yeast is the chance to harness its robust homologous recombination machinery to assemble vectors *in situ*. When using this technique, one should, however, be aware that some transformants may carry re-ligated vector, inflating the usable transformation efficiency (Chandran & Shapland, 2017). For our example TetX library, we transformed variant *TETX* gene sequences along with linear copies of a yeast shuttle vector to place *TETX* expression under control of the constitutive *ADH1* promoter.

3.2.2 Small-scale library testing

Small-scale pilot library experiments are recommended to investigate the quality of the library and determine the reagents required before up-scaling. Key considerations that can be optimized in a small-scale experiment include transformation efficiency, mutation rate, and mutation type. During small-scale testing, it is useful to test various ratios of vector and insert DNA for their transformation efficiencies to select conditions that maximize yield during larger experiments. For our TetX pilot study, insert to vector ratios of 2:1 and 1:1 were transformed into yeast by the lithium acetate/single-stranded carrier DNA/polyethylene glycol method (Gietz & Schiestl, 2007; Gietz & Woods, 2002), then inoculated as serial dilutions on agar plates selecting only for plasmid maintenance. Colony counts revealed that an insert to vector ratio of 2:1 yielded the most transformants in our experimental conditions. We recommend, however, testing even more ratios, for example, up to an insert to vector ratio of 5:1.

Using Sanger sequencing, small-scale transformation tests are an excellent opportunity to compare the achieved mutation rate with the original target. PCR conditions can then be adjusted to titrate this rate if necessary.

Sanger sequencing of yeast transformant colonies from our TetX pilot library found mutation rates of 1.5 nucleotide substitutions per kilobase, resulting in an average 1.2 mutations per gene. Encouragingly, no clones carrying empty re-ligated vector were detected.

As well as mutation rate, the type of mutations occurring in a small set of EP-PCR transformants will provide key information on the strength of your mutational strategy. For our example TetX library, it was found that for 25% of those mutants sequenced, insertions and deletions occurred in a single 9bp adenine homopolymer region of the gene (Fig. 3). If we proceeded with large-scale experiments, these indels would result in many missense mutants that would be non-functional and escape selection—overwhelming rarer loss of function mutants that arose from point mutations. In this case, it was simple to redesign the synthetic *TETX* gene template with two AAA codons replaced by the synonymous AAG to disrupt the homopolymer. This example demonstrates the importance of small-scale tests before attempting more substantial EP-PCR protocols.

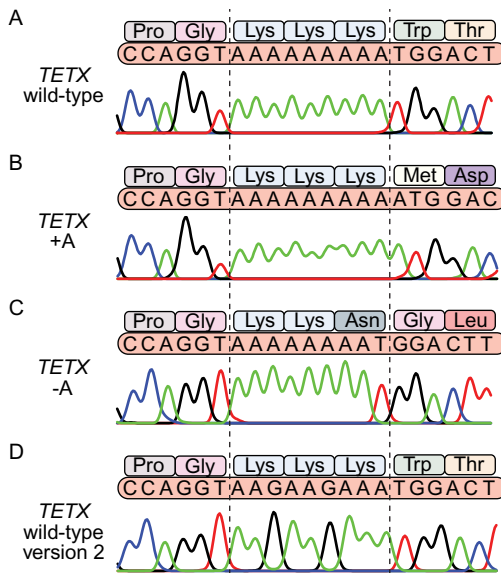


Fig. 3 Small-scale tests are crucial to the success of large-scale EP-PCR libraries. In this example, an unexpectedly high (25%) rate of insertions and deletions occurring in a single adenine homopolymer sequence (A) flooded the library with nonsense loss of function mutations (B and C). To correct this, it was trivial to replace two AAA codons with the synonymous AAG in the synthetic gene template (D).

3.2.3 Large-scale library screening

Large-scale library construction parameters should be extrapolated from results achieved during the small-scale pilot experiment. The small-scale experiment should give information on the lowest amount of DNA required to still deliver the most transformants possible that can be analyzed with downstream techniques. This is important as, although rare, increasing the DNA amount can lead to multiple variants being introduced to the same cell, decoupling the phenotype-genotype link. It is therefore recommended to scale up the whole transformation reaction, and not just the DNA amount. We recommend up-scaling using the lithium acetate/single-stranded carrier DNA/polyethylene glycol yeast transformation method which should yield tens of millions of clones (Gietz & Schiestl, 2007; Gietz & Woods, 2002). This process is highly contamination sensitive due to large volumes of rich yeast media undergoing repeated incubations, and the use of laminar air flow is recommended.

Once transformed, the library should be allowed to undergo additional doublings before plating on selective media. This is an important step that provides a clonal population of each mutant. Having multiple cells containing the same mutant ensures that even those that occur rarely in the population have a chance of being sampled. It also means that the entire diversity of the library can be captured on multiple selection conditions, and crucially on a non-selective condition to provide an un-selected representation of the library, which is necessary for calculating the enrichment of mutations. Additionally, multiple biological replicates should be constructed and selected. Comparing mutant enrichment between replicates using purpose-built software such as Enrich2 will ensure the planned library diversity is captured and can also decrease noise and false positives from inconsistent escape mutants (Rubin et al., 2017).

In our recent work, additional doublings of the library allowed TetX mutants that could not inactivate different tetracyclines (ATC or doxycycline (DOX)) to be selected. This was done using media lacking tyrosine and lysine, to select for TetX expression and biosensor plasmids, and lacking uracil and histidine, to select for reporter activation because of enzyme function loss. Furthermore, we also conducted some selection experiments using media lacking tyrosine and lysine, supplemented with ATC and 5-FOA—to select for functional TetX variants only—demonstrating the flexibility of this selection system (Scott et al., 2018). Aliquots of the library were also taken following rescue of transformants and after the

final additional doublings, to calculate final transformation efficiency and library size by serial dilution survival assays on media selecting for plasmid maintenance only.

3.3 Mutant library analysis

3.3.1 *Single colony analysis*

In conjunction with large-scale library analysis using deep sequencing, it is useful to isolate mutant strains with interesting phenotypes by traditional replica plating techniques. Doing so will offer a chance to study these mutants' activity in greater detail and against a greater variety of substrates. The enrichment state of these phenotypes in deep sequencing data can also serve as validation for selection, sequencing, and analysis strategies. Isolated mutants with studied activity levels offered a valuable point of comparison with our *TETX* deep sequencing data, their enrichment in the larger data set confirmed the validity of the deep sequencing data and indicated the relative level of activity loss that corresponded to various enrichment values. Following selection from a portion of our TetX library, yeast colonies that escaped selection were inoculated on selective agar plates with alternative enzyme substrates, in this case ATC and DOX. By comparing growth between these plates, it was possible to identify mutant enzymes with a preference for one substrate. Sequencing these mutations and modeling them with molecular visualization software provided crucial structure-function insights (Scott et al., 2018). Care must be taken to retain part of each sampled colony on the original selection plate to ensure their representation during deep sequencing.

3.3.2 *Whole population analysis*

3.3.2.1 Generation of sequencing data

As it is not possible to individually sequence and phenotypically screen each mutant, deep sequencing the whole mutant population and comparing this with an unselected population can reveal which mutations are enriched after a selection condition. There are many sequencing technologies and protocols suitable for this purpose; however, choices here will be governed by availability, cost, and experimental needs. For shorter gene sequences or smaller libraries, Sanger sequencing many individual mutants in parallel remains a reliable and robust option. However, for longer gene sequences and larger populations containing rare mutant phenotypes, next-generation

sequencing technologies are more appropriate. Of these technologies, sequencing by synthesis (Illumina), pyrosequencing (454), and sequencing by ligation (SOLiD) reportedly give the highest accuracy (99.9% for all three) (Liu et al., 2012).

For our example TetX library, we chose to use Illumina MiSeq with paired-end reads, due to its ready availability and low cost per base. Since deep mutational scanning is a quantitative method relying on detection of rare phenotypes, we aimed to maximize the coverage depth achieved in the *TETX* open reading frame using sub-amplicon sequencing. This is a technique where primers tagged with known Illumina compatible sequences are used to amplify a series of short products spanning the target (Hiatt, Patwardhan, Turner, Lee, & Shendure, 2010; Starita et al., 2013). These amplicons can then be applied directly to the flow cell as paired-end reads. For our example TetX library, we found that 250bp sub-amplicon reads provided the greatest coverage depth and most accurate results. These reads were generated from mutants in an unselected condition, escape mutants from ATC and from DOX selections (selection for non-functional TetX), and escape mutants from selection with ATC and 5-FOA (selection for functional TetX).

3.3.2.2 Analysis of sequencing data

After optimization of a sequencing strategy and sequencing of your library, you will need software to convert raw read data into useful information about the mutants that make up your library. In general, the process of analysis will include pre-processing of reads for quality, alignment of reads to the reference DNA sequence, and finally calling mutations and calculating enrichment scores for mutations compared to an unselected population. Naturally, the choice of bioinformatics tool(s) will depend on the sequencing technology used, characteristics of the library, characteristics of the target gene(s), and the nature of your research questions. For many applications, obvious choices include the widely used open-source toolkits SAMtools (Li et al., 2009) and GATK (McKenna et al., 2010). If these platforms do not meet the specific needs of your experiment, there are specialized open-source tools available from code repositories developed by researchers who have needed to extend the reach of their deep sequencing and deep mutational scanning experiments. Enrich2 and DMStools are examples of purpose-built software that will treat sequencing data with a statistically

principled approach to generate plots for deep mutational scanning data visualization (Bloom, 2015; Rubin et al., 2017). Recently, machine learning algorithms have been developed to incorporate variant function prediction of large mutant datasets (Gray, Hause, Luebeck, Shendure, & Fowler, 2018).

For our example TetX library, we performed initial read quality and adaptor trimming, as well as extraction and realignment of subamplicons, using the proprietary software Geneious (Kearse et al., 2012). Enrich2 was then used to calculate enrichment scores for mutations across the protein sequence, and these enrichment scores visualized graphically as well as mapped onto protein structure using PyMOL (Schrödinger LLC, 2018).

Mutational bias can also be investigated at this stage. When the frequency of all possible base substitutions in the *TETX* library was calculated, TA→CG transversions were most frequently observed, followed by GC→AT transversions (Fig. 4). This trend is consistent with the known mutation bias of Taq and with the relatively low GC content of the target gene (Vanhercke et al., 2005). Encouragingly, a low but consistent rate of most other substitutions was detected in the library, perhaps resulting from our novel approach to reduce mutation bias.

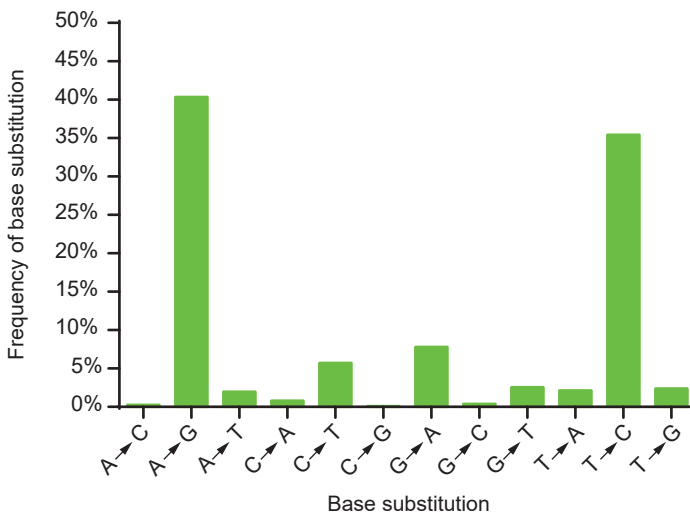


Fig. 4 Frequencies of base substitutions occurring in an EP-PCR library following selection with an artificial yeast genetic circuit are generally consistent with expectations for Taq polymerase. Low but consistent levels of rare mutations were attributable to the use of a low fidelity gene synthesis template for the reaction.



4. Summary

Artificial genetic circuits in yeast can provide a valuable tool to link mutant genotype to protein function. In doing so, the analysis of tens of millions of variants can be achieved in parallel, a technique known as deep mutational scanning. This chapter describes this process using the analysis of tetracycline resistance protein TetX variants, as an example. We describe in detail how to engineer genetic selections for protein function, how to create libraries of mutant proteins, and how deep sequencing can reveal the relevance of specific residues for protein activity. Investigation of the structure and function of TetX provides insights into how new drugs could be made to combat this resistance mechanism, illustrating the power of this methodology.

Acknowledgments

This work was supported by fellowships, scholarships and grants from the Australian Research Council (DP180101656 to A.F. and O.R.), the National Health and Medical Research Council (APP1154646 to A.F. and APP1154932 to O.R.), and the Cancer Council Western Australia (to A.F. and O.R.).

References

- Araya, C. L., & Fowler, D. M. (2011). Deep mutational scanning: Assessing protein function on a massive scale. *Trends in Biotechnology*, 29, 435–442.
- Bloom, J. D. (2015). Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16, 168.
- Blount, B. A., Weenink, T., & Ellis, T. (2012). Construction of synthetic regulatory networks in yeast. *FEBS Letters*, 586, 2112–2121.
- Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., & Robins, K. (2012). Engineering the third wave of biocatalysis. *Nature*, 485, 185–194.
- Bose, J. L. (2016). Chemical and UV mutagenesis. In J. Bose (Ed.), *Methods in molecular biology: Vol. 1373. The genetic manipulation of Staphylococci* (pp. 111–115). New York, NY: Humana Press.
- Brophy, J. A. N., & Voigt, C. A. (2014). Principles of genetic circuit design. *Nature Methods*, 11, 508–520.
- Cadwell, R. C., & Joyce, G. F. (1992). Randomization of genes by PCR mutagenesis. *PCR Methods and Applications*, 2, 28–33.
- Chandran, S., & Shapland, E. (2017). Efficient assembly of DNA using yeast homologous recombination (YHR). In R. Hughes (Ed.), *Methods in molecular biology: Vol. 1472. Synthetic DNA* (pp. 187–192). New York, NY: Humana Press.
- Chen, S.-J., Lin, G., Chang, K.-J., Yeh, L.-S., & Wang, C.-C. (2008). Translational efficiency of a non-AUG initiation codon is significantly affected by its sequence context in yeast. *The Journal of Biological Chemistry*, 283, 3173–3180.
- Cheng, Y., Grigorieff, N., Penczek, P. A., & Walz, T. (2015). A primer to single-particle cryo-electron microscopy. *Cell*, 161, 438–449.

- Cunningham, B., & Wells, J. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, *244*, 1081–1085.
- Dietrich, J. A., McKee, A. E., & Keasling, J. D. (2010). High-throughput metabolic engineering: Advances in small-molecule screening and selection. *Annual Review of Biochemistry*, *79*, 563–590.
- Fashena, S. J., Serebriiskii, I. G., & Golemis, E. A. (2000). LexA-based two-hybrid systems. *Methods in Enzymology*, *328*, 14–26.
- Fields, S., & Song, O. (1989). A novel genetic system to detect protein–protein interactions. *Nature*, *340*, 245–246.
- Filipovska, A., & Rackham, O. (2008). Building a parallel metabolism within the cell. *ACS Chemical Biology*, *3*, 51–63.
- Filipovska, A., Razif, M. F. M., Nygård, K. K. A., & Rackham, O. (2011). A universal code for RNA recognition by PUF proteins. *Nature Chemical Biology*, *7*, 425–427.
- Firth, A. E., & Patrick, W. M. (2008). GLUE-IT and PEDEL-AA: New programmes for analyzing protein diversity in randomized libraries. *Nucleic Acids Research*, *36*, W281–W285.
- Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., et al. (2010). High-resolution mapping of protein sequence–function relationships. *Nature Methods*, *7*, 741–746.
- Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: A new style of protein science. *Nature Methods*, *11*, 801–807.
- Fowler, D. M., Stephany, J. J., & Fields, S. (2014). Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature Protocols*, *9*, 2267–2284.
- Gietz, R. D., & Schiestl, R. H. (2007). Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols*, *2*, 38–41.
- Gietz, R. D., & Woods, R. A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods in Enzymology*, *350*, 87–96.
- Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., & Fowler, D. M. (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Systems*, *6*, 116–124.e3.
- Gupta, K., & Varadarajan, R. (2018). Insights into protein structure, stability and function from saturation mutagenesis. *Current Opinion in Structural Biology*, *50*, 117–125.
- Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C., & Shendure, J. (2010). Parallel, tag-directed assembly of locally derived short sequence reads. *Nature Methods*, *7*, 119–122.
- Invitrogen. (2015). *Evaluation of linear synthetic DNA fragments from separate suppliers*. Thermo Fisher Scientific.
- Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., et al. (2000). Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 1143–1147.
- Jessop-Fabre, M. M., Jakočiūnas, T., Stovicek, V., Dai, Z., Jensen, M. K., Keasling, J. D., et al. (2016). EasyClone-MarkerFree: A vector toolkit for marker-less integration of genes into *Saccharomyces cerevisiae* via CRISPR–Cas9. *Biotechnology Journal*, *11*, 1110–1117.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*, 1647–1649.
- Keren, L., Hausser, J., Lotan-Pompan, M., Vainberg Slutskin, I., Alisar, H., Kaminski, S., et al. (2016). Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell*, *166*, 1282–1294.e18.

- Keren, L., Zackay, O., Lotan-Pompan, M., Barenholz, U., Dekel, E., Sasson, V., et al. (2013). Promoters maintain their relative activity levels under different growth conditions. *Molecular Systems Biology*, 9, 701.
- Lederer, T., Kintrup, M., Takahashi, M., Sum, P. E., Ellestad, G. a., & Hillen, W. (1996). Tetracycline analogs affecting binding to Tn10-Encoded Tet repressor trigger the same mechanism of induction. *Biochemistry*, 35, 7439–7446.
- Lee, M. E., DeLoache, W. C., Cervantes, B., & Dueber, J. E. (2015). A highly characterized yeast toolkit for modular, multipart assembly. *ACS Synthetic Biology*, 4, 975–986.
- Leung, D. W., Chen, E., & Goeddel, D. V. (1989). A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Techniques*, 1, 11–15.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
- Licitra, E. J., & Liu, J. O. (1996). A three-hybrid system for detecting small ligand-protein receptor interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93, 12817–12821.
- Lin-Goerke, J. L., Robbins, D. J., & Burczak, J. D. (1997). PCR-based random mutagenesis using manganese and reduced dNTP concentration. *BioTechniques*, 23, 409–412.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012, 1–11.
- Liu, W., Yuan, J. S., & Stewart, C. N., Jr. (2013). Advanced genetic tools for plant biotechnology. *Nature Reviews Genetics*, 14, 781–793.
- Lutz, S. (2010). Beyond directed evolution—semi-rational protein engineering and design. *Current Opinion in Biotechnology*, 21, 734–743.
- Marion, D. (2013). An introduction to biological NMR spectroscopy. *Molecular & Cellular Proteomics: MCP*, 12, 3006–3025.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20, 1297–1303.
- Medina-Cucurella, A. V., Steiner, P. J., Faber, M. S., Beltrán, J., Borelli, A. N., Kirby, M. B., et al. (2019). User-defined single pot mutagenesis using unamplified oligo pools. *Protein Engineering, Design and Selection*, 32, 41–45.
- Moore, I. F., Hughes, D. W., & Wright, G. D. (2005). Tigecycline is modified by the flavin-dependent monooxygenase TetX. *Biochemistry*, 44, 11829–11835.
- Morrison, K. L., & Weiss, G. A. (2001). Combinatorial alanine-scanning. *Current Opinion in Chemical Biology*, 5, 302–307.
- Muteeb, G., & Sen, R. (2010). Random mutagenesis using a mutator strain. In J. Brame (Ed.), *Methods in molecular biology (Methods and protocols): Vol. 634. In vitro mutagenesis protocols* (pp. 411–419). Totowa, NJ: Humana Press.
- Naseri, G., Balazadeh, S., Machens, F., Kamranfar, I., Messerschmidt, K., & Mueller-Roeber, B. (2017). Plant-derived transcription factors for orthologous regulation of gene expression in the yeast *Saccharomyces cerevisiae*. *ACS Synthetic Biology*, 6, 1742–1756.
- Park, J., Gasparrini, A. J., Reck, M. R., Symister, C. T., Elliott, J. L., Vogel, J. P., et al. (2017). Plasticity, dynamics, and inhibition of emerging tetracycline resistance enzymes. *Nature Chemical Biology*, 13, 730–736.
- Ramos, J. L., Martínez-Bueno, M., Molina-Henares, A. J., Terán, W., Watanabe, K., Zhang, X., et al. (2005). The TetR family of transcriptional repressors. *Microbiology and Molecular Biology Reviews: MMBR*, 69, 326–356.
- Rogers, J. K., Taylor, N. D., & Church, G. M. (2016). Biosensor-based engineering of biosynthetic pathways. *Current Opinion in Biotechnology*, 42, 84–91.
- Rubin, A. F., Gelman, H., Lucas, N., Bajjalieh, S. M., Papenfuss, A. T., Speed, T. P., et al. (2017). A statistical framework for analyzing deep mutational scanning data. *Genome Biology*, 18, 150.

- Schrödinger LLC. (2018). *The PyMOL molecular graphics system*. New York, NY: Schrödinger, LLC.
- Scott, L. H., Mathews, J. C., Flematti, G. R., Filipovska, A., & Rackham, O. (2018). An artificial yeast genetic circuit enables deep mutational scanning of an antimicrobial resistance protein. *ACS Synthetic Biology*, 7, 1907–1917.
- Shi, Y. (2014). A glimpse of structural biology through X-ray crystallography. *Cell*, 159, 995–1014.
- Starita, L. M., Pruneda, J. N., Lo, R. S., Fowler, D. M., Kim, H. J., Hiatt, J. B., et al. (2013). Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 110, E1263–E1272.
- Steiner, K., & Schwab, H. (2012). Recent advances in rational approaches for enzyme engineering. *Computational and Structural Biotechnology Journal*, 2, e201209010.
- Tang, S.-Y., & Cirino, P. C. (2011). Design and application of a mevalonate-responsive regulatory protein. *Angewandte Chemie International Edition*, 50, 1084–1086.
- Tang, S.-Y., Qian, S., Akinterinwa, O., Frei, C. S., Gredell, J. A., & Cirino, P. C. (2013). Screening for enhanced triacetic acid lactone production by recombinant *Escherichia coli* expressing a designed triacetic acid lactone reporter. *Journal of the American Chemical Society*, 135, 10099–10103.
- Van Criekeing, W., & Beyaert, R. (1999). Yeast two-hybrid: State of the art. *Biological Procedures Online*, 2, 1–38.
- van Montfort, R. L. M., & Workman, P. (2017). Structure-based drug design: Aiming for a perfect fit. *Essays in Biochemistry*, 61, 431–437.
- Vanhercke, T., Ampe, C., Tirry, L., & Denolf, P. (2005). Reducing mutational bias in random protein libraries. *Analytical Biochemistry*, 339, 9–14.
- Volkers, G., Damas, J. M., Palm, G. J., Panjikar, S., Soares, C. M., & Hinrichs, W. (2013). Putative dioxygen-binding sites and recognition of tigecycline and minocycline in the tetracycline-degrading monooxygenase TetX. *Acta Crystallographica. Section D, Biological Crystallography*, 69, 1758–1767.
- Volkers, G., Palm, G. J., Weiss, M. S., Wright, G. D., & Hinrichs, W. (2011). Structural basis for a new tetracycline resistance mechanism relying on the TetX monooxygenase. *FEBS Letters*, 585, 1061–1066.
- Wallis, C. P., Filipovska, A., & Rackham, O. (2018). A modified yeast three-hybrid system enabling both positive and negative selections. *Biotechnology Letters*, 40, 1127–1134.
- Wallis, C. P., Richman, T. R., Filipovska, A., & Rackham, O. (2018). Tighter ligand binding can compensate for impaired stability of an RNA-binding protein. *ACS Chemical Biology*, 13, 1499–1505.
- Weber, W., & Fussenegger, M. (2010). Synthetic gene networks in mammalian cells. *Current Opinion in Biotechnology*, 21, 690–696.
- Yan, Y., Liu, Q., Zang, X., Yuan, S., Bat-Erdene, U., Nguyen, C., et al. (2018). Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. *Nature*, 559, 415–418.
- Yang, W., Moore, I. F., Koteva, K. P., Bareich, D. C., Hughes, D. W., & Wright, G. D. (2004). TetX is a flavin-dependent monooxygenase conferring resistance to tetracycline antibiotics. *The Journal of Biological Chemistry*, 279, 52346–52352.