Title
Towards Comprehensive and Programmable Protein Mutagenesis

Author
Higgins, Sean

Publication Date
2018

Peer reviewed|Thesis/dissertation

Towards Comprehensive and Programmable Protein Mutagenesis

By

Sean A Higgins

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Molecular and Cell Biology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor David F. Savage, Chair
Professor Ming C. Hammond
Professor Evan W. Miller
Professor John E. Dueber

Fall 2018

Abstract

Towards Comprehensive and Programmable Protein Mutagenesis

by

Sean A Higgins

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor David F. Savage, Chair

A fundamental goal of protein biochemistry is to determine the sequence-function relationship, but the vastness of sequence space makes comprehensive evaluation of this landscape difficult. Advances in DNA synthesis and sequencing now allow researchers to assess the functional impact of thousands of amino acid substitutions in a single experiment, however, the quality and diversity of these mutations controls the breadth of knowledge gained by these emerging methods. Comprehensive and programmable protein mutagenesis is critical for understanding structure-function relationships and improving protein function. However, current techniques enabling comprehensive protein mutagenesis are based on PCR and require in vitro reactions involving specialized protocols and reagents. This has complicated efforts to rapidly and reliably produce desired comprehensive protein libraries. Here we demonstrate that plasmid recombineering is a simple and robust in vivo method for the generation of protein mutants for both comprehensive library generation as well as programmable targeting of sequence space. Using the fluorescent protein iLOV as a model target, we build a complete mutagenesis library and find it to be specific and comprehensive, detecting 99.8% of our intended mutations. We then develop a thermostability screen and utilize our comprehensive mutation data to rapidly construct a targeted and multiplexed library that identifies significantly improved variants, thus demonstrating rapid protein engineering in a simple protocol.

Beyond simple amino acid substitutions, protein topology is also well-established as a key mechanism by which large, complex multi-domain proteins evolve highly specialized functions. While rationally constructed protein deletions have long been essential to elucidating biochemical properties, current techniques are insufficient for a comprehensive approach. Here we develop a method for constructing fitness landscapes for even the largest and most complex proteins, comprehensively surveying functional deletions in the RNA-guided DNA binding protein dCas9, the foundation for powerful genome editing and modifying technologies. CRISPR proteins are highly complex with numerous distinct domains responsible for activities such as guide RNA binding, DNA recognition, DNA unwinding, specificity sensing and ultimately the cleavage of each DNA strand. We exploit the fitness landscape to revert functionality and step backward in domain evolution, comprehensively minimizing dCas9 and screening for an essential function. We demonstrate the power of this technique by revealing the minimal RNA guided DNA binding module at 64% of the full CRISPR-Cas9 platform, providing many new opportunities for fusions and delivery. This exploration also uncovers evidence for a DNA unwinding mechanism in a domain heretofore viewed as dispensable in Cas9. These results

highlight the power of comprehensive protein deletions to clearly elucidate the boundaries of a central function.

Together, amino acid substitution and topological mutation (encompassing deletions, insertions, and circular permutations) comprise all possible genetic protein modifications. This work has served to develop simple and robust methods, which remain programmable and comprehensive, for both substitution and topological mutagenesis. The construction of high-quality protein libraries is a foundational step for applications in the fundamentals of protein biochemistry, disease prediction, and protein engineering. Ultimately, understanding the general principles of protein sequence-function landscapes - enabled by massively parallel experimentation - will allow computational methods to synergize with programmable mutagenesis and vastly improve the search for novel fitness variants.

Acknowledgements

Graduate school has been a formative period of my life, and I am grateful for the support and mentorship of many people.

I am indebted to Professor David Savage for the opportunities and experiences I have had in pursuing this work. His mentorship and leadership combine to create a productive and strongly cohesive team. I remember discussing MAGE during my MCB interview, and I would go on to begin exploring recombineering pilot experiments during my first rotation a few months later.

I wish to thank all the members of the Savage lab for contributing to the team, both professionally and socially. It has been a pleasure to take part as the lab continues to grow and develop. Many of you directly helped me in discussions of experimental design, taught me to use dozens of instruments, and aided me in data analysis approaches. A non-exhaustive list: Ben Oakes, Avi Flamholz, Dana Nadler, Rayka Yokoo, Stacy-anne Morgan, Rachel Hood, Caleb Cassidy-Amstutz, Luke Oltrogge, Arik Shams, Rob Nichols, and Emeric Charles. Many other mentors throughout UC Berkeley also directly supported my work, including Rob Egbert, Zachary Hallberg, and Christof Fellmann.

A special thanks to the UC Berkeley MCB Class of 2013 and the friendships we have built. Your optimism, creativity, and dedication create the feedback loops that power scientific advancement.

I could not have completed this work without Katie.

I dedicate this work to my father, Steve Higgins.

Table of Contents

List of Figures

List of Tables:

Chapter 1

Introduction

† The work presented in this chapter has previously been published in the following review article: Higgins, S.A., and Savage, D.F. (2018). Protein Science by DNA Sequencing: How Advances in Molecular Biology Are Accelerating Biochemistry. Biochemistry. 57 (1), 38-46.

Abstract

A fundamental goal of protein biochemistry is to determine the sequence-function relationship, but the vastness of sequence space makes comprehensive evaluation of this landscape difficult. However, advances in DNA synthesis and sequencing now allow researchers to assess the functional impact of every single mutation in many proteins, but challenges remain in library construction and the development of general assays applicable to a diverse range of protein functions. This perspective briefly outlines the technical innovations in DNA manipulation which allow massively parallel protein biochemistry, then summarizes the methods currently available for library construction and the functional assays of protein variants. Areas in need of future innovation are highlighted with a particular focus on library construction, including *in vivo* mutagenesis methods and topological protein modifications. Assay development and the use of computational analysis is also discussed in effectively traversing the sequence-function landscape. Finally, applications in the fundamentals of protein biochemistry, disease prediction, and protein engineering are presented.

1.1 Protein science by DNA sequencing: how advances in molecular biology are accelerating biochemistry

Protein function is encoded in the sequence of amino acids making up the polypeptide chain, and decades of biochemical studies have sought to define the principles underlying the sequence-function relationship. Altering even a single amino acid, though, requires choosing from thousands of theoretical experiments due to the vast, combinatorial nature of sequence space, forcing the researcher to predict which mutations will be most informative (Mandecki 1998). Mutational experiments are traditionally informed by biochemical principles and empirical data yet nonetheless often fail. Recent advances in molecular biology, however, enable a comprehensive investigation of functional determinants by allowing every possible single mutation to be made and evaluated (Fowler et al. 2010).

Emerging approaches leverage advances in DNA sequencing and molecular biology to create massive libraries (e.g. >106) of mutant proteins, map the functional importance of each residue and, in some cases, to engineer enhanced function into proteins (Fowler and Fields 2014). Such experiments generally consist of building a mutant library, challenging protein variants with some form of functional assay that preserves a genotype-phenotype link (either in vivo or in vitro), and leveraging the bandwidth of next-generation DNA sequencing (NGS) to quantitatively evaluate the function of each variant (Figure 1.1). Thus, questions in protein biochemistry are increasingly posed in the form of a DNA sequencing experiment.

The increased scope of massively parallel biochemistry experiments has led to a number of issues that must be resolved. The first experimental phase, library building, must take into account the available methods for constructing genetic diversity (Wrenbeck, Faber, and Whitehead 2017; Zheng, Xing, and Zhang 2017). As discussed below, molecular biology has progressed to the point where substitution libraries (i.e. where amino acids may be mutated, or substituted, with other amino acids) may be produced with high specificity and at moderate cost, but more complex libraries require specialized protocols. Once constructed, the library must be assayed. Conceptually, assays are designed in such a way that a protein variant's function impacts its abundance within the population of other library members, which can be quantitatively defined as a term known as fitness (Araya and Fowler 2011). Linking function and fitness is recalcitrant to general solutions due to the diversity of protein functions (Wrenbeck, Faber, and Whitehead 2017), and assay development remains the lowest bandwidth stage in experimental design. Here, bandwidth refers to the maximum possible number of variants that can pass through an assay in a reasonable fashion. Assays can be categorized as either a selection or a screen. A selection refers to assays in which high fitness variants are enriched autonomously in cells through growth dependence, while screens rely on measurement and physical separation to enrich desired variants. Finally, the change in abundance, necessarily accomplished by competition between the variants, can be measured in the third phase: DNA sequencing. NGS allows the fitness evaluation of millions of protein variants simultaneously, under the assumption that the most enriched sequences correlate with the most functional variants (Fowler et al. 2011). The factors that determine the most functional variants will vary based on the assay, but generally include characteristics such as stability, catalytic activity, substrate affinity and specificity, and affinity.

This chapter will first outline the techniques currently available for high-throughput protein science, including recent advances in manipulating DNA and various options for massively parallel functional interrogation of proteins. The engaged reader is also referred to

recent reviews on systematic protein mutagenesis (Wrenbeck, Faber, and Whitehead 2017; Zheng, Xing, and Zhang 2017). The second section describes ways in which large experimental datasets can provide insight into biological questions about the fundamental properties of proteins (Fowler and Fields 2014), the consequences of human genetic variation (Shendure and Fields 2016), and the engineering of novel proteins for specific functions (Wrenbeck, Faber, and Whitehead 2017).

1.2 Technical aspects of high throughput biochemistry

Proteins evolve via two different general mechanisms: mutagenic substitutions to individual amino acids and larger, topological changes to gene structure (Figure 1.2). From an experimental viewpoint, substitution mutagenesis is the better explored of the two and ranges from site directed mutagenesis (Hutchison et al. 1978) to recent innovations including deep mutational scanning experiments, where e.g. all possible single mutations across a gene can be evaluated (Araya and Fowler 2011; Fowler and Fields 2014). In contrast, topological changes such as insertions, deletions, gene duplications, and circular-permutations have received only limited experimental attention despite evidence that these mutations are central to the evolution of protein complexity (Bhattacharyya et al. 2006). For example, one conclusion of the human genome project was that the human proteome diversity arises from recombination of conserved domains into new protein architectures (Lander et al. 2001), similar to predictions of the Exon Theory of Genes (Gilbert 1987). Subsequent work found that many multi-domain proteins are the result of gene insertions (Aroul-Selvam, Hubbard, and Sasidharan 2004), and early efforts to create synthetic allosteric protein switches found success by utilizing domain insertion (Guntas and Ostermeier 2004; Guntas et al. 2005). Relatedly, it is now believed that the complexity and evolvability of eukaryotic cell signaling circuits results from the underlying modularity of protein domains (Bhattacharyya et al. 2006).

Two breakthroughs have allowed the development of a comprehensive mutagenesis approach to protein biochemistry. These innovations are the most recent in an exponentially increasing capacity to read and write DNA (Carr and Church 2009; Esvelt and Wang 2013). First, advances in DNA synthesis now allow the production of > 50,000 unique 200-mer oligonucleotides and have been used to synthesize dozens of genes, totaling tens of kilo-basepairs (bp) of DNA, at a cost of < $0.01/bp of final synthesized sequence (Kosuri et al. 2011). The technology behind this throughput, Oligo Library Synthesis (OLS), is enabled by advances in massively parallel solid surface phosporamidite chemistry (LeProust et al. 2010), where side reactions are the limiting factor in synthesis fidelity and ultimately constrain oligonucleotide length. Future advances in synthesis chemistry will likely improve these economics. Second, the development of NGS enables routine sequencing of > 1 billion reads of 150 bp, at a cost of $7 per million bp (Goodwin, Mcpherson, and Mccombie 2016). Key innovations of NGS include amplification of library fragments to allow sufficient signal for base-calling and the use of solid surfaces to enable massively parallel sequencing (Mardis 2013). The scale of NGS technologies is steadily advancing, with upcoming commercial systems capable of 10 billion high quality reads per run. One important technical limitation to overcome is that of sequencing read length. Epistasis cannot be detected, for example, if the linkage of two mutations lies further apart than the read length, which is typically 300 bp. Overcoming this issue requires either increasing the read length or generating a lookup table of variant sequences, and both these options are currently being pursued (Wrenbeck, Faber, and Whitehead 2017), but challenges remain.

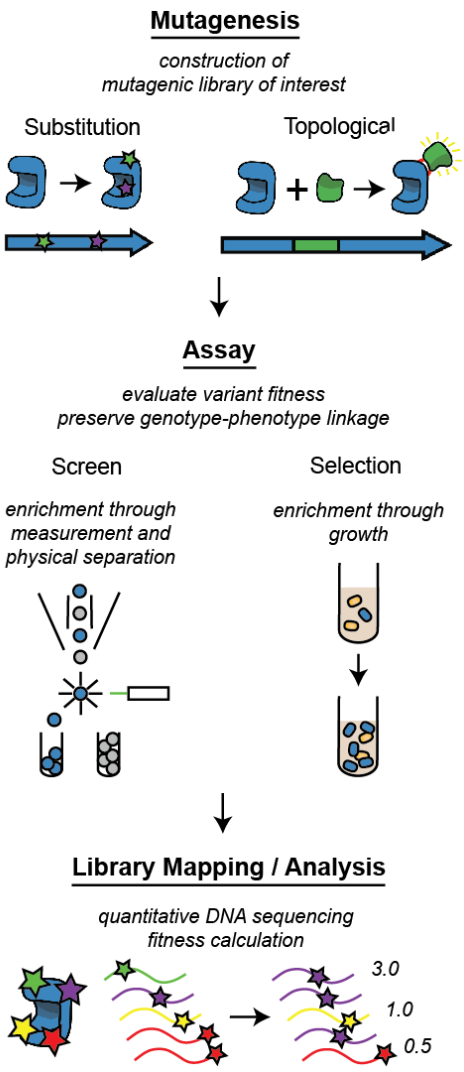


Figure 1.1: Massively parallel investigations of protein function consist of three phases. First, libraries of protein variants must be constructed. These variants can contain mutations which consist of either amino acid substitutions or larger topological changes, such as insertions (shown), deletions, or circular permutations. Second, protein variants are evaluated for a specific function in an assay step. Broadly, an assay can consist of measurement and physical separation of desired variants in a screen, or growth of an organism harboring the protein variant where replication fitness is dependent on the target protein's function. Finally, both the naïve library and libraries which have passed through an assay are sequenced by NGS to identify and enumerate variants contained within each. An enrichment ratio can then be calculated for each variant in order to quantitatively assign a fitness effect to those particular mutations.

Systems enabling longer read length such as PacBio (Rhoads and Au 2015) and Nanopore (Loose 2017), for example, currently suffer from lower bandwidth and accuracy in comparison to NGS.

1.3 Substitution mutagenesis

There are a now a number of methods that allow programmed creation of protein libraries (Figure 1.2) (Wrenbeck, Faber, and Whitehead 2017; Zheng, Xing, and Zhang 2017). In general, libraries are built on an episomal replicon, such as a plasmid, so as to easily amplify and manipulate DNA, shuttle between hosts, and verify phenotype-genotype linkage through retransformation. Recently, several techniques including PALS (Kitzman et al. 2015), PFunkel (Firnberg and Ostermeier 2012), and Nicking Mutagenesis (Wrenbeck et al. 2016), have been developed to produce comprehensive substitution mutagenesis libraries in vitro by leveraging short oligonucleotide chemical synthesis coupled with downstream molecular biology. The engaged reader is referred to recent reviews on the topic (Wrenbeck, Faber, and Whitehead 2017; Zheng, Xing, and Zhang 2017). Additionally, recombineering (Copeland, Jenkins, and Court 2001) in vivo presents an alternative approach for comprehensive substitution mutagenesis (Higgins, Ouonkap, and Savage 2017). Ultimately, it is likely that chemical synthesis methods, which offer control over DNA sequence composition, will be the optimal future choice for library generation. In this vein, DNA synthesis-based libraries have already begun to be applied to peptides, such as in a recent paper from the Baker group describing the computationally-guided enumeration and assay of thousands of miniproteins (folded polypeptides less than 50 amino acids) to investigate the sequence determinants of folding and stability(Rocklin et al. 2017). The computational design of minimal proteins is sure to be a rich future area of biochemistry and is highlighted below.

An alternative emerging strategy for studying and improving protein is by the continuous introduction of mutations by site-specific mutators in vivo (Packer and Liu 2015). Although the methods above offer significant control over library composition, they are limited by ex vivo library creation and the need to transform the library into the desired cell type (i.e. limited transformation yield). Additionally, though it may change in the future, synthesis errors and oligo length remain a limitation for library construction. Mutagenesis methods that function entirely in vivo theoretically offer the largest library sizes, the least labor-intensive protocols, and greater accessibility in terms of cost and specialized reagents. However, targeted mutagenesis in vivo is considerably more difficult, as the origin of sequence diversity must be localized to a specific target area in the genome. Various existing approaches leverage an orthogonal error-prone polymerase (Fabret et al. 2000), a targeting glycosylase (Finney-Manchester and Maheshri 2013), an error-prone Ty1 reverse transcriptase (Crook et al. 2016), phage-assisted continuous evolution (PACE) (Esvelt, Carlson, and Liu 2011), or a cytidine deaminase-Cas9 fusion (Komor et al. 2016) to accomplish mutagenesis at a particular target gene. Each of these methods suffers from significant drawbacks compared to in vitro protocols as detailed in a recent review (Zheng, Xing, and Zhang 2017). In general these approaches do not cover the full codon mutational space (Kitzman et al. 2015), are not programmable, and require significant customization.

1.4 Topological mutagenesis

While substitution mutations can modify the existing function of a protein, topological changes – i.e. changes in connectivity - can introduce entirely new functions by taking large steps in sequence and functional space as is seen in the evolution of multi-domain proteins. For example, circular permutations are topological changes where a protein is split and the N-

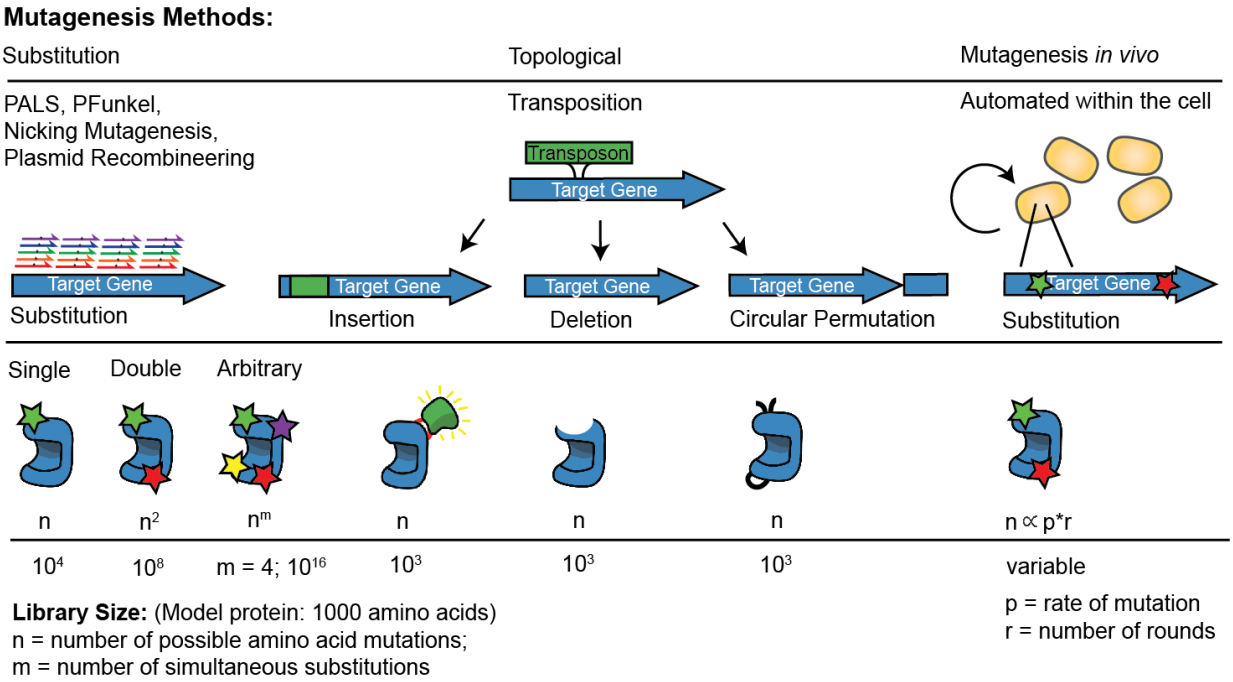


Figure 1.2: A variety of mutagenesis methods exist for constructing large libraries of protein variants, and can be categorized as either substitution, topological, or in vivo mutagenesis. Substitution methods based on synthetic oligonucleotides enable programmable control over library composition and bias, and have been used to build comprehensive substitution libraries. Note that random mutagenesis methods cannot cover the entirety of amino acid space due to the unlikelihood of triplet codon mutation. The combinatorial nature of sequence space results in an nm library size for substitution libraries, where n is the total number of possible mutations (i.e. length of protein x number of amino acids) and m is the number of intended mutations. Practically speaking, comprehensive mutagenesis libraries of m > 2 are exceedingly difficult to achieve. In contrast, topological mutations including insertions, deletions, and circular permutations result in library sizes equal to the length of a protein, and are generated using transposon-based methods. Finally, in vivo mutagenesis methods rely on random mutations generated in a variety of ways. These library sizes will vary based on the rate of mutagenesis (p) and the number of rounds of mutagenesis (r).

terminal portion is fused to the C-terminus via a linker, leaving a new N- and C-terminus elsewhere in the protein. Circular permutations have been exploited in a variety of studies to generate altered binding affinity (Cheltsov, Barber, and Ferreira 2001), improved catalytic activity (Qian and Lutz 2005), and new biosensors (Okada, Ota, and Ito 2009) and enzymes (Reitinger et al. 2010). However, topological mutations are more complicated to create than substitutions. For example, topological mutations will often require a so called linker peptide to reconnect protein domains. Early studies of TEM β-lactamase circular permutation found that varying both the length and amino acid composition of the connecting peptide drastically affected protein function (Osuna, Pérez-Blancas, and Soberón 2002), and recent work constructing biosensors using domain insertion has continued to show that linker screening can improve protein function (Nadler et al. 2016). Moreover, insertions, deletions, and circular permutation all require specialized molecular biology protocols. An early approach to comprehensive topological mutations was based on random DNaseI cleavage of a plasmid

containing the target gene, but the use of a nonspecific endonuclease creates numerous deleterious truncation and duplication events (Graf and Schachman 1996; Guntas and Ostermeier 2004).

More recently, transposons have emerged as a reliable tool for introducing topological protein changes (Shah and Kim 2016). For example, the random integration of the Mu transposon into a target gene, which can be catalyzed in vitro with purified transposase, has been used in downstream molecular biology applications to generate both comprehensive protein domain insertion (Edwards et al. 2010) and circular permutation (Mehta, Liu, and Silberg 2012) libraries. Additional work has engineered the transposon itself to minimize any vestigial transposon 'scar' sequence (Nadler et al. 2016; A. M. Jones et al. 2016) that could affect protein function (Pierre et al. 2015; Shah and Kim 2016; Nadler et al. 2016). In the case of domain insertion, these approaches have been used to systematically map the potential for fusion protein engineering in several proteins, such as the RNA-guided endonuclease Cas9, and to identify variants of beta-lactamase (Edwards et al. 2010), Cas9 and the green fluorescent protein (GFP) that are allosterically regulated by small molecules (Nadler et al. 2016; Oakes et al. 2016). Transposons have also been used to systematically study the effect of deletions on protein function (Arpino et al. 2014; Morelli et al. 2017). Specifically, a method employing Type IIS restriction enzymes for trinucleotide deletion was developed (D. D. Jones 2005) and used to investigate registry shifts in GFP, resulting in the discovery of a variant with higher fluorescence in cells (Arpino et al. 2014). Another recent study used transposons to generate extensive truncation variants of the artificial RNA ligase enzyme 10C and used mRNA display as an in vitro selection to identify functional variants that were nearly 20% shorter (Morelli et al. 2017).

One important consideration is that transposon-based methods, which produce randomized libraries, are not programmable in the manner that oligonucleotide-based methods are (Shah and Kim 2016), i.e., subsets of positions cannot be selectively targeted. Additionally, in the case of insertions and deletions, 2/3rds of a transposon library will contain a frameshifted gene, and only one of two insertion orientations will encode the expected protein sequence. Thus, only 1/6th of a transposon-created insertion libraries are interesting variants. These drawbacks further limit the library sizes that can be generated, but more importantly they fundamentally exclude any type of rational exploration of topological sequence space. In the future, it is likely that new mutagenesis approaches will allow specific libraries to be constructed which will be as tunable as substitution methods (Tullman et al. 2016). In particular, unique restriction sites could be programmably introduced into target genes, facilitating downstream molecular biology for generating topological mutations.

1.5 Assay Design and Data Interpretation

Good assays link fitness to function and result in an increased or decreased abundance of each variant in accordance with its performance. Abundance can later be quantified by NGS. Unfortunately, there are no generic approaches for assaying protein function, and the assay choice depends greatly on the protein (Figure 1.3). Assay development is therefore a core technical limitation that requires customization and tuning, reduces effective library size (by limiting the number of variants that can be tested), and restricts the total diversity of sequence space that can be explored in one experiment.

Selections offer excellent throughput (106 – 1013), low cost, and minimal labor requirements but require that a link be formed between function and growth (i.e. connecting protein fitness and cell fitness) (Packer and Liu 2015). This is typically non-trivial to achieve except for the simplistic cases of positive selection, such as resistance of a target protein to its cognate drug (Firnberg et al. 2014) or the enrichment of high affinity binders. In the case of binding, the physical link between a protein and its encoding gene can be accomplished through yeast display (Cherf and Cochran 2015), phage display(Wu et al. 2016), or ribosome display (Hanes and Plückthun 1997). For proteins with other functions, recent results suggest that a systems-level understanding of physiology may open routes to novel selections. For example, recent work showed that certain enzymes can be functionally linked to replication through dependence on a sole nitrogen source (Wrenbeck, Azouz, and Whitehead 2017). Here, the aliphatic amide hydrolase AmiE catalyses hydrolysis of the amide to its corresponding carboxylic acid. This reaction produces ammonium as a bioavailable nitrogen source, and results in a robust selection assay after tuning enzyme expression. In general, enzymes can be amenable to selection schemes if their function impacts metabolism, and a recently developed computational tool, SelFi, seeks to streamline the identification of assays for a particular target enzyme (Hassanpour et al. 2017). At its core, SelFi seeks to identify synthesis pathways that lead from a desired enzymatic product to a metabolite suitable for selection within an organism. Selections are subject to cheating, escape, and contamination, which can complicate experiments. An alternative to the above is to avoid the evolution of unintended solutions by creating minimal replication systems. One such system, termed compartmentalized partnered replication (CPR), is an ex vivo system capable of selecting for any genetic element that can alter the expression or function of Taq DNA polymerase (Ellefson et al. 2014). CPR sequesters a library of genetic parts into emulsion droplets that are then differentially amplified by a PCR step. The production of polymerase, however, is dependent on the function of the partner gene. Because the amplification step is dependent on a minimal number of components, the complexity of selection is vastly reduced and thereby eliminates pathways for escape mutants to arise. CPR, though, is thus inherently limited to proteins that alter the polymerase expression.

For a protein whose function cannot be straightforwardly tied to replicative fitness, the experimenter must instead use a screen to individually measure and physically bin each protein variant. This is increasingly practical with the improvement of and broader accessibility to high-throughput technologies such as fluorescence-activated cell sorting (FACS). As in a selection, a protein's function can be linked to a secondary function which is amenable to these high throughput screening technologies. For example, a general method has been developed to link glycosyltransferases activity to cellular fluorescence, enabling throughputs of > 107 variants (Yang et al. 2010). Here, fluorescently labeled sugars such as lactose are transported into the cell by a transport protein. Enzymatically modified sugars lose transport competence, remaining trapped inside the cell, while unreacted dye is lost during a wash. FACS sorting of brighter cells

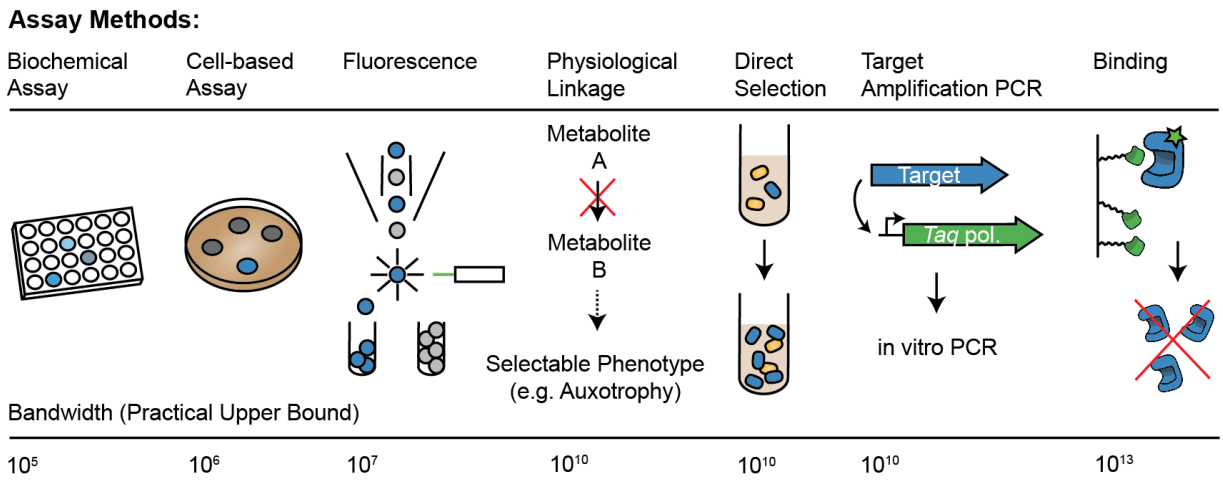


Figure 1.3: Due to the wide diversity of protein function, universal methods to evaluate libraries of proteins are difficult to develop. However, existing approaches may be categorized by the specific method used to link protein function to protein fitness. The lowest bandwidth approach, biochemical assay, relies on individual sequential measurement and spatial separation of variants at the macro scale. Next, cell-based assays can be used to screen individual colonies on plates, e.g. calorimetric enzyme activity. Fluorescence-based approaches can make use of FACS to screen even larger numbers of cells. Notably, these assays are not limited to only fluorescent proteins, and systems can be designed such that fluorescence is dependent on function of the protein target. Similarly, physiological linkage of cellular growth to a reliance on target protein function enables even higher bandwidth. Additionally, in vitro PCR amplification of a target protein can be linked to target protein function, such that the most functional target protein variants become the most amplified in the population. The highest throughput cellular assays rely on direct selection, where the target protein natively impacts cell growth. Finally, binding assays represent the maximum throughput available for protein screening, as each individual variant is physically linked to its genetic information.

thus enriches for functionally active enzymes. Two different colors of fluorescent markers can be used simultaneously to avoid the selection of dye binding sites, a previously observed escape mechanism. Alternatively, the function of metabolic pathways can be assayed using a class of protein tools known as metabolite biosensors (Morgan et al. 2016). Biosensors link the abundance of small molecules to a measurable phenotype, such as fluorescence intensity, and are increasingly being used to assay enzymes and engineer improved function into metabolic pathways (Rogers and Church 2016).

Due to the diversity of protein functions it is difficult to develop general assays suitable for many possible targets. Nevertheless, analytical chemistry and NGS provide two plausible approaches, provided that genotype-phenotype linkage is maintained and that throughput is high enough to cover desired library sizes. For example, the Agilent RapidFire mass spectrometry system is capable of sample measurements in 15 s for small molecules and 10 s for peptides (Vanderporten et al. 2013). This technology has been used to measure glycolysis reactants in order to quantify the effectiveness of inhibitory molecules as anticancer drugs, achieving a rate enabling 10,000 samples per day (Rye and LaMarr 2015). Plate-based assays with sophisticated robotics can reduce measurement time to ~ 1 s, while water-in-oil emulsions have achieved > 2000 individual assays per second (Agresti et al. 2010). Alternatively, NGS enables throughput that is more than sufficient for generalizable assays, and in principle can quantify function for

any protein that directly controls transcription or could be engineered to. Such an approach requires linking individual transcripts to individual proteins, which could be accomplished through transcript barcoding.

In order to determine the fitness of individual protein variants, the abundance of each variant is measured by NGS. A number of statistical packages have been released to interpret the deep sequencing datasets produced by these assays (Wrenbeck, Faber, and Whitehead 2017). Briefly, a typical approach is to identify variant sequences and enumerate them in pre- and post-assay libraries, creating an enrichment ratio for each variant between selected and unselected libraries. This ratio forms the basis for understanding the sign and strength of a variant's functional change. Many library generation techniques, such as those using transposons, display some biases, and the enrichment ratio is critical to separating variants with enhanced function from those that are simply more abundant in the library by chance (Nadler et al. 2016). Because an enrichment ratio represents the slope of abundance change over time, multiple time points or rounds of an assay can be fit by a linear regression, in which case the slope of the line corresponds to functional activity. Enrich2 improves upon previous approaches by 1) normalizing for wild-type in each time point and 2) weighting regression time points based on variant counts (Fowler et al. 2011). These corrections help to control for non-linear behavior and reduce sampling error.

A recent statistical guide to deep mutational scanning experiments suggests that increasing the number of time points and experimental duration are strong approaches to increasing precision (Matuszewski et al. 2016). These suggestions are, however, based in the consideration of initial naïve libraries in which all studied mutants are well-represented and that population and sample size are large compared with the number of mutants and sequencing depth. Because assay bandwidths are limited by the specific approach (Figure 1.3), oversampling reduces the number of unique variants that can be examined (Persikov et al. 2014). The implication is that library quality is now of paramount importance. For well-studied proteins, it can be preferable to use programmable techniques so that the library composition and size can be designed precisely to match the assay bandwidth. For less well-understood proteins, it can be preferable to use mutagenesis technique that are as unbiased as possible so as to minimize the assay throughput spent re-evaluating the same variant (Higgins, Ouonkap, and Savage 2017).

1.5 Objective of this study

This dissertation is framed as two main parts, where the overall purpose is to improve the methods available for performing programmable and comprehensive protein mutagenesis. The first part, Chapter 2, focuses on the first type of general protein mutation: amino acid substitution. The second part, Chapter 3, focuses on the second type of general protein mutation: topological mutations. As discussed above, topological mutations have received much less attention in the literature, despite initial success in both expanding biochemical knowledge and engineering proteins for enhanced function. In particular, Chapter 3 deals with a type of mutation never before performed comprehensively on a large, complex, multi-domain protein: deletions. The novel method therein devised is also capable of the other two possible topological mutations, namely, protein insertions and circular permutations. The methods developed in this work serve as a foundation for programmably and comprehensively modifying a protein in all possible ways.

Chapter 2

Rapid and Programmable Protein Mutagenesis Using Plasmid Recombineering

† The work presented in this chapter has previously been published in the following article: Higgins, S.A., Ouonkap, S.V.Y., and Savage, D.F. (2017). Rapid and Programmable Protein Mutagenesis Using Plasmid Recombineering. ACS Synthetic Biology. 6 (10), 1825-1833

Abstract

Comprehensive and programmable protein mutagenesis is critical for understanding structure-function relationships and improving protein function. However, current techniques enabling comprehensive protein mutagenesis are based on PCR and require in vitro reactions involving specialized protocols and reagents. This has complicated efforts to rapidly and reliably produce desired comprehensive protein libraries. Here we demonstrate that plasmid recombineering is a simple and robust in vivo method for the generation of protein mutants for both comprehensive library generation as well as programmable targeting of sequence space. Using the fluorescent protein iLOV as a model target, we build a complete mutagenesis library and find it to be specific and comprehensive, detecting 99.8% of our intended mutations. We then develop a thermostability screen and utilize our comprehensive mutation data to rapidly construct a targeted and multiplexed library that identifies significantly improved variants, thus demonstrating rapid protein engineering in a simple protocol.

2.1 Introduction

Directed mutagenesis of a desired protein is an important technique both for understanding structure-function relationships as well as improving protein function for research, biotechnology, and medical applications. For example, techniques like deep mutational scanning, where every position in a protein is mutated to all possible amino acids, can be applied to understand key variants associated with disease (Majithia et al. 2016), while targeted mutagenesis of proteins such as Green Fluorescent Protein (GFP) have expanded our capacity to visualize many biological processes (Heim, Cubitt, and Tsien 1995). The ability to generate comprehensive mutation libraries and programmed libraries focused on specific locations or amino acids is crucial to these applications.

In order to address these needs, many in vitro based approaches have been developed. Firnberg and Ostermeier have built libraries composed almost entirely of single mutations using specialized protocols based on uracil-containing template DNA (Firnberg and Ostermeier 2012), while Melnikov and Mikkelsen constructed a comprehensive library by splitting one gene into many different regions small enough to be synthesized on a programmable microarray, followed by multiplexed in vitro recombination (Melnikov et al. 2014). Belsare and Lewis have demonstrated targeted, combinatorial library construction using alternating cycles of fragment and joining PCR (Belsare et al. 2016). Recently, Wrenbeck and Whitehead introduced nicking mutagenesis, using specialized nucleases to selectively degrade the wild type (WT) template DNA (Wrenbeck et al. 2016).

An alternative approach would be to incorporate synthetic oligonucleotides in vivo directly into a gene of interest in a programmable fashion. In E. coli, oligonucleotides introduced into the cell via electroporation can recombine with the genome or resident plasmids with the help of the lambda phage protein Beta, in a process termed recombineering (Copeland, Jenkins, and Court 2001). Mechanistically, it is thought that Beta-bound oligonucleotides anneal to the replication fork of replicating deoxyribonucleic acid (DNA) and are subsequently incorporated into the daughter strand, thus directly encoding mutations into a new DNA molecule (Mosberg, Lajoie, and Church 2010). Recombineering is therefore a compelling method for genetic manipulation. Cheap and easily obtained standard oligonucleotides are the only varying input and the protocol - mixing oligonucleotides in pooled reactions - is straightforward.

This process was shown to be capable of mutating the E. coli genome for rapid metabolic engineering in a process termed Multiplexed Automated Genome Engineering (MAGE), which used multiple rounds of recombineering to increase the penetrance of mutations (Wang et al. 2009). Other work has demonstrated that thousands of pooled, barcoded oligonucleotides can be used, in parallel, to modify the expression of > 95% of E. coli genes and map their effect on fitness (Warner et al. 2010). More recent studies have combined recombineering with the programmable DNA nuclease Cas9, as a means of enforcing mutational penetrance, to mutate tens of thousands of loci in parallel with high efficiency (Garst et al. 2016).

Despite its success in genome engineering, recombineering of plasmids is relatively uncharacterized (Thomason et al. 2007). Plasmid recombineering (PR) is of particular interest in protein mutagenesis as plasmids are easily shuttled between different strains and organisms for cloning and screening. Notably, recombineering strains achieve enhanced mutation efficiency by knocking out mismatch repair and possess a higher genome-wide mutation rate (Turrientes et al. 2013), which can complicate screens or selections sensitive to suppressor mutations. The use of plasmids, however, uncouples protein variation from any background mutation in the genome.

Importantly, Nyerges et al. found that recombineering itself produces few if any off-target mutations (Nyerges et al. 2016). Thomason et al. have previously demonstrated that PR is capable of generating mutations, insertions, and deletions with efficiencies comparable to genomic recombineering. We reasoned that the principles of MAGE – multiplexed reactions and multiple mutation rounds – would be applicable to PR as well.

To benchmark comprehensive PR for protein engineering we sought to measure the efficiency, bias, and overall performance of saturation mutagenesis on the small protein iLOV. iLOV is a 110 residue protein derived from the Light, Oxygen, Voltage (LOV) domains of the A. thaliana phototropin 2 protein (Chapman et al. 2008). The native LOV domain binds flavin mononucleotide (FMN) and uses this co-factor as a photosensor to direct downstream signal transduction. Mutational analysis has revealed that a cysteine to alanine substitution in the FMN binding site interrupts the native photocycle and instead dramatically increases the protein's fluorescent properties. iLOV is an ideal candidate for further engineering because fluorescent proteins that don't require molecular oxygen for chromophore maturation are a desirable alternative to green fluorescent protein. In previous experiments, DNA shuffling was used to isolate iLOV, a variant that has six amino acid mutations relative to the wild-type phototropin 2 LOV2 sequence and an improved fluorescence quantum yield of 0.44 (Chapman et al. 2008). Additional approaches to engineer further improved iLOV variants have also relied on error-prone PCR and DNA shuffling, missing much of the possible sequence-space (Christie et al. 2012). Due to the potential utility of iLOV and its comparatively limited engineering relative to other fluorescent proteins, we hypothesized iLOV could serve as an excellent model system for exploring the utility of PR. Finally, the gene length of iLOV is exceptionally short (330 bp) and analysis of iLOV libraries is suited to deep sequencing. Current paired-end sequencing covers the entirety of the open reading frame and can accurately identify all mutations to a single sequenced plasmid. This provides insight into the mechanisms and utility of recombineering.

Here we demonstrate that PR is capable of constructing both comprehensive protein libraries and targeted mutagenesis libraries focusing on a small section of sequence space. We built a complete mutagenesis library of iLOV and found it to be specific and unbiased, detecting 99.8 % of our intended mutations. We explored this fitness landscape in the context of thermostability using a plated-based screen that allowed us to identify many desirable thermostabilizing mutations. To demonstrate the iterative and programmable nature of our platform, we designed and built a multiplexed library focused on these mutational hotspots and isolated significantly more stable variants. In total, this work demonstrates that plasmid recombineering is a rapid and robust method for the generation of protein mutants for both unbiased, comprehensive libraries and programmable targeting of specific regions in sequence space.

2.2 Materials and Methods

2.2a Strains and Media

Strain EcNR2 (Addgene ID: 26931)(H. H. Wang et al. 2009) was used for generating PR libraries in plasmid pSAH031 (Addgene ID: 90330). For thermostability screening and protein expression, iLOV libraries were cloned into pTKEI-Dest (Addgene ID: 79784)(Nadler et al. 2016) using Golden Gate cloning(Engler, Kandzia, and Marillonnet 2008) with restriction enzyme BsmbI (NEB) and transformed into either Tuner (Novagen) or XJ b Autolysis E. coli

(Zymo Research). Unless otherwise stated, strains were grown in standard LB (Teknova) supplemented with kanamycin (Fisher) at 60 μg/mL.

2.2b Recombineering Library Construction

Libraries were constructed using a modified protocol from Wang 2011(H. Wang and Church 2011). Briefly, 110 oligonucleotides (Table 2.1) or 25 thermostabilizing oligonucleotides (Table 2.2) were mixed and diluted in water. A final volume of 50 μL of 2 μM oligonucleotides, plus 10 ng of pSAH031, was electroporated into 1 mL of induced and washed EcNR2 using a 1 mm electroporation cuvette (BioRad GenePulser). A Harvard Apparatus ECM 630 Electroporation System was used with settings 1800 kV, 200 Ω, 25 μF. Three replicate electroporations were performed, then individually allowed to recover at 30° C for 2 hr in 1 mL of SOC (Teknova) without antibiotic. LB and kanamycin was then added to 6 mL final volume and grown overnight. Cultures were miniprepped (QIAprep Spin Miniprep Kit) and monomer plasmids were isolated by agarose gel electrophoresis and gel extraction (QIAquick Gel Extraction Kit) to remove multimer plasmids (Thomason et al. 2007). The three replicates were then combined, completing a round of PR.

2.2c Library Sequencing and Analysis

The iLOV open reading frame was amplified from PR libraries by PCR to add indices and priming sequences for deep sequencing (Table 2.3). PCR products were sequenced on Illumina platform sequencers (MiSeq and HiSeq) through the Berkeley Genomics Sequencing Laboratory. Sequencing data were analyzed with a custom MATLAB pipeline. Briefly, reads were filtered to remove those that were of low quality, frameshifted, or did not exactly match the annealing portion of the amplifying primers. A detailed description of sequencing analysis can be found below. Finally, full-length reads were compared with the iLOV target sequence for mutation analysis.

Sequencing analysis was performed with custom MATLAB scripts available online at https://github.com/savagelab. 250 nucleotide paired end reads were used to sequence the iLOV gene (330 bp). Sequencing of the forward reads began with a five nucleotide variable region to facilitate cluster identification. Positions 6 to 27 of the forward read are composed of the following sequence, corresponding to the annealing region used for primer amplification of the libraries: ‘TCATTAATGCACGTCTCTGTCC’. Likewise, positions 1 to 18 of the reverse reads correspond to the reverse primer ‘ATGGTGATGGTGACCGCT’. The first 187 nucleotides from each read were used to construct the sequence of the iLOV gene for each molecule. Regions of overlap between forward and reverse reads were used to estimate the sequencing error rate for the final quality passed reads (see below). Reads were filtered as follows: required to match the forward read primer region and required a quality score > 15 (Sanger / Illumina 1.9 encoding) for every nucleotide up to position 187, inclusive. This resulted in the following distribution of reads: quality passed reads/total reads. Negative control: 458,749/2,055,337. Round 1: 566,418/2,417,937. Round 2: 398,167/1,785,121. Round 3: 419,328/2,409,483. Round 4: 322,727/1,838,945. Round 5: 479,688/2,305,297. Of the quality-passed reads, some were identified as frameshifted due to indels if they contained at least four consecutive nucleotide mutations, and the following total reads were thus identified: Negative control: 1,386, Round 1: 11,118, Round 2: 9,465, Round 3: 15,094, Round 4: 11,473, Round 5: 21,671.

To estimate the error rate of sequencing, two independent approaches were taken. Oligonucleotides were designed to target the gene within a high copy ColE1 plasmid, and a negative control sample of this plasmid was prepared and sequenced in parallel with the iLOV libraries. The per-nucleotide mutation rate was calculated to be 0.025% versus 0.31% for the negative control and Round 1 library, respectively. Probabilistically, a mutation rate of 0.025% per nucleotide would yield an overall nucleotide mutation rate of 7.9% in the iLOV library ([1 - (1-(0.00025))^{330}]). Because every nucleotide mutation does not lead to an amino acid mutation, the nonsynonymous mutation rate will be lower than this value, as is observed in the total nonsynonymous mutation rate of 5.9% for the negative control (Table 2.4).

In order to validate this estimate for the error rate, we used a second approach to directly measure the error of per-nucleotide base-calling during sequencing of the Round 1 library. By measuring discordant base calls in the first 30 base-pairs of overlap between our forward and reverse sequencing reads, we obtain a sequencing error rate of 0.028% per nucleotide in the Round 1 library. This value predicts that 8.8% of our sequenced reads in the Round 1 library contain codon mutations due to sequencing errors. Despite this background error level, we find that the most important parameter of the library - amino acid mutation coverage - is not significantly affected by controlling for sequencing errors.

Nucleotide and amino acid sequences for iLOV and tLOV can be found in Table 2.5.

2.2d In silico Effective Library Size Simulation

Simulations were performed using a custom MATLAB script. The simulations focused on amino acid positions 5 – 40, for which we possess comprehensive frequency data from the equimolar oligonucleotide and normalized PR experiments. Successive sampling, with replacement, of these 36 positions was performed, with each position's likelihood commensurate with the observed mutation frequency. Once 34 out of 36 positions had been observed in the simulation, which we arbitrarily define as 'well-sampled' (94.4% of the library diversity), the total number of samples taken was recorded as the effective library size. This process was then repeated 104 times for each library to generate a distribution of effective library sizes.

2.2e Thermostability Screening

Colonies were screened on 10 cm dishes containing standard LB-agar with kanamycin. Tuner cells expressing the library were found to be fluorescent in the absence of induction after 24 hours. Plates were incubated at 60° C for 2 hours, after which the vast majority of colonies were no longer fluorescent. Approximately 500,000 colonies were screened, of which 244 remained fluorescent. These colonies were pooled, miniprepped, and deep sequenced to identify protein mutations. This DNA was also transformed into XJb cells to recover individual variants. 93 colonies were then grown overnight in 96-well deep well plates with 1 mL of LB + kanamycin supplemented with 100 μM Isopropyl β-D-1-thiogalactopyranoside (IPTG) and 3 mM arabinose at 37° C. Cultures were frozen and thawed to lyse the cells, then clarified by centrifugation. Supernatant was analyzed in a StepOnePlus™ Real-Time PCR System (Applied Biosystems) to estimate a Tm for each variant.

2.2f Protein Expression and in vitro Characterization

iLOV variants were expressed and lysed in XJb cells as above but in 100 mL volume. Excess FMN (Sigma) was added to the lysate to ensure all proteins possessed ligand. iLOV variants were then purified by nickel affinity chromatography using HisPur™ Ni-NTA Resin (Thermo Scientific). Variants were filtered using Vivaspin 6 3,000 molecular weight cut-off (Sartorius) with phosphate buffered saline (PBS) pH 7.4 (Gibco). Variants were further purified by size exclusion chromatography using a NGC Chromatography System (Bio-Rad). Purified proteins were stored at 4° C in PBS. For quantum yield determination, emission between 460 nm and 600 nm was measured in a FluoroLog Spectrophotometer (Horiba), and 450 nm absorbance was measured in an Infinity M1000 PRO monochromator (Tecan). Emission curves were integrated in MATLAB and normalized for absorbance. Thermostability measurements were made in a Nano differential scanning calorimeter (TA Instruments).

2.2g Accession Codes

Sequence Read Archive: Sequencing data have been deposited under accession numbers SAMN07204029, SAMN07204030, SAMN07204032, SAMN07204042, SAMN07204073, and SAMN07204112.

2.3 Results

2.3a A Single Round of PR Generates a Comprehensive Mutation Library

To generate a comprehensive mutation library of iLOV, oligonucleotides were designed to target the gene within a high copy ColE1 plasmid (Figure 2.1). The target plasmid contained a promoterless iLOV coding region to prevent growth biases between mutants possessing different fitness during multiple rounds of transformation and outgrowth. Note that this step does introduce some additional complexity, such as additional time between library generation and screening and the potential to lose library diversity. This step is not inherent to the recombineering method, however, but was chosen in order to most accurately investigate the naive iLOV library. 10 ng of target plasmid was mixed with an equimolar mixture of 110 recombineering oligonucleotides, one primer for each codon in iLOV (Figure 2.2). These oligonucleotides were 60 bp long and complementary to the lagging strand, which was previously demonstrated to be more efficient than targeting the leading strand (Lim, Min, and Jung 2008). Oligos contained a centrally located NNM mutation codon, (where N = A/C/G/T and M = A/C) which encodes all amino acids except methionine and tryptophan. Tryptophan, in particular, is known to quench flavin fluorescence in flavoproteins (Callis and Liu 2006) and was excluded from the library. Although modified oligonucleotides have been shown to enhance recombineering efficiency, e.g. phosphorothioation, standard oligonucleotides were used to minimize cost and complexity (H. H. Wang et al. 2009).

Initial experiments confirmed that PR can be used to generate diverse libraries in a programmable fashion. The plasmid and oligonucleotide mixture was first electroporated into the recombineering strain EcNR2 (H. H. Wang et al. 2009), grown overnight and miniprepped. As observed in Thomason et al. (Thomason et al. 2007) we found that a fraction of the plasmids had converted into multimers two or three times the length of the target plasmid. Because we

intended to perform multiple rounds of PR, we chose to gel extract the monomeric form of the plasmid. Notably, this step is unnecessary if the target gene is subsequently isolated by PCR or

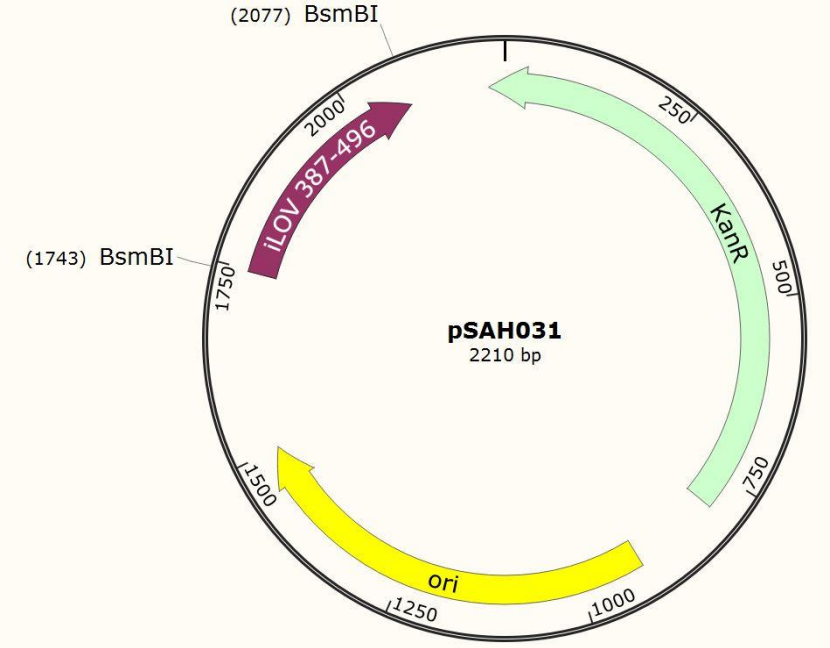


Figure 2.1: Plasmid map of pSAH031 (Addgene ID: 90330). High-copy number ColE1 bacterial plasmid used as recombineering target for library generation. Note that the kanamycin resistance gene is compatible with the recombineering strain EcNR2, which contains genomic resistance cassettes for chloramphenicol and ampicillin. The iLOV gene lacks a promotor and start codon to prevent growth biases during library construction, and is flanked by BsmBI sites for golden gate cloning directly into an expression plasmid.

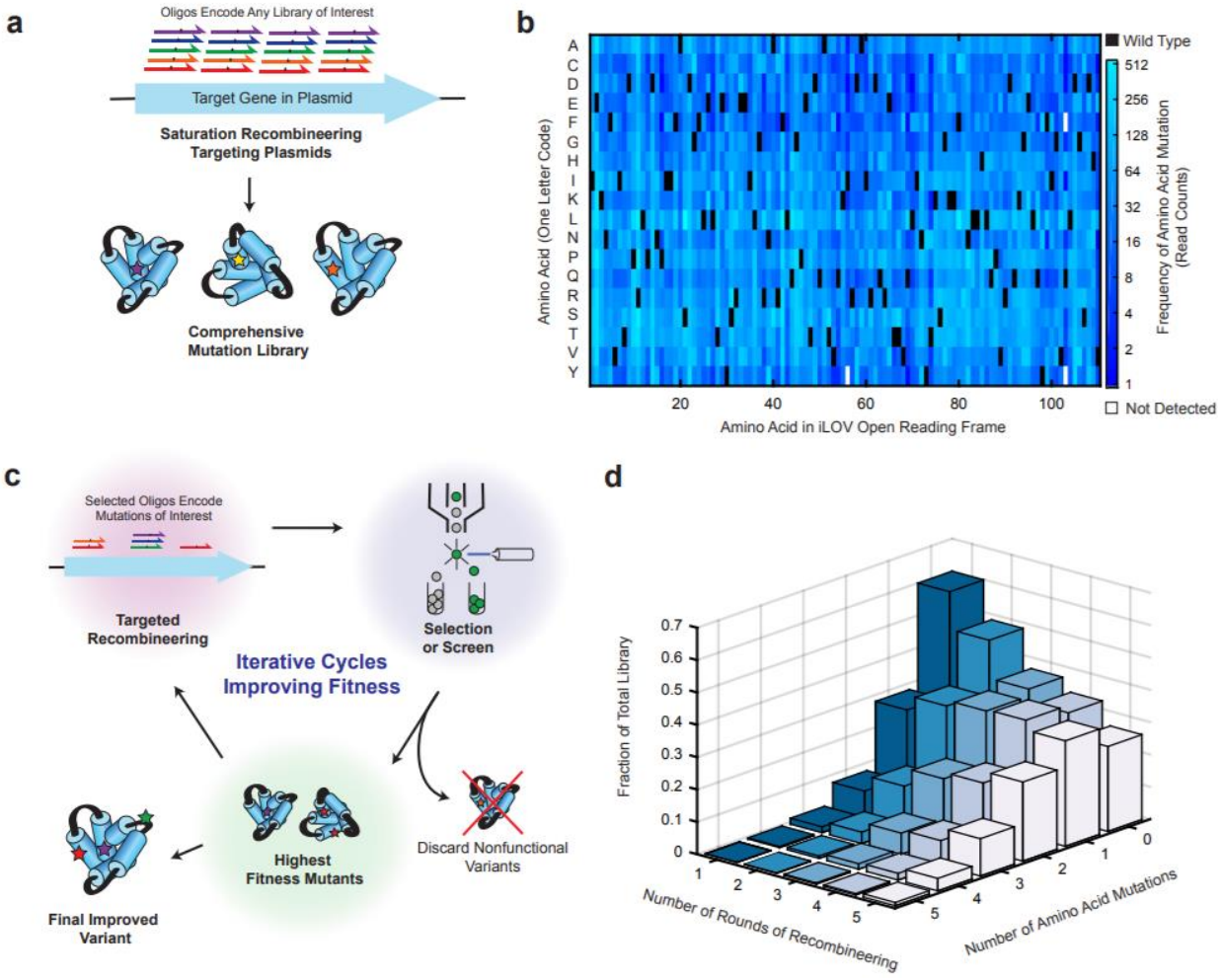
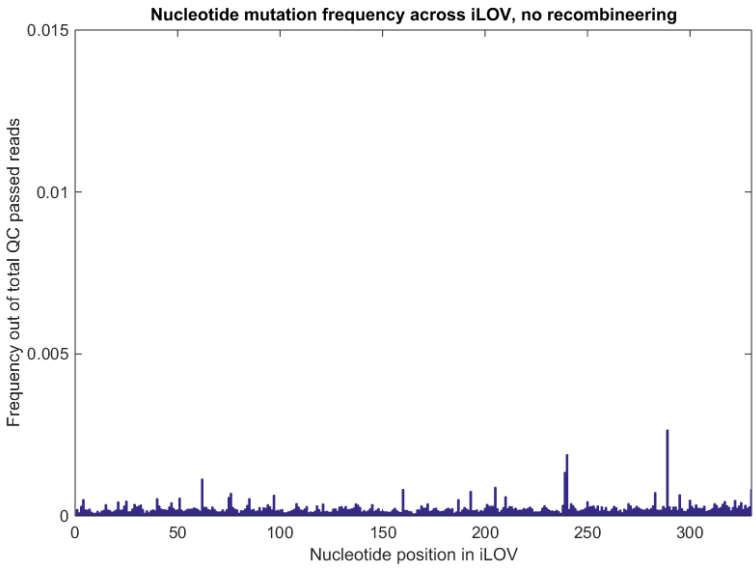


Figure 2.2: Plasmid Recombineering (PR) of iLOV generates a specific and comprehensive mutation library. (A) Cartoon of comprehensive PR accomplished using synthetic oligonucleotides tiled across the target gene. (B) Frequency of single amino acid mutations mapped by residue and location in the Round 1 library. Black indicates the WT iLOV residue. White indicates no detected reads. (C) Cartoon of programmable PR targeting a specific sequence space. Sampling of the highest fitness mutants can inform subsequent library design, with recombineering oligos specifically targeting mutations of interest. (D) Mutational distribution of the library after additional rounds of recombineering.

subcloning, especially since multimeric forms are often successfully modified plasmid molecules (Thomason et al. 2007). Deep sequencing of the recovered library, hereafter termed the Round 1 library, revealed substantial mutagenesis compared to the non-recombineered plasmid (Figure 2.3). Further analysis revealed that, while the majority of reads were WT iLOV sequence, 29% of reads contained a single amino acid mutation (Table 2.4). Note that the true recombineering efficiency for introducing a single nonsynonymous mutations is likely ~ 23% after accounting for sequencing errors relative to the negative control. These mutations covered every position in the protein and nearly all targeted amino acid conversions were observed (Figure 2.2B). 1867 out of the possible 1870 single residue mutations were detected in the Round 1 library (Table 2.4). Sequencing errors do not significantly contribute to the coverage of single nonsynonymous mutations (Figure 2.4).

Further founds of PR were used to increase the penetrance of mutations. Although the Round 1 library covered targeted mutations comprehensively, it was roughly 61% WT iLOV sequence. Furthermore, the programmable nature of PR allows the construction of more targeted libraries with combinatorial multiplexing of high fitness mutations, such as would be useful in directed evolution (Figure 2.2C). Four additional rounds of recombineering were performed to reduce the WT fraction of the library and investigate the distribution of variants with 2+ mutations. The number of reads with codon mutations increased substantially with further rounds of recombineering (Figure 2.2D). In the Round 5 library, single mutations were the most common (33%), with the remainder composed of WT sequences (26%) and sequences containing 2+ mutations (41%).

a



b

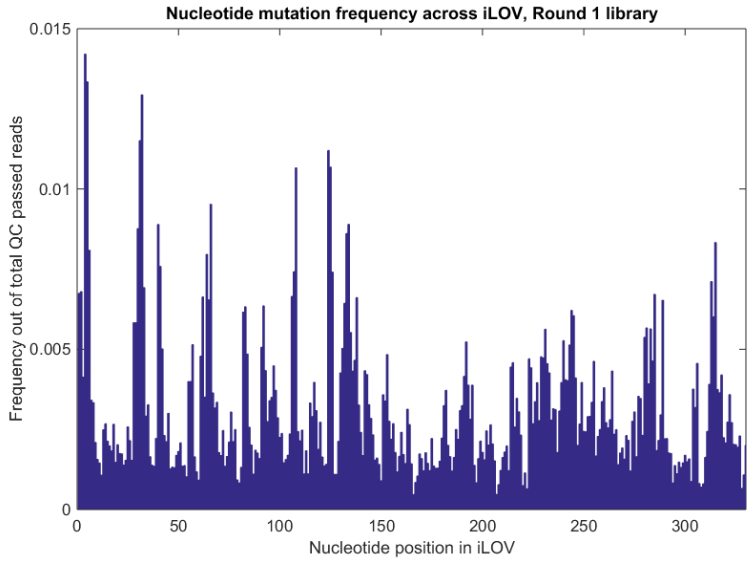


Figure 2.3: Comparative nucleotide mutation frequency demonstrates that the majority of detected mutations are due to recombineering. Mutation frequency is relative to the total number of quality control (QC) passed reads, and is 0.025% versus 0.31% per nucleotide for the negative control versus Round 1 library, respectively. The 95% confidence intervals for these measurements are all smaller than +/- 0.001% (normal approximation to the binomial distribution). (A) Nucleotide mutation frequency detected in the iLOV gene without recombineering, 458,749 QC passed reads. (B) Nucleotide mutation frequency detected in the iLOV gene with one round of recombineering (the Round 1 library), 566,418 QC passed reads.

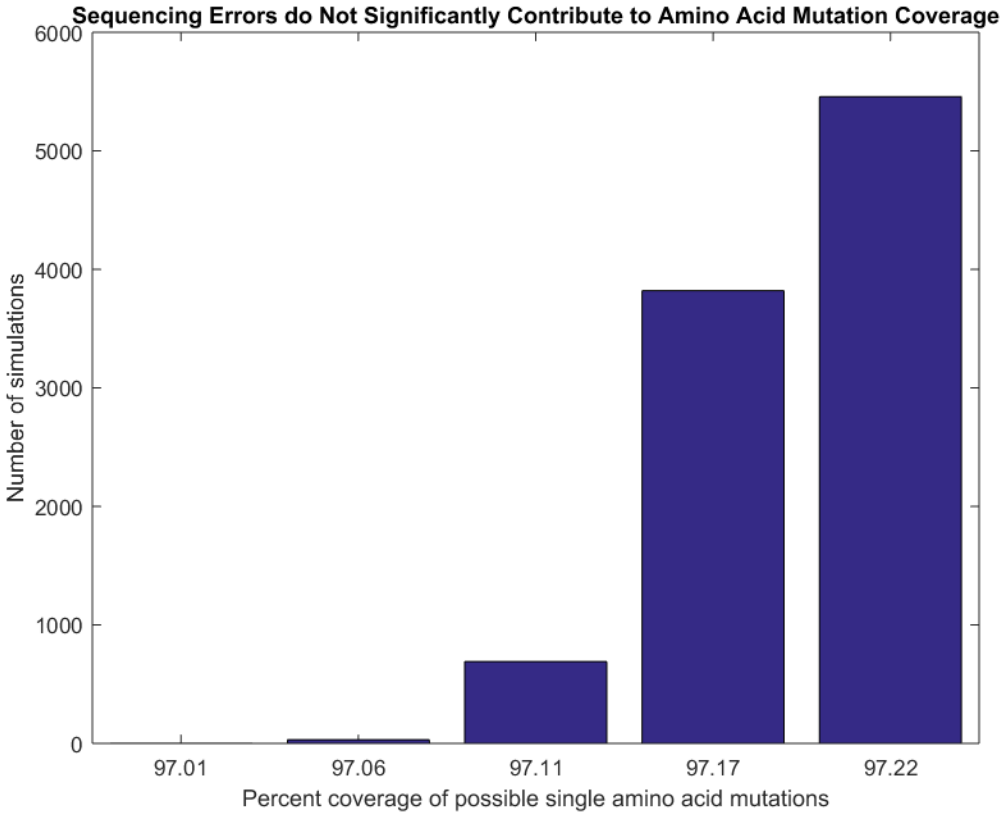


Figure 2.4: Accounting for sequencing errors does not significantly reduce the coverage of amino acid mutations among single amino acid mutants in the Round 1 library. Coverage has been calculated at a threshold of five reads. Distribution of mutation coverage for 104 simulations in silico. Coverage using all reads is 97.22%. To account for sequencing errors, a fraction of reads were randomly discarded from the pool in order to simulate an appropriate loss of mutation diversity. 20% of the 1 nonsynonymous mutation reads were discarded based on the assumption of a similar sequencing error rate in Round 1 as in the negative control, since they were prepared and sequenced in parallel. However, only reads with a single base pair mutation were eligible to be discarded. This is appropriate because sequencing errors are almost entirely composed of single base pair codon mutations due to the probabilistically independent nature of mis-calling a base. These parameters result in 51% of single base pair mutant reads being discarded in each simulation. Finally, coverage was calculated for mutants observed at least five times in order to further exclude low-frequency reads arising from sequencing errors. The lack of significant outliers along the distribution confirms that very few of the amino acid mutations observed are dependent on reads arising from sequencing errors.

2.3b Recombineering libraries can be finely controlled to alter the composition of mutations

In the absence of prior information, the ideal mutagenesis technique would produce every mutant targeted with equal frequency. Because each variant is initially present in the same amount, a uniform library requires the least amount of screening (or selection) in order to isolate improved mutants. A non-uniform method, in contrast, might produce a highly variable distribution of mutants and requires more screening or selection in order to fully explore sequence space. We therefore analyzed our sequencing data to characterize the uniformity and sequence preferences of PR libraries.

We detected two distinct types of bias: positional bias and mismatch bias. In positional bias, oligonucleotides targeted to different positions in the coding sequence incorporate with characteristically different efficiencies (regardless of the mutation produced at that position). In mismatch bias, oligonucleotides that are more similar to the WT sequence (e.g. differing only at the first base of the targeted codon) incorporated with high efficiency while more divergent oligos (e.g. different at all three positions) incorporated with lower efficiency. Positional bias is evident when comparing the frequencies of single codon mutations across all 110 codons in iLOV (Figure 2.5A). Despite the nearly 60 bp of homology between the oligonucleotide and the template plasmid, oligos targeting adjacent codons can exhibit a 2-3 fold different incorporation. The largest such discrepancy is found at codon 42, which is mutated 5.7 times more frequently than codon 41. Variation in efficiency has been observed in other recombineering studies and was somewhat correlated with the oligonucleotide binding energy (H. H. Wang et al. 2009). While our data was not clearly correlated with binding energy (Figure 2.6), a biological replicate revealed replicable positional bias (Figure 2.7), suggesting the presence of an underlying physical mechanism for positional bias.

Comprehensive mutagenesis applications, such as deep mutational scanning, ideally begin with uniformly distributed mutants in a naïve library. We hypothesized that the positional bias observed in the Round 1 library – constructed using equimolar mutagenic oligonucleotides – could be corrected by altering the mixture of oligonucleotides used in the electroporation step. A second library was therefore constructed by normalizing the concentration of each oligonucleotide in the library according to the frequency of mutations obtained at the corresponding position in Round 1. This library was constructed using PR, and the first 40 amino acids were deep sequenced. The normalized library was indeed more uniform, with a largest adjacent codon discrepancy of 2.1 fold, compared to 3.3 fold in the equimolar replication library (Figure 2.5B).

In order to understand the importance of this normalization in a quantitative fashion, we performed a simulation to evaluate the pragmatic impact of oligonucleotide normalization on effective library size – the number of samples that must be taken from a library to achieve a desired representation of the library diversity. In the ideal case, i.e. mutants are found in the library with equal frequency, if one wishes to sample, say, 95% of the diversity contained within the library, we must screen roughly three times the number of distinct members. That is, the effective library size is threefold the targeted library size. Our equimolar and normalized libraries are not uniformly distributed, but random sampling from our sequenced data in silico allows calculation of the effective library size. Our simulation revealed a substantial improvement in sampling efficiency, from a mean effective library size of 139 in the original library to 107 in the

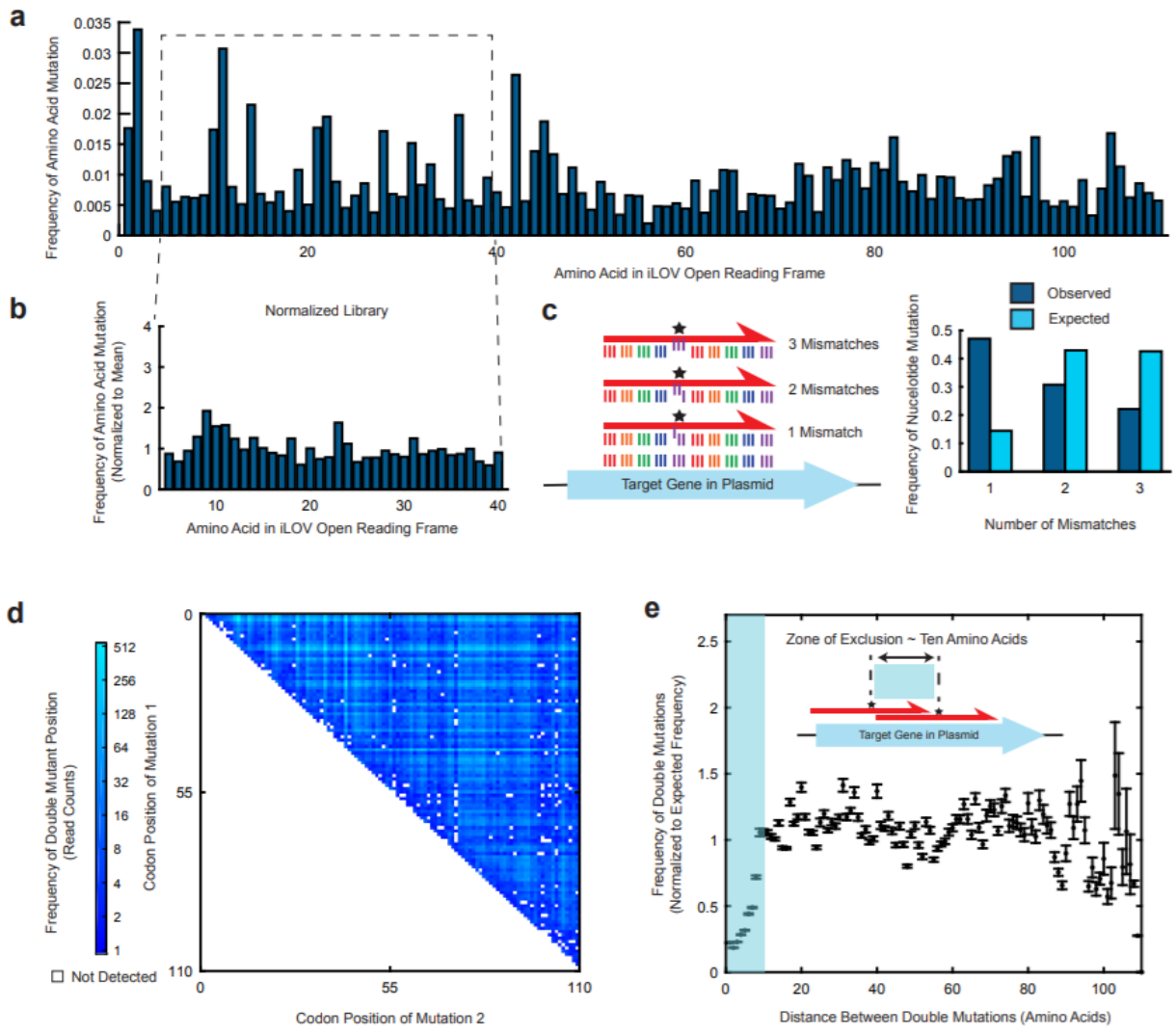
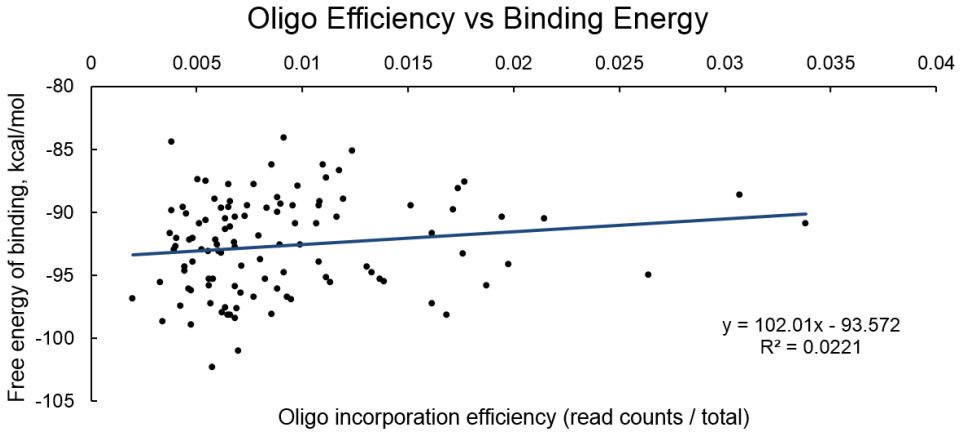


Figure 2.5: The recombineering mechanism produces moderate bias within the library and can be manipulated by varying the delivered oligonucleotide concentration. (A) Frequency of single amino acid mutations across iLOV in the Round 1 library. (B) Biological replicate of Round 1 library with normalized oligonucleotide concentrations. (C) Observed vs. expected distribution of one, two, or three nucleotide mismatches among single amino acid mutations. The 95% confidence intervals for these measurements are all smaller than +/- 0.0025 (normal approximation to the binomial distribution). (D) Frequency of double mutations in the Round 5 library. White = not detected. Highly represented rows and columns arise from positional bias. (E) Frequency of double mutations in the Round 5 library as a function of the pairwise distance between double mutations. For each distance, data has been normalized for the number of possible double mutations. Error bars, standard deviation.

a



b

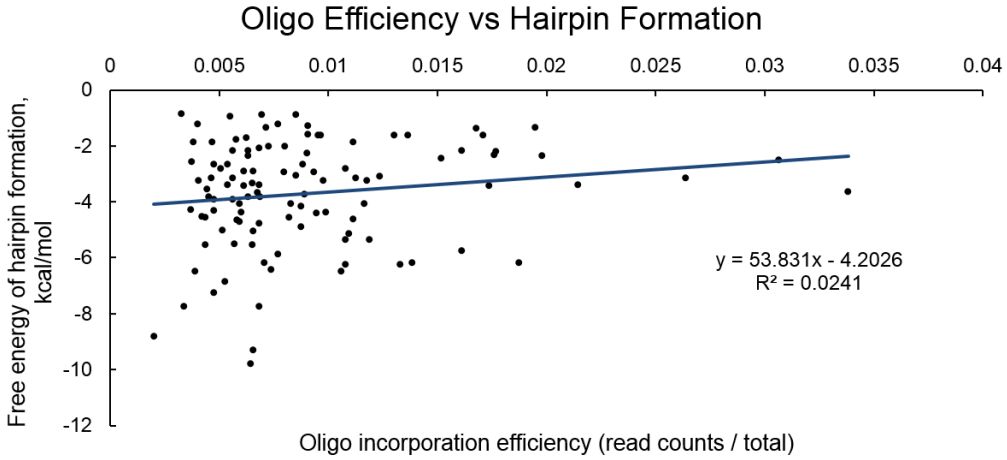


Figure 2.6: Oligonucleotide incorporation frequency is not linked to simple binding energy or hairpin formation. A). Incorporation frequency for each oligonucleotide versus free energy in kcal/mol of binding. p = 0.113 ANOVA. Binding energy was calculated through MATLAB function 'oligoprop' in bioinformatics toolbox, using nearest neighbor parameters from [7] SantaLucia Jr., J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proceedings of the National Academy of Science USA 95, 1460–1465. B). Incorporation frequency for each oligonucleotide versus free energy in kcal/mol of hairpin formation. p = 0.13 ANOVA. Hairpin formation was calculated using UNAfold (Markham 2008).

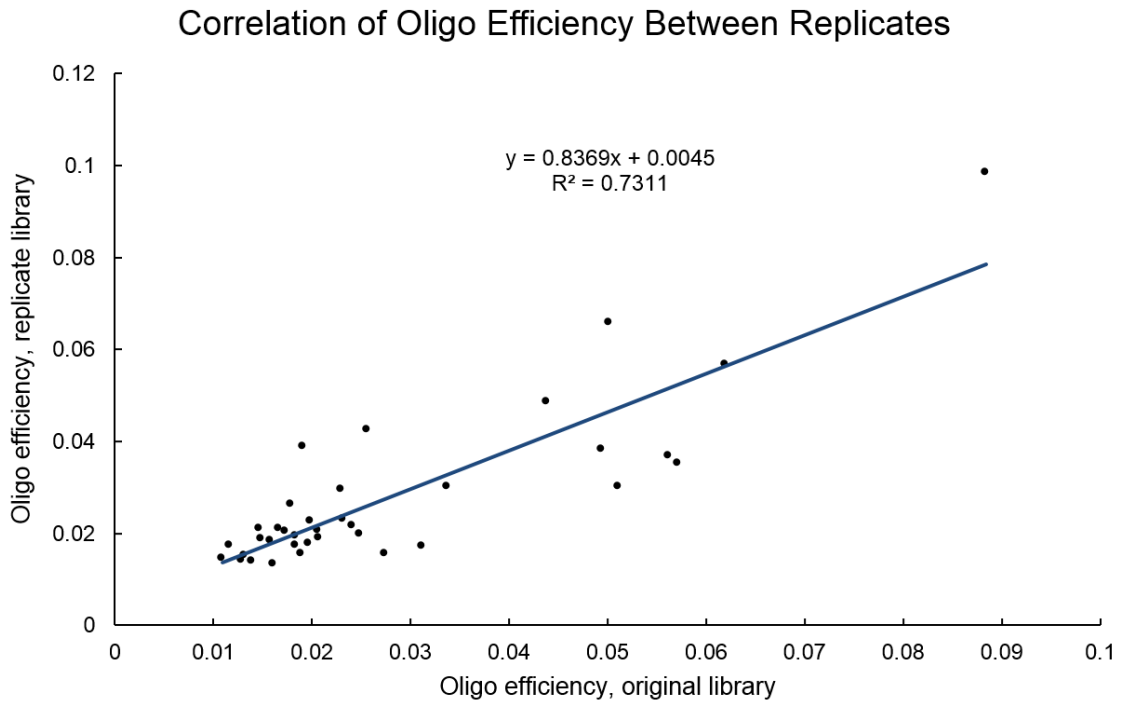


Figure 2.7: A biological replicate NNM library was independently generated and oligo efficiencies were found to correlate, indicating an underlying physical mechanism affecting recombineering efficiency. Correlation of efficiencies for oligos 5 – 40 between the initial library and replicate library (p = 6.4E-11 ANOVA).

normalized library, which is quite close to that of an ideal uniformly distributed library (Figure 2.8).

The presence of mismatch bias in the library demonstrates that recombineering favors incorporation of oligos that are more similar to the WT template sequence. Our oligonucleotides all contained one, two, or three nucleotide mismatches relative to WT iLOV. After computationally enumerating all possible oligonucleotides for each codon in iLOV, we expected that 14% of oligonucleotides would contain one mismatch, 43% two, and 43% three. In contrast, oligonucleotide-template pairs with two or three mismatches were observed at only 31% and 22% respectively, while single nucleotide mismatches were overrepresented by more than threefold at 47% (Figure 2.5C). This value is inflated due to the presence of sequencing errors, of which the vast majority are single mismatches, but is still significantly higher than expected given the error rate in the negative control (Table 2.4). Together, mismatch bias and positional bias are substantial sources of variation in the library, accounting for roughly 60% of model variance (Figure 2.9).

While a single round of PR generates many single mutants, additional rounds shift the distribution to increasing mutation numbers. The Round 5 library, for example, is 25% double mutants. These mutations are well represented, with 5817 out of a possible 5940 locations detected (Figure 2.5D). However, it is clear that the same positional bias seen in the Round 1 library is preserved in subsequent rounds. This is to be expected if recombineering events are statistically independent, and can be seen by the correlation of double-mutation hotspots with

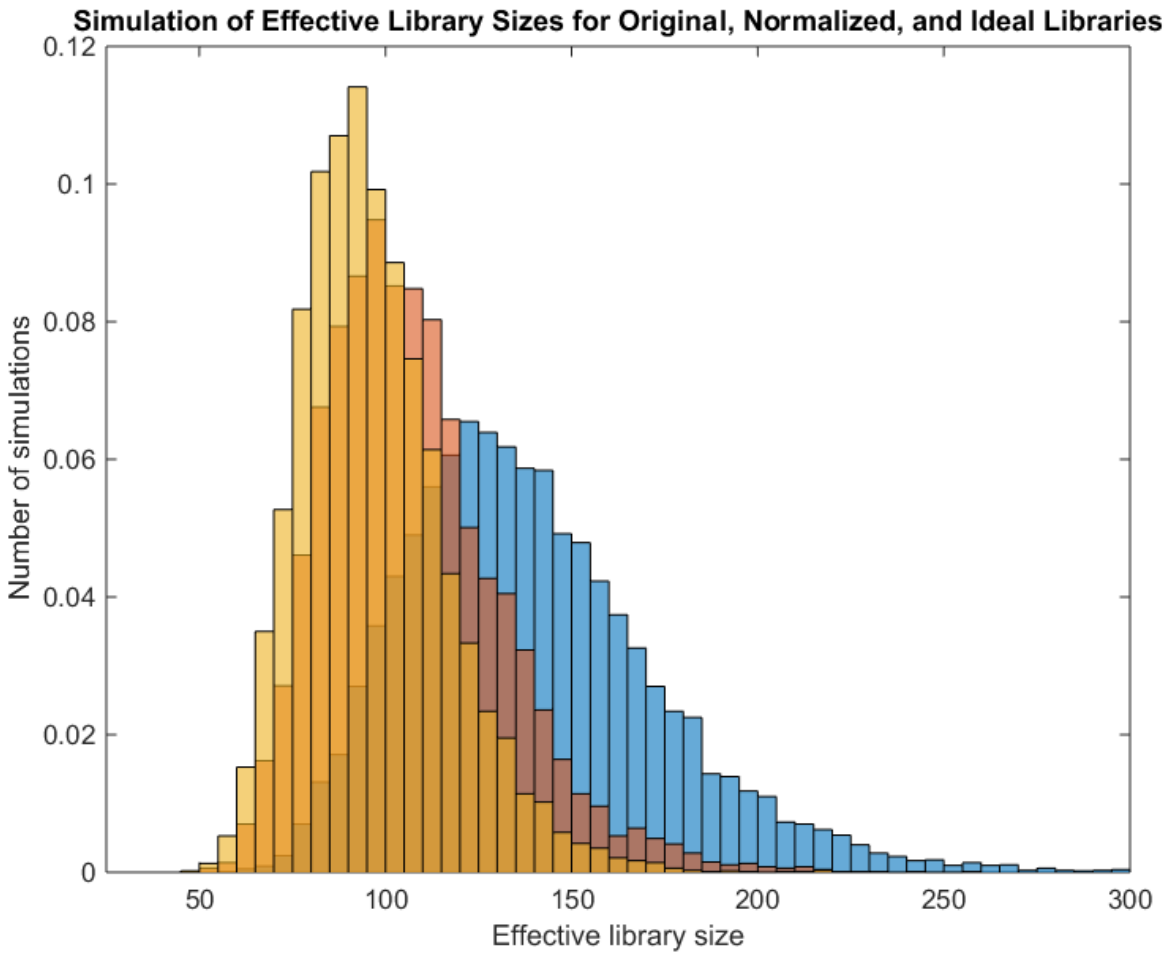
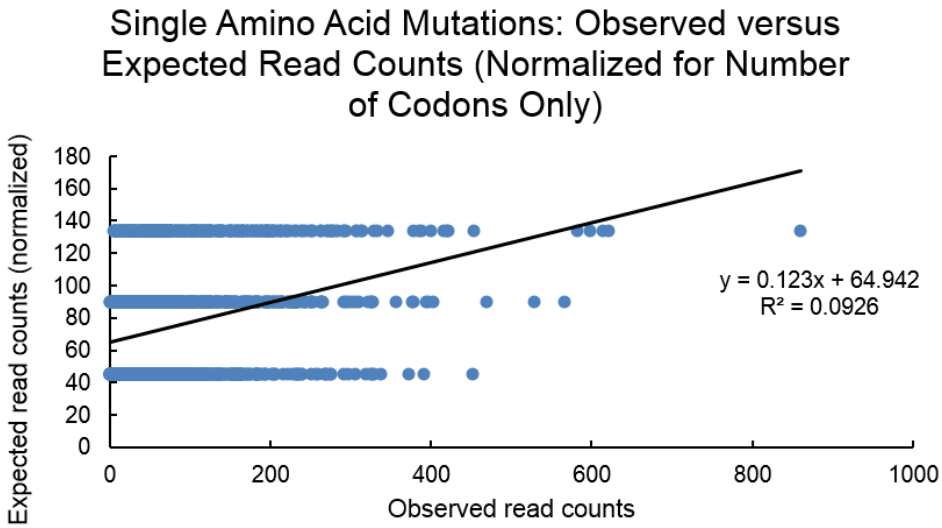
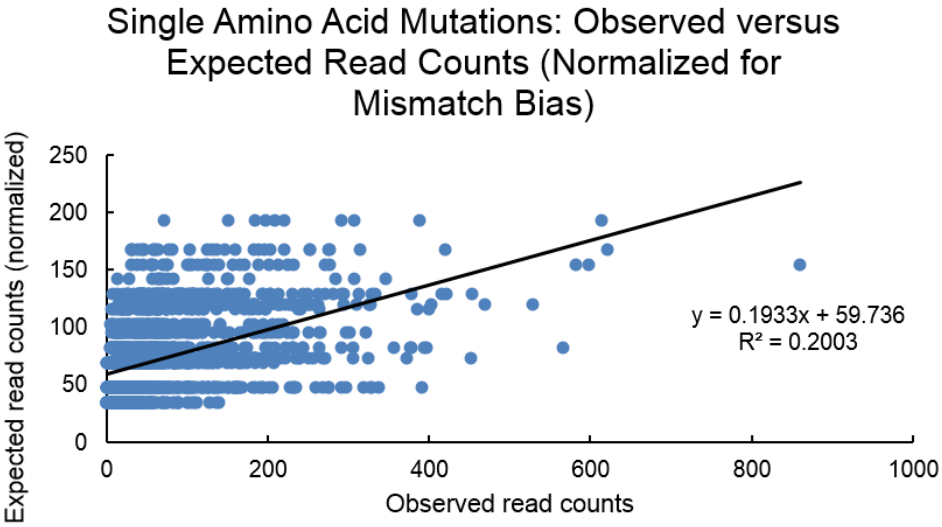


Figure 2.8: Simulation of effective library size reveals substantial improvement in screening effectiveness using a normalized library. Distribution of effective library sizes for 104 simulations in silico for the original library (blue) vs the normalized library (red) vs an ideal uniformly distributed library (orange). $p < 10^{-10}$ by two tailed z-test comparing the original vs normalized distribution.

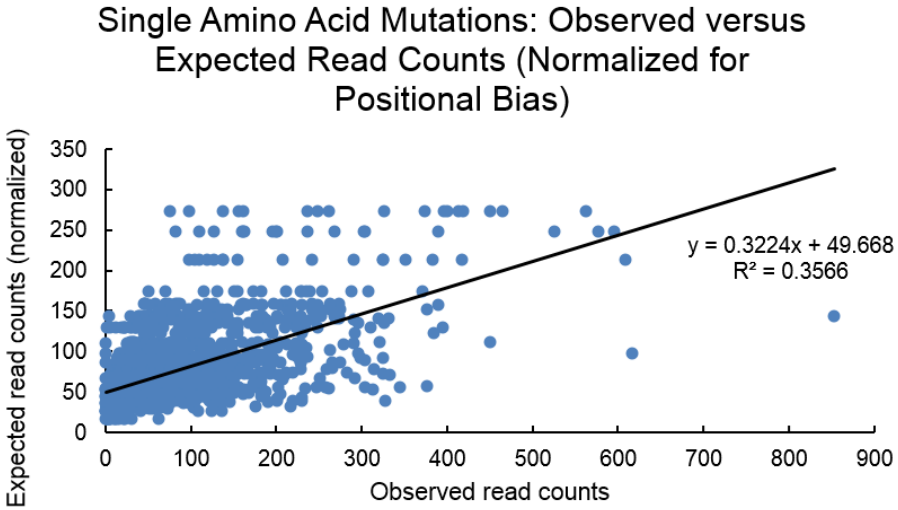
a


Single Amino Acid Mutations: Observed versus Expected Read Counts (Normalized for Number of Codons Only)
Expected read counts (normalized)
Observed read counts
y = 0.123x + 64.942
R² = 0.0926
0
20
40
60
80
100
120
140
160
180
0
200
400
600
800
1000

b


Single Amino Acid Mutations: Observed versus Expected Read Counts (Normalized for Mismatch Bias)
Expected read counts (normalized)
Observed read counts
y = 0.1933x + 59.736
R² = 0.2003
0
50
100
150
200
250
0
200
400
600
800
1000

c



d

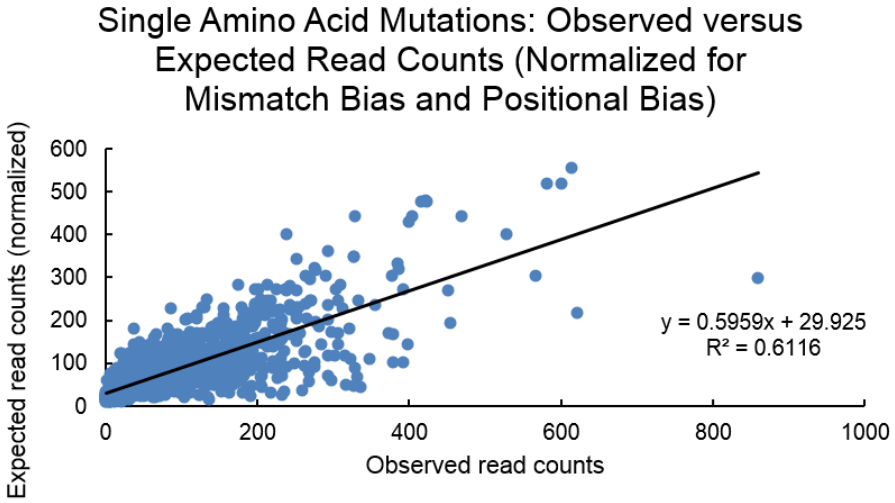


Figure 2.9: Correlation plots of observed read counts versus expected read counts for single amino acid mutations in the Round 1 library. Each data point corresponds to one amino acid (i.e. all codons are lumped together) at one position in the iLOV sequence. A) Correlation plot of observed read counts versus expected read counts, normalized only for the number of codons programmed for each amino acid. Note that the finite number of codons from an NNM-mutagenizing primer generates a highly discretized expected read count. B) As in (A), except normalized for the observed mismatch bias (i.e. on a per-codon basis). C) As in (A), except normalized for positional bias. D) As in (A), except normalized for both mismatch and positional bias. The coefficient of determination (i.e. R2) indicates that including mismatch and positional bias can account for 61% of the variance between the counts of expected and observed sequences at the amino acid level.

single mutation positional bias (Figure 2.10). This data suggests that double mutations in a normalized library will be far more uniformly distributed.

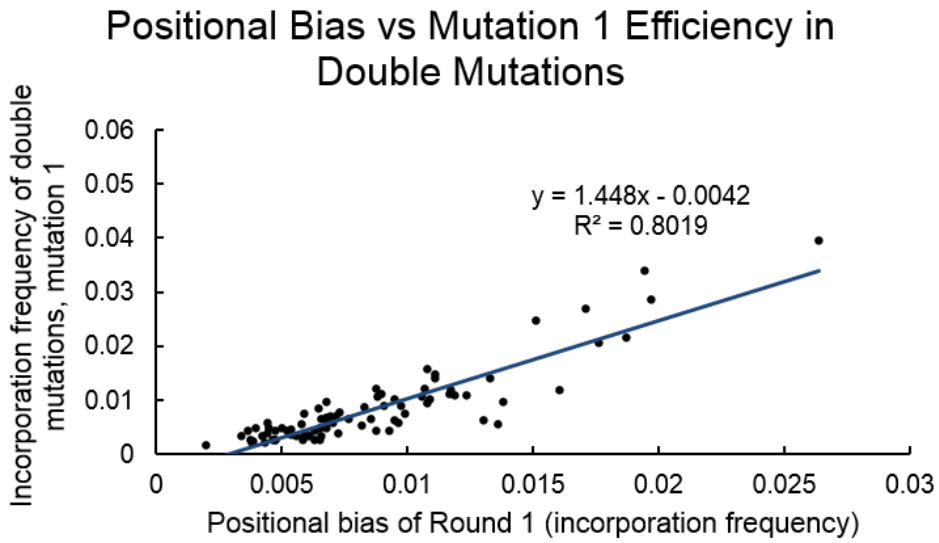
The double mutation data also indicates that some recombineering events are not perfectly independent. A plot of the pairwise distance between all detected double mutants reveals an uneven distribution in frequency (Figure 2.5E). Specifically, double mutations are less likely to be within 30 bp (i.e. 10 amino acids) of one another. This effect becomes more pronounced with additional rounds of recombineering (Figure 2.11). In the Round 5 library, this 'zone of exclusion' is significant enough that double mutations a few amino acids apart are nearly four times less frequent than double mutations with a much larger separation (e.g. three amino acid gap versus 10 amino acid gap). We hypothesize that this bias is due to the mechanism of recombineering which requires oligonucleotides to anneal to a complementary locus via homology arms flanking the NNM mutagenic codon. In this mechanism, sequential oligonucleotides would 'overwrite' previous mutations due to incorporation of the most recent oligonucleotide's homology arms, which extend 30 bp on either side of the central NNM. Another potential mechanism that disfavors incorporation of nearby double mutants is mutational reduction of oligo incorporation efficiency. In this hypothetical mechanism, mutations generated in early rounds of PR could reduce the homology and, therefore, incorporation efficiency of oligos in subsequent PR rounds.

2.3c Screening the iLOV library identifies mutations conferring thermostability

Previous screening and structural work indicates that a well-packed binding site for the FMN fluorophore may lead to improved photochemical properties of iLOV, such as photostability, by limiting the dynamics of the FMN chromophore and its ability to dissipate energy following excitation (Christie et al. 2012). Additionally, searches for improved LOV-based fluorescent reporters have turned up homologous variants such as CreiLOV that, while brighter (50% greater quantum yield), exhibit substantial toxicity upon expression (Mukherjee et al. 2015). More generally, Tawfik and colleagues have theorized that thermostable proteins serve as more fruitful starting points for engineering and directed evolution (Tokuriki and Tawfik 2009). In this view, variants with greater thermostability can better tolerate mutations, increasing the likelihood of observing mutations that improve protein function in a manner unrelated to thermostability. We thus investigated the thermostability of our library, which contained nearly every single amino acid mutation and a small fraction of possible double mutations.

We created a plate-based assay to screen for thermostabilized variants in the Round 5 library. The library was plated at high colony density onto standard LB-Agar, grown overnight, then incubated at 60 °C for two hours (Figure 2.12A). This treatment completely abrogates the fluorescence of WT iLOV as well as nearly all mutants (Figure 2.12B). Some colonies remained fluorescent, however, and these were recovered and expressed in 96-well plate format. Cultures expressing these library members were lysed, clarified, and analyzed for thermostability. Fluorescence measurements were taken every 0.5° C during a temperature ramp from 25° C to 95° C and used to calculate the melting temperature (Tm) for each clone, the temperature at which 50% of maximal fluorescence is retained (Figure 2.12C). All 93 assayed clones demonstrated a substantial increase in thermostability relative to iLOV (Figure 2.12D). Improvements in Tm ranged from two to nearly ten degrees C.

a



b

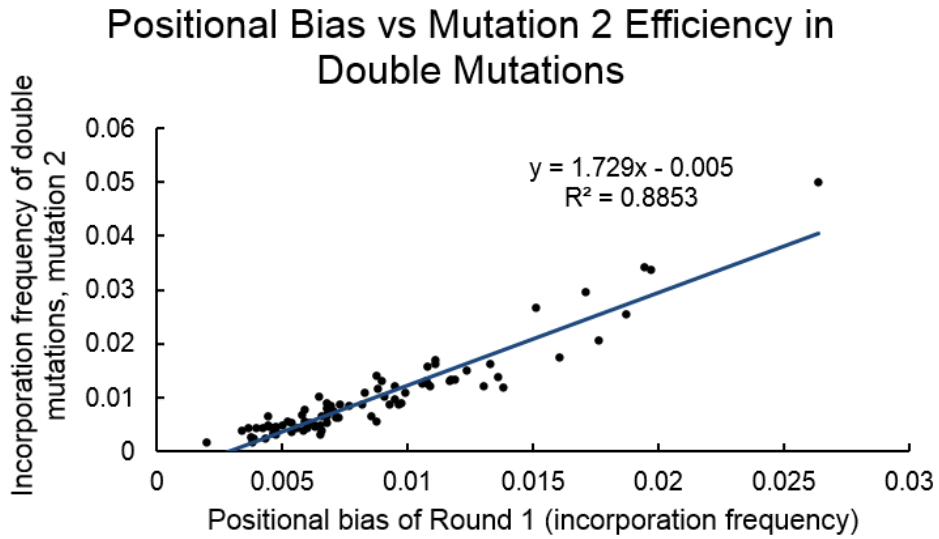


Figure 2.10: The non-uniformity of double mutations occurs largely due to positional bias of successive rounds of recombineering. Reducing positional bias (i.e. by generating a normalized library) would thus increase the uniformity of double mutations. A) Correlation plot of positional bias at each amino acid of iLOV vs the bias observed in mutation 1 of double mutations. Frequencies of double mutations have been normalized for the total number of possible mutations for each distance category in order to facilitate comparison with positional bias. Data for the first and last 14 amino acids have been excluded because normalization by small numbers artificially increases the data variability, while the 'zone of exclusion' also suppresses mutations in the extreme top-left and bottom-right of the double mutation heatmap (Figure 2D) ($p < 8.6E-30$ ANOVA). B) Data as in A, but displaying mutation 2 of double mutations ($p < 6.5E-39$ ANOVA).

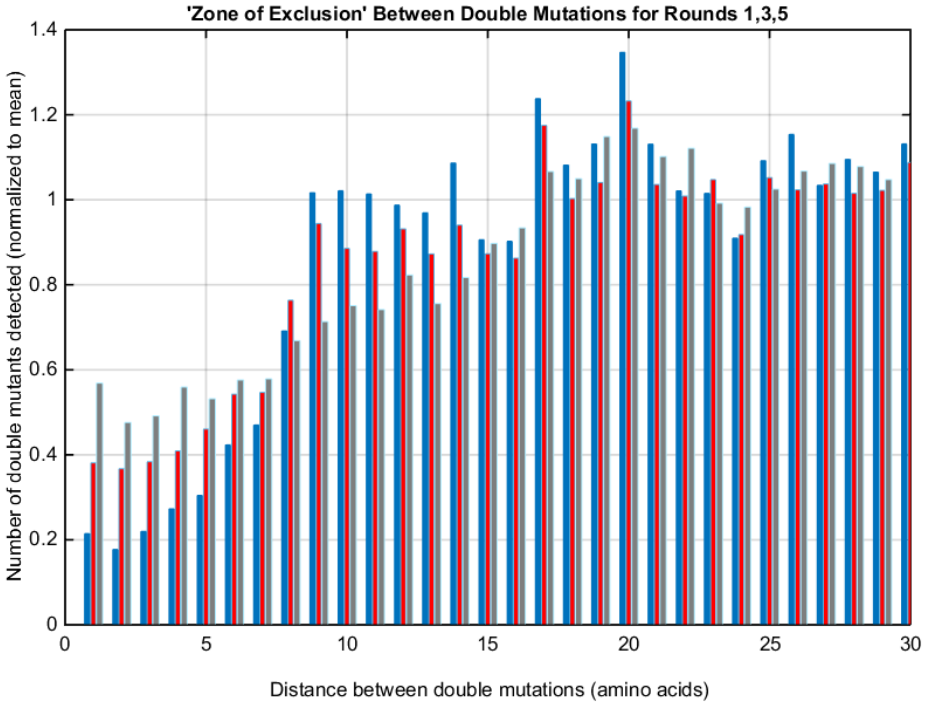


Figure 2.11: The 'zone of exclusion' for nearby amino acid mutations becomes more pronounced with increasing rounds of PR. Pairwise distance between double mutations for rounds 1 (Gray), 3 (Red), and 5 (Blue). Data has been normalized for the total number of possible mutations for each distance category, then normalized to the mean in order to facilitate comparison between rounds.

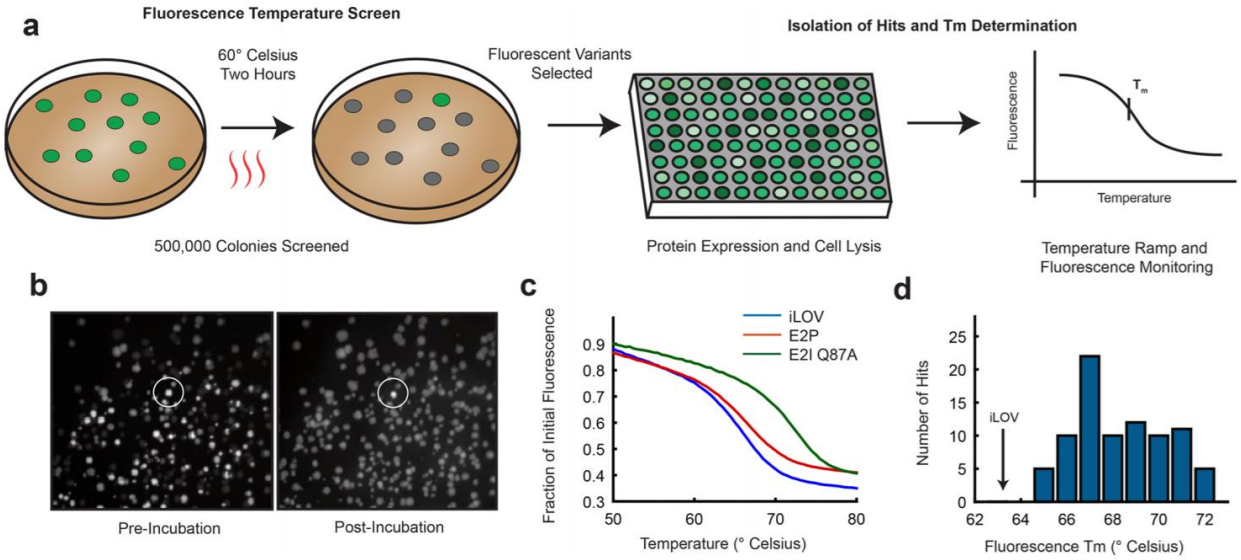


Figure 2.12: A plate-based thermostability screen identifies mutations that improve iLOV fluorescence at elevated temperatures. (A) Cartoon of the thermostability screen assay and subsequent hit validation procedure. (B) Representative fluorescent images of library colonies before and after temperature challenge. (C) Representative fluorescent thermal melt curves of lysate for two thermostable hits. (D) Histogram of Tms for 93 thermostabilized iLOV variants.

2.3d Recombineering based multiplexing allows rapid and robust directed evolution

While comprehensive single mutant libraries are ideal for exploring structure-function relationships in an unbiased manner, hypothesis-driven investigation requires targeted libraries for exploring the effect of mutations at specific sites of interest. Because PR targets only sites programmed by synthetic oligonucleotides, we hypothesized that PR could generate a specific library in a cost-effective and straightforward protocol.

To demonstrate the rapid multiplexing capability of PR, we designed a second library containing the top 25 most frequent thermostabilizing mutations from the initial plate screen (Figure 2.13A). Several of these mutations consisted of alternative amino acids at the same position, and in these cases a different oligonucleotide was designed for each. The encoding oligonucleotides were designed such that homology arms would stop short of neighboring mutations so as not to overwrite them (Figure 2.14).

This multiplexed library resulted in striking improvements to thermostability, with the best variants having Tm values nearly 20° C greater than iLOV (Figure 2.13B). Isolating and re-cloning these variants verified that despite the significant increase in Tms, the shape of their melt curves was not significantly different from that of iLOV, even for the most thermostable mutant (Figure 2.13C). It was noted during the screening process that one variant in particular, hereafter referred to as thermostable LOV (tLOV), seemed to produce abnormally bright lysate under high expression conditions. tLOV and iLOV were expressed and purified in parallel and their absorbance and emission were characterized in vitro. To accurately quantify relative quantum yield, the emission curves were normalized for absorption at 450 nm and integrated. tLOV was found to be approximately 10% brighter (Figure 2.15), and sequencing revealed the presence of four mutations scattered throughout the protein, all introduced by PR (Figure 2.13D). Notably, none of these locations are directly within the FMN binding pocket, and their contribution to thermostability or quantum yield improvement are not obvious. Thus, it would have been difficult to rationally design tLOV. Moreover, a targeted mutagenesis technique is required for isolating quadruple mutants reliably: iLOV is a small protein, but a quadruple mutant library would contain $> 10^{13}$ variants, well beyond our current screening capacities. Finally, thermostability was verified by differential scanning calorimetry of iLOV vs tLOV (Figure 2.16).

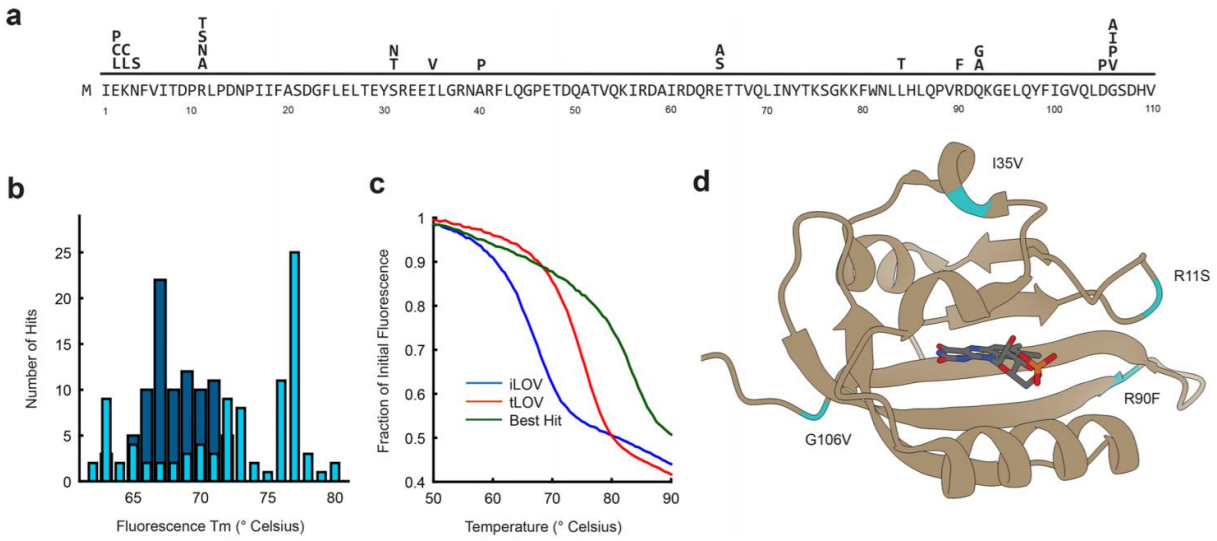


Figure 2.13: Multiplexing thermostabilizing mutations rapidly identifies doubly improved iLOV variants. (A) 25 thermostabilizing mutations mapped to the iLOV protein sequence. (B) Histogram of initial thermostable hits (Blue) superimposed on hits from the multiplexed library (Cyan). (C) Representative melt curves of two thermostabilized variants compared to iLOV. (D) Crystal structure of iLOV (PDB 4EES) indicating the location of four mutations found in tLOV: R11S, I35V, R90F, G106V.

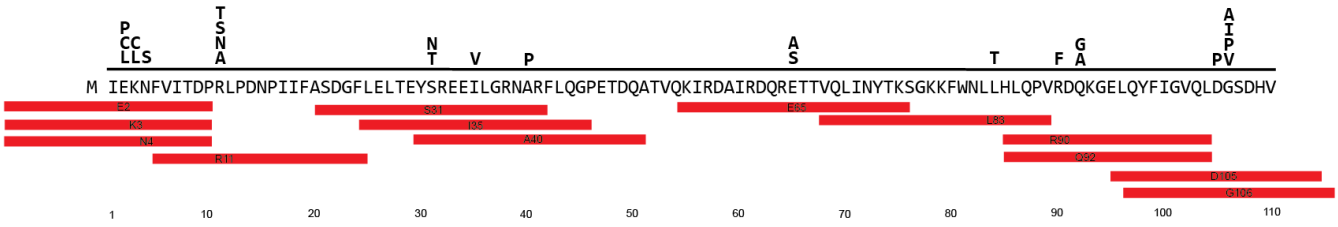


Figure 2.14: The multiplexed mutation library was designed to incorporate the 25 most frequent thermostabilizing mutations from the initial thermostability plate screen. These oligos were designed in such a way as to minimize the 'overwriting' that might occur from the homology arms of sequential oligonucleotides, illustrated above. For some oligonucleotides this means the mutagenic codon is located asymmetrically.

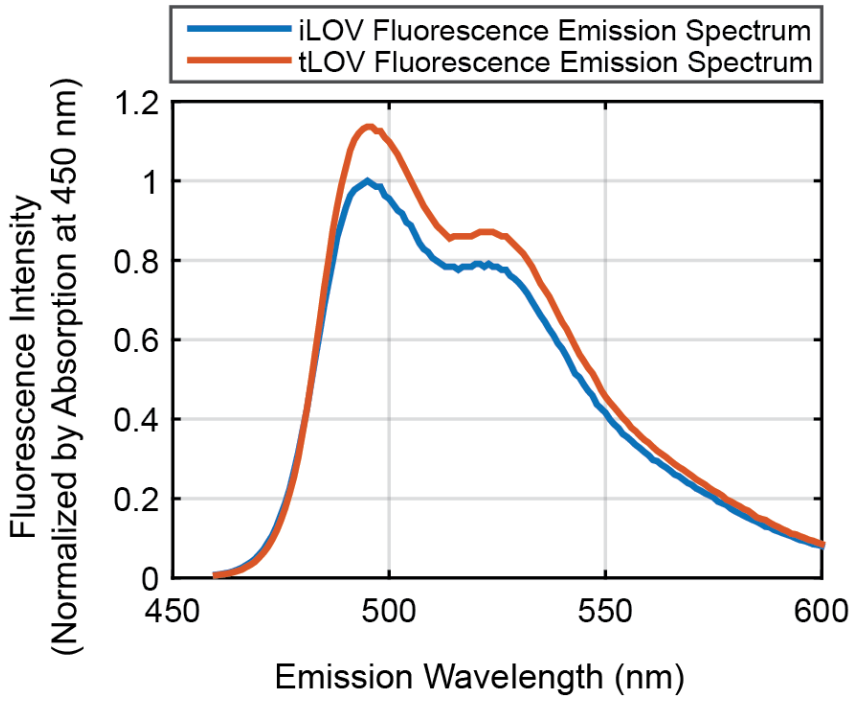


Figure 2.15: Emission spectrum of purified iLOV vs. tLOV, normalized for absorbance at 450 nm. Integration of peak area indicates tLOV is approximately 10% brighter, in addition to a 10 °C improvement in fluorescence Tm.

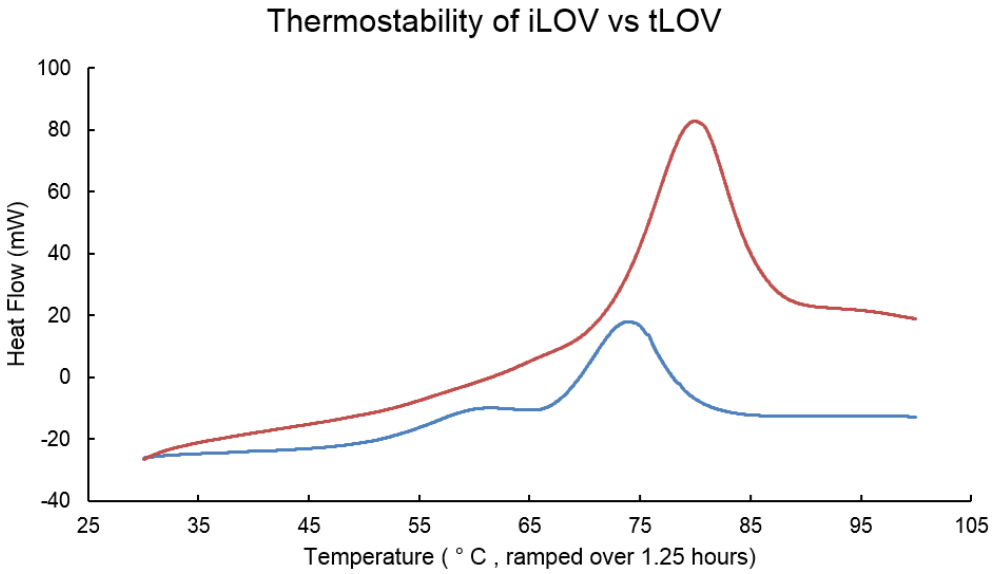


Figure 2.16: Differential scanning calorimetry of purified proteins confirms the thermostability of iLOV (Blue) relative to tLOV (Red). The fluorescence Tm of iLOV and tLOV is approximately 65 °C and 75 °C for the purified proteins, respectively. Interestingly, this appears to occur before the peak heat capacity of protein unfolding for tLOV, and in between two peaks of the biphasic iLOV heat capacity.

2.4 Discussion

An ideal method for generating comprehensive protein libraries would be simple and robust, enabling both complete and targeted mutagenesis without a change in reagents. In this study, we demonstrate that PR can be effectively used for both comprehensive and programmable mutagenesis of the fluorescent protein iLOV, using methods that are generalizable to any gene of interest.

Previous recombineering studies have successfully built libraries in the genome, demonstrating specificity and fine control of mutational composition. Wang et al.(H. H. Wang et al. 2009) used multiple rounds of recombineering to mutagenize six consecutive nucleotides using 90 bp oligonucleotides and found a 75% mutation rate after five rounds. We find quite similar behavior in the sequencing analysis of the iLOV recombineered plasmid libraries constructed here. Our Round 5 library consisted of 74% mutant variants, and the mutations were well distributed in sequence space. The modest positional bias could be substantially ameliorated by altering the ratios of oligonucleotides added to the electroporation mixture. This resulted in a nearly uniform distribution of mutations such that the effective library size was almost ideal (Figure 2.8). This approach could easily be used to accommodate more complex libraries with weighted mutation frequencies at various locations. The correlation of replicate libraries strongly suggests an underlying physical mechanism for differing oligonucleotide recombineering efficiencies. It will be important to understand this effect in order to predict efficiencies rather than rely on empirical data as used here, which required additional sequencing and labor. Regardless of these modifications, the reagent cost and experimental effort remain low – standard 60 bp oligonucleotides and simple cycles of electroporation, growth, and plasmid isolation.

Additional biases resulting from the mechanism of recombineering were detected and while their magnitude was smaller, their effect on library size and screening can be significant. Previous work has found that recombineering efficiency drops sharply with the size of the modification made (H. H. Wang et al. 2009). Here, too, single nucleotide mutations were observed to be more common than expected. This effect alters the distribution of codons present in the final library, with the template codons determining this frequency shift. Notably, mismatch bias is intrinsic to the annealing of oligonucleotides and is likely present for in vitro methods as well.

The ‘zone of exclusion’ around a first mutation generates another mode of bias. A second mutation is less likely to appear inside this zone than outside of it. This effect became more pronounced with increasing rounds of recombineering. Is it likely this results from oligonucleotides’ homology arms overwriting earlier mutations during later rounds of PR. In other words, we hypothesized that homology arms are capable of introducing revertant mutations. We thus predict that the length of homology arms would impact the length of the ‘zone of exclusion.’ As some amount of homology is absolutely required for recombineering, no form of recombineering is suitable for efficiently generating adjacent mutations from different oligonucleotides. This limitation can likely be overcome by using single oligonucleotides encoding sequential mutations but at the cost of reduced efficiency due to mismatch bias. Wang et al.(H. H. Wang et al. 2009) observed similar mutation rates for between one and four mismatches, but found efficiency dropped by ~half when incorporating mismatches of five or ten nucleotides. Again, if this effect fundamentally stems from the annealing of oligonucleotides then it is likely present for in vitro methods as well.

In vitro methods represent the most powerful alternatives for generating comprehensive mutation libraries. The chief advantage of these approaches is that they can generate a library composed mostly of mutated variants. Ostermeier and colleagues accomplished this by utilizing specialized protocols based on uracil-containing template DNA (Firnberg and Ostermeier 2012), while Wrenbeck and Whitehead selectively degrade the WT strand using single-strand nicking endonucleases (Wrenbeck et al. 2016). Such methods have been used to generate nearly comprehensive libraries of genes for exploring the entirety of the fitness landscape (Firnberg et al. 2014), and to comprehensively examine the possible evolutionary pathways leading from one allele to another (Steinberg and Ostermeier 2016).

In directed protein evolution, iterative rounds of mutagenesis can be used to multiplex fitness-improving mutations. PCR-based protocols (Firnberg and Ostermeier 2012), direct gene synthesis (Melnikov et al. 2014), and some other recombineering techniques(Garst et al. 2016) excel at generating libraries composed of single mutations. However, in many applications, 2+ mutations are desired at non-contiguous locations. Recent work has developed a PCR-based method to accomplish this goal in vitro (Belsare et al. 2016), and we hypothesized that PR was well suited to serve as a complementary approach in vivo, doing away with cloning altogether. To this end, we comprehensively explored the iLOV single mutation sequence-space for thermostability, selected the fitness enhancing mutations, and demonstrated the utility of PR for advanced protein engineering by multiplexing many different single and double mutations at discontinuous sites across iLOV in a second library. The ability to select and easily mutate numerous specific and non-contiguous locations across a protein is highly useful for a variety of techniques that utilize experimental or phylogenetic data to computationally predict and enhance enzymes (Heinzelman et al. 2009), explore epistatic interactions (Olson, Wu, and Sun 2014), or even scan SNPs in human proteins for disease prediction (Majithia et al. 2016).

iLOV engineering has been relatively limited in comparison to other fluorescent proteins (Chapman et al. 2008). Because the domain has been taken out of its natural structural context, we hypothesized that its thermal stability could be increased. Consistent with this idea, many mutations were found to improve the thermal stability of iLOV up to a robust, 10° C increase in Tm. These improvements were then stacked by multiplexing the 25 most common mutations from the first screen. As a measure of convenience, the same pooled, single electroporation protocol was used, although iterative rounds of transformation and outgrowth set a lower limit on throughput – typically ~16 hours per cycle. One particular variant among the thermostabilized pool, tLOV, was found to be ~10% brighter that iLOV in vitro. This result is consistent with previous work demonstrating that constraining the FMN fluorophore can improve the photochemical properties of iLOV (Christie et al. 2012). It would be interesting to perform comprehensive mutagenesis of the thermostabilized iLOV mutants in search of further improvements to the protein's brightness or red/blue spectral shifting, as increased thermostability has been hypothesized to permit greater exploration of function-altering mutations (Tokuriki and Tawfik 2009).

In summary, we have demonstrated that PR retains many of the ideal properties of genome recombineering, including specificity and programmability. We found that PR was suited for the construction of both comprehensive and targeted libraries, and that the simplicity of the protocol led to rapid and reliable screening experiments. In particular, PR is suitable for cycles of iterative design, construction, and sampling of genetic libraries, requiring no specialized reagents or protocols. We developed a thermostability screen of the fluorescent protein iLOV and used the resulting mutation data to rapidly construct a multiplexed library that

identified significantly improved variants, including the first enhancement to the protein's brightness since its development.

2.5 Tables

iLOV_NNM_1	CGCGGGTCGGTGATAACGAAGTTTTTTTCKNNGGACAGAGACGtgcattaatgaatcggc
iLOV_NNM_2	AAGCGCGGGTCGGTGATAACGAAGTTTTTKNNGATGGACAGAGACGtgcattaatgaatc
iLOV_NNM_3	GGTAAGCGCGGGTCGGTGATAACGAAGTTKNNTTCGATGGACAGAGACGtgcattaatga
iLOV_NNM_4	TCAGGTAAGCGCGGGTCGGTGATAACGAAKNNTTTTTCGATGGACAGAGACGtgcattaa
iLOV_NNM_5	TTGTCAGGTAAGCGCGGGTCGGTGATAACKNNGTTTTTTTCGATGGACAGAGACGtgcat
iLOV_NNM_6	GGATTGTCAGGTAAGCGCGGGTCGGTGATKNNGAAGTTTTTTTTCGATGGACAGAGACGtg
iLOV_NNM_7	ATTGGATTGTCAGGTAAGCGCGGGTCGGTKNNAACGAAGTTTTTTTCGATGGACAGAGAC
iLOV_NNM_8	ATGATTGGATTGTCAGGTAAGCGCGGGTCKNNGATAACGAAGTTTTTTTCGATGGACAGA
iLOV_NNM_9	AAGATGATTGGATTGTCAGGTAAGCGCGGKNNGGTGATAACGAAGTTTTTTTCGATGGAC
iLOV_NNM_10	GCAAAGATGATTGGATTGTCAGGTAAGCGKNNGTCGGTGATAACGAAGTTTTTTTCGATG
iLOV_NNM_11	GAGGCAAAGATGATTGGATTGTCAGGTAAKNNCGGGTCGGTGATAACGAAGTTTTTTTCG
iLOV_NNM_12	TCTGAGGCAAAGATGATTGGATTGTCAGGKNNGCGCGGGTCGGTGATAACGAAGTTTTTT
iLOV_NNM_13	CCATCTGAGGCAAAGATGATTGGATTGTCKNNTAAGCGCGGGTCGGTGATAACGAAGTTT
iLOV_NNM_14	AAACCATCTGAGGCAAAGATGATTGGATTKNNAGGTAAGCGCGGGTCGGTGATAACGAAG
iLOV_NNM_15	AGAAAACCATCTGAGGCAAAGATGATTGGKNNGTCAGGTAAGCGCGGGTCGGTGATAACG
iLOV_NNM_16	TCGAGAAAACCATCTGAGGCAAAGATGATKNNATTGTCAGGTAAGCGCGGGTCGGTGATA
iLOV_NNM_17	AGCTCGAGAAAACCATCTGAGGCAAAGATKNNTGGATTGTCAGGTAAGCGCGGGTCGGTG
iLOV_NNM_18	GTTAGCTCGAGAAAACCATCTGAGGCAAAKNNGATTGGATTGTCAGGTAAGCGCGGGTCG
iLOV_NNM_19	TCGGTTAGCTCGAGAAAACCATCTGAGGCKNNGATGATTGGATTGTCAGGTAAGCGCGGG
iLOV_NNM_20	TATTCGGTTAGCTCGAGAAAACCATCTGAKNNAAAGATGATTGGATTGTCAGGTAAGCGC
iLOV_NNM_21	CTGTATTCGGTTAGCTCGAGAAAACCATCKNNGGCAAAGATGATTGGATTGTCAGGTAAG
iLOV_NNM_22	CGGCTGTATTCGGTTAGCTCGAGAAAACCKNNTGAGGCAAAGATGATTGGATTGTCAGGT
iLOV_NNM_23	TCACGGCTGTATTCGGTTAGCTCGAGAAAKNNATCTGAGGCAAAGATGATTGGATTGTCA
iLOV_NNM_24	TCTTCACGGCTGTATTCGGTTAGCTCGAGKNNACCATCTGAGGCAAAGATGATTGGATTG
iLOV_NNM_25	ATTTCTTCACGGCTGTATTCGGTTAGCTCKNNAAAACCATCTGAGGCAAAGATGATTGGA
iLOV_NNM_26	AGAATTTCTTCACGGCTGTATTCGGTTAGKNNGAGAAAACCATCTGAGGCAAAGATGATT
iLOV_NNM_27	CCCAGAATTTCTTCACGGCTGTATTCGGTKNNCTCGAGAAAACCATCTGAGGCAAAGATG
iLOV_NNM_28	CGACCCAGAATTTCTTCACGGCTGTATTCKNNTAGCTCGAGAAAACCATCTGAGGCAAAG
iLOV_NNM_29	TTACGACCCAGAATTTCTTCACGGCTGTAKNNGGTTAGCTCGAGAAAACCATCTGAGGCA
iLOV_NNM_30	GCATTACGACCCAGAATTTCTTCACGGCTKNNTTCGGTTAGCTCGAGAAAACCATCTGAG
iLOV_NNM_31	CGAGCATTACGACCCAGAATTTCTTCACGKNNGTATTCGGTTAGCTCGAGAAAACCATCT
iLOV_NNM_32	AAACGAGCATTACGACCCAGAATTTCTTCKNNGCTGTATTCGGTTAGCTCGAGAAAACCA
iLOV_NNM_33	AGGAAACGAGCATTACGACCCAGAATTTCKNNACGGCTGTATTCGGTTAGCTCGAGAAAA

iLOV_NNM_34	TGCAGGAAACGAGCATTACGACCCAGAATKNNTTCACGGCTGTATTCGGTTAGCTCGAGA
iLOV_NNM_35	CCCTGCAGGAAACGAGCATTACGACCCAGKNNTTCTTCACGGCTGTATTCGGTTAGCTCG
iLOV_NNM_36	GGGCCCTGCAGGAAACGAGCATTACGACCKNNAATTTCTTCACGGCTGTATTCGGTTAGC
iLOV_NNM_37	TCCGGGCCCTGCAGGAAACGAGCATTACGKNNCAGAATTTCTTCACGGCTGTATTCGGTT
iLOV_NNM_38	GTTTCCGGGCCCTGCAGGAAACGAGCATTKNNACCCAGAATTTCTTCACGGCTGTATTCG
iLOV_NNM_39	TCGGTTTCCGGGCCCTGCAGGAAACGAGCKNNACGACCCAGAATTTCTTCACGGCTGTAT
iLOV_NNM_40	TGGTCGGTTTCCGGGCCCTGCAGGAAACGKNNATTACGACCCAGAATTTCTTCACGGCTG
iLOV_NNM_41	GCTTGGTCGGTTTCCGGGCCCTGCAGGAAKNNAGCATTACGACCCAGAATTTCTTCACGG
iLOV_NNM_42	GTAGCTTGGTCGGTTTCCGGGCCCTGCAGKNNACGAGCATTACGACCCAGAATTTCTTCA
iLOV_NNM_43	ACCGTAGCTTGGTCGGTTTCCGGGCCCTGKNNGAAACGAGCATTACGACCCAGAATTTCT
iLOV_NNM_44	TGAACCGTAGCTTGGTCGGTTTCCGGGCCKNNCAGGAAACGAGCATTACGACCCAGAATT
iLOV_NNM_45	TTCTGAACCGTAGCTTGGTCGGTTTCCGGKNNCTGCAGGAAACGAGCATTACGACCCAGA
iLOV_NNM_46	ATTTTCTGAACCGTAGCTTGGTCGGTTTCKNNGCCCTGCAGGAAACGAGCATTACGACCC
iLOV_NNM_47	CGAATTTTCTGAACCGTAGCTTGGTCGGTKNNCGGGCCCTGCAGGAAACGAGCATTACGA
iLOV_NNM_48	TCGCGAATTTTCTGAACCGTAGCTTGGTCKNNTTCCGGGCCCTGCAGGAAACGAGCATTA
iLOV_NNM_49	GCATCGCGAATTTTCTGAACCGTAGCTTGKNNGGTTTCCGGGCCCTGCAGGAAACGAGCA
iLOV_NNM_50	ATTGCATCGCGAATTTTCTGAACCGTAGCKNNGTCGGTTTCCGGGCCCTGCAGGAAACGA
iLOV_NNM_51	CGAATTGCATCGCGAATTTTCTGAACCGTKNNTTGGTCGGTTTCCGGGCCCTGCAGGAAA
iLOV_NNM_52	TCGCGAATTGCATCGCGAATTTTCTGAACKNNAGCTTGGTCGGTTTCCGGGCCCTGCAGG
iLOV_NNM_53	TGATCGCGAATTGCATCGCGAATTTTCTGKNNCGTAGCTTGGTCGGTTTCCGGGCCCTGC
iLOV_NNM_54	CGCTGATCGCGAATTGCATCGCGAATTTTKNNAACCGTAGCTTGGTCGGTTTCCGGGCCC
iLOV_NNM_55	TCACGCTGATCGCGAATTGCATCGCGAATKNNCTGAACCGTAGCTTGGTCGGTTTCCGGG
iLOV_NNM_56	GTCTCACGCTGATCGCGAATTGCATCGCGKNNTTTCTGAACCGTAGCTTGGTCGGTTTCC
iLOV_NNM_57	GTAGTCTCACGCTGATCGCGAATTGCATCKNNAATTTTCTGAACCGTAGCTTGGTCGGTT
iLOV_NNM_58	ACGGTAGTCTCACGCTGATCGCGAATTGCKNNGCGAATTTTCTGAACCGTAGCTTGGTCG
iLOV_NNM_59	TGAACGGTAGTCTCACGCTGATCGCGAATKNNATCGCGAATTTTCTGAACCGTAGCTTGG
iLOV_NNM_60	AATTGAACGGTAGTCTCACGCTGATCGCGKNNTGCATCGCGAATTTTCTGAACCGTAGCT
iLOV_NNM_61	ATCAATTGAACGGTAGTCTCACGCTGATCKNNAATTGCATCGCGAATTTTCTGAACCGTA
iLOV_NNM_62	TTGATCAATTGAACGGTAGTCTCACGCTGKNNGCGAATTGCATCGCGAATTTTCTGAACC
iLOV_NNM_63	TAGTTGATCAATTGAACGGTAGTCTCACGKNNATCGCGAATTGCATCGCGAATTTTCTGA
iLOV_NNM_64	GTGTAGTTGATCAATTGAACGGTAGTCTCKNNCTGATCGCGAATTGCATCGCGAATTTTC
iLOV_NNM_65	TTGGTGTAGTTGATCAATTGAACGGTAGTKNNACGCTGATCGCGAATTGCATCGCGAATT
iLOV_NNM_66	CTTTTGGTGTAGTTGATCAATTGAACGGTKNNCTCACGCTGATCGCGAATTGCATCGCGA
iLOV_NNM_67	CCGCTTTTGGTGTAGTTGATCAATTGAACKNNAGTCTCACGCTGATCGCGAATTGCATCG

iLOV_NNM_68	TTACCGCTTTTGGTGTAGTTGATCAATTGKNNGGTAGTCTCACGCTGATCGCGAATTGCA
iLOV_NNM_69	TTTTTACCGCTTTTGGTGTAGTTGATCAAKNNAACGGTAGTCTCACGCTGATCGCGAATT
iLOV_NNM_70	AATTTTTTACCGCTTTTGGTGTAGTTGATKNNTTGAACGGTAGTCTCACGCTGATCGCGA
iLOV_NNM_71	CAAAATTTTTTACCGCTTTTGGTGTAGTTKNNCAATTGAACGGTAGTCTCACGCTGATCG
iLOV_NNM_72	TTCCAAAATTTTTTACCGCTTTTGGTGTAKNNGATCAATTGAACGGTAGTCTCACGCTGA
iLOV_NNM_73	AGGTTCCAAAATTTTTTACCGCTTTTGGTKNNGTTGATCAATTGAACGGTAGTCTCACGC
iLOV_NNM_74	AACAGGTTCCAAAATTTTTTACCGCTTTTKNNGTAGTTGATCAATTGAACGGTAGTCTCA
iLOV_NNM_75	TGCAACAGGTTCCAAAATTTTTTACCGCTKNNGGTGTAGTTGATCAATTGAACGGTAGTC
iLOV_NNM_76	AGATGCAACAGGTTCCAAAATTTTTTACCKNNTTTGGTGTAGTTGATCAATTGAACGGTA
iLOV_NNM_77	TGAAGATGCAACAGGTTCCAAAATTTTTTKNNGCTTTTGGTGTAGTTGATCAATTGAACG
iLOV_NNM_78	GGTTGAAGATGCAACAGGTTCCAAAATTTKNNACCGCTTTTGGTGTAGTTGATCAATTGA
iLOV_NNM_79	ACCGGTTGAAGATGCAACAGGTTCCAAAAKNNTTTACCGCTTTTGGTGTAGTTGATCAAT
iLOV_NNM_80	CGTACCGGTTGAAGATGCAACAGGTTCCAKNNTTTTTTACCGCTTTTGGTGTAGTTGATC
iLOV_NNM_81	TCGCGTACCGGTTGAAGATGCAACAGGTTKNNAAATTTTTTACCGCTTTTGGTGTAGTTG
iLOV_NNM_82	TGGTCGCGTACCGGTTGAAGATGCAACAGKNNCCAAAATTTTTTACCGCTTTTGGTGTAG
iLOV_NNM_83	TTTTGGTCGCGTACCGGTTGAAGATGCAAKNNGTTCCAAAATTTTTTACCGCTTTTGGTG
iLOV_NNM_84	CCTTTTTGGTCGCGTACCGGTTGAAGATGKNNCAGGTTCCAAAATTTTTTACCGCTTTTG
iLOV_NNM_85	TCGCCTTTTTGGTCGCGTACCGGTTGAAGKNNCAACAGGTTCCAAAATTTTTTACCGCTT
iLOV_NNM_86	AGCTCGCCTTTTTGGTCGCGTACCGGTTGKNNATGCAACAGGTTCCAAAATTTTTTACCG
iLOV_NNM_87	TGAAGCTCGCCTTTTTGGTCGCGTACCGGKNNAAGATGCAACAGGTTCCAAAATTTTTTA
iLOV_NNM_88	TACTGAAGCTCGCCTTTTTGGTCGCGTACKNNTTGAAGATGCAACAGGTTCCAAAATTTT
iLOV_NNM_89	AAGTACTGAAGCTCGCCTTTTTGGTCGCGKNNCGGTTGAAGATGCAACAGGTTCCAAAAT
iLOV_NNM_90	ATAAAGTACTGAAGCTCGCCTTTTTGGTCKNNTACCGGTTGAAGATGCAACAGGTTCCAA
iLOV_NNM_91	CCGATAAAGTACTGAAGCTCGCCTTTTTGKNNGCGTACCGGTTGAAGATGCAACAGGTTC
iLOV_NNM_92	ACGCCGATAAAGTACTGAAGCTCGCCTTTKNNGTCGCGTACCGGTTGAAGATGCAACAGG
iLOV_NNM_93	TGCACGCCGATAAAGTACTGAAGCTCGCCKNNTTGGTCGCGTACCGGTTGAAGATGCAAC
iLOV_NNM_94	AGCTGCACGCCGATAAAGTACTGAAGCTCKNNTTTTTGGTCGCGTACCGGTTGAAGATGC
iLOV_NNM_95	TCCAGCTGCACGCCGATAAAGTACTGAAGKNNGCCTTTTTGGTCGCGTACCGGTTGAAGA
iLOV_NNM_96	CCATCCAGCTGCACGCCGATAAAGTACTGKNNCTCGCCTTTTTGGTCGCGTACCGGTTGA
iLOV_NNM_97	CTCCCATCCAGCTGCACGCCGATAAAGTAKNNAAGCTCGCCTTTTTGGTCGCGTACCGGT
iLOV_NNM_98	TCACTCCCATCCAGCTGCACGCCGATAAAKNNCTGAAGCTCGCCTTTTTGGTCGCGTACC
iLOV_NNM_99	TGGTCACTCCCATCCAGCTGCACGCCGATKNNGTACTGAAGCTCGCCTTTTTGGTCGCGT
iLOV_NNM_100	ACATGGTCACTCCCATCCAGCTGCACGCCKNNAAAGTACTGAAGCTCGCCTTTTTGGTCG
iLOV_NNM_101	CTCACATGGTCACTCCCATCCAGCTGCACKNNGATAAAGTACTGAAGCTCGCCTTTTTGG

iLOV_NNM_102	TCGCTCACATGGTCACTCCCATCCAGCTGKNNGCCGATAAAGTACTGAAGCTCGCCTTTT
iLOV_NNM_103	CTCTCGCTCACATGGTCACTCCCATCCAGKNNCACGCCGATAAAGTACTGAAGCTCGCCT
iLOV_NNM_104	CGTCTCTCGCTCACATGGTCACTCCCATCKNNCTGCACGCCGATAAAGTACTGAAGCTCG
iLOV_NNM_105	cagCGTCTCTCGCTCACATGGTCACTCCCKNNCAGCTGCACGCCGATAAAGTACTGAAGC
iLOV_NNM_106	cgtcagCGTCTCTCGCTCACATGGTCACTKNNATCCAGCTGCACGCCGATAAAGTACTGA
iLOV_NNM_107	gcccgtcagCGTCTCTCGCTCACATGGTCKNNCCCATCCAGCTGCACGCCGATAAAGTAC
iLOV_NNM_108	caagcccgtcagCGTCTCTCGCTCACATGKNNACTCCCATCCAGCTGCACGCCGATAAAG
iLOV_NNM_109	agacaagcccgtcagCGTCTCTCGCTCACKNNGTCACTCCCATCCAGCTGCACGCCGATA
iLOV_NNM_110	agcagacaagcccgtcagCGTCTCTCGCTKNNATGGTCACTCCCATCCAGCTGCACGCCG

Table 2.1: Recombineering oligonucleotides used in construction of the iLOV comprehensive library. 110 oligonucleotides target each of the 110 codons of the iLOV open reading frame in plasmid pSAH031. The cost of this oligo library is approximately $1000 at 10 nmole scale synthesis in plate form through Integrated DNA Technologies.

SAH_186	CGGGTCGGTGATAACGAAGTTTTTCGGGATGGACAGAGACGtgcattaatgaatcggcca	E2P
SAH_187	CGGGTCGGTGATAACGAAGTTTTTAAGGATGGACAGAGACGtgcattaatgaatcggcca	E2L
SAH_188	CGGGTCGGTGATAACGAAGTTTTTGCAGATGGACAGAGACGtgcattaatgaatcggcca	E2C
SAH_189	CGGGTCGGTGATAACGAAGTTGCATTCGATGGACAGAGACGtgcattaatgaatcggcca	K3C
SAH_190	CGGGTCGGTGATAACGAAGTTAAGTTCGATGGACAGAGACGtgcattaatgaatcggcca	K3L
SAH_191	CGGGTCGGTGATAACGAAGCTTTTTTCGATGGACAGAGACGtgcattaatgaatcggcca	N4S
SAH_192	AAAACCATCTGAGGCAAAGATGATTGGATTGTCAGGTAACGTCGGGTCGGTGATAACGAA	R11T
SAH_193	AAAACCATCTGAGGCAAAGATGATTGGATTGTCAGGTAAGCTCGGGTCGGTGATAACGAA	R11S
SAH_194	AAAACCATCTGAGGCAAAGATGATTGGATTGTCAGGTAAATTCGGGTCGGTGATAACGAA	R11N
SAH_195	AAAACCATCTGAGGCAAAGATGATTGGATTGTCAGGTAAAGCCGGGTCGGTGATAACGAA	R11A
SAH_196	GCATTACGACCCAGAATTTCTTCACGattGTATTCGGTTAGCTCGAGAAAACCATCTGAG	S31N
SAH_197	GCATTACGACCCAGAATTTCTTCACGggtGTATTCGGTTAGCTCGAGAAAACCATCTGAG	S31T
SAH_198	CTGCAGGAAACGAGCATTACGACCCAGaacTTCTTCACGGCTGTATTCGGTTAGCTCGAG	I35V
SAH_199	TCGGTTTCCGGGCCCTGCAGGAAACGcggATTACGACCCAGAATTTCTTCACGGCTGTAT	A40P
SAH_200	GTGTAGTTGATCAATTGAACGGTAGTtgcACGCTGATCGCGAATTGCATCGCGAATTTTC	E65A
SAH_201	GTGTAGTTGATCAATTGAACGGTAGTgctACGCTGATCGCGAATTGCATCGCGAATTTTC	E65S
SAH_202	TACCGGTTGAAGATGCAAggtGTTCCAAAATTTTTTACCGCTTTTGGTGTAGTTGATCAA	L83T
SAH_203	CTGCACGCCGATAAAGTACTGAAGCTCGCCTTTTTGGTCaaaTACCGGTTGAAGATGCAA	R90F
SAH_204	CAGCTGCACGCCGATAAAGTACTGAAGCTCGCCTTTtgcGTCGCGTACCGGTTGAAGATG	Q92A
SAH_205	CAGCTGCACGCCGATAAAGTACTGAAGCTCGCCTTTaccGTCGCGTACCGGTTGAAGATG	Q92G
SAH_206	CGTCTCTCGCTCACATGGTCACTCCCcggCAGCTGCACGCCGATAAAGTACTGAAGCTCG	D105P
SAH_207	cagCGTCTCTCGCTCACATGGTCACTtgcATCCAGCTGCACGCCGATAAAGTACTGAAGC	G106A
SAH_208	cagCGTCTCTCGCTCACATGGTCACTgatATCCAGCTGCACGCCGATAAAGTACTGAAGC	G106I
SAH_209	cagCGTCTCTCGCTCACATGGTCACTcggATCCAGCTGCACGCCGATAAAGTACTGAAGC	G106P
SAH_210	cagCGTCTCTCGCTCACATGGTCACTaacATCCAGCTGCACGCCGATAAAGTACTGAAGC	G106V

Table 2.2: Recombineering oligonucleotides used in the construction of the multiplexed thermostability library. Each oligonucleotide targets one amino acid mutation, as indicated in the rightmost column. The cost of this oligo library was approximately $330 at 25 nmole scale from Integrated DNA Technologies.

SAH_162	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-NNNNNtcattaatgcaCGTCTCTGTCC
SAH_163	CAAGCAGAAGACGGCATACGAGAT-GCGTGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcgtcagCGTCTCTCGCT
SAH_164	CAAGCAGAAGACGGCATACGAGAT-ATAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcgtcagCGTCTCTCGCT
SAH_178	CAAGCAGAAGACGGCATACGAGAT-TAGGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcgtcagCGTCTCTCGCT
SAH_179	CAAGCAGAAGACGGCATACGAGAT-CCGTATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcgtcagCGTCTCTCGCT
SAH_180	CAAGCAGAAGACGGCATACGAGAT-CGGCCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcgtcagCGTCTCTCGCT
SAH_181	CAAGCAGAAGACGGCATACGAGAT-AGCGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcgtcagCGTCTCTCGCT
SAH_182	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-NNNNNcacacaaggagatataccatgtcc
SAH_183	CAAGCAGAAGACGGCATACGAGAT-GATGCTGATGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTatggtgatggtgaccgct

Table 2.3: PCR primers used to add barcodes and priming sequences for Illumina sequencing. SAH_162: forward primer for sequencing pSAH031. Consecutive degenerate bases increase diversity to facilitate cluster identification during sequencing. SAH_163, SAH_164, SAH_178, SAH_179, SAH_180, SAH_181: reverse primers for sequencing pSAH031, encoding different indices for identifying Rounds 1-5. SAH_182: forward primer for sequencing pTKEI-Dest. SAH_183: reverse primer for sequencing pTKEI-Dest.

	Negative Control	Round 1	Round 5
Sequencing reads after quality filtering (fold coverage)	458,749	566,418 (286)	479,688 (242)
Number of transformants		$> 5.0 * 10^7$	$> 1.0 * 10^7$
Number of mutated codons		110	110
Nonsynonymous mutation library size		1870	1870
Percent frameshifted reads	0.3	2	4.5
Percent non-frameshifted reads with:			
0 nonsynonymous mutations	94.1	60.7	26.3
1 nonsynonymous mutation	5.7	28.8	32.6
2 nonsynonymous mutations	0.2	8.3	24.5
3 nonsynonymous mutations	0	1.8	11.5
4+ nonsynonymous mutations	0	0.4	4.9
Mismatch bias: among reads with 1 nonsynonymous mutation, percentage of mutant codons with:			
1-bp substitution	99.1	47.1	27.6
2-bp substitution	0.8	30.8	42.1
3-bp substitution	0.1	22.2	30.4
Percentage of possible codon substitutions observed			
1-bp substitution		100.0	100.0
2-bp substitution		100.0	100.0
3-bp substitution		99.9	100.0
All substitutions		99.9	100.0
Observed coverage of possible single amino acid substitutions		99.8	99.4
Observed coverage of possible single mutant codons		99.7	99.1
Coverage of possible single amino acid substitutions with ≥5 reads		97.2	95.6
Coverage of possible single mutant codons with ≥5 reads		94.1	92.2

Table 2.4: Plasmid recombineering library coverage statistics.

iLOV	Nucleotide Sequence	ATCGAAAAAAACTTCGTTATCACCGACCCGCGCTTACCTGACAATCCAATCATCTT TGCCTCAGATGGTTTTCTCGAGCTAACCGAATACAGCCGTGAAGAAATTCTGGGT CGTAATGCTCGTTTCCTGCAGGGCCCGGAAACCGACCAAGCTACGGTTCAGAAAA TTCGCGATGCAATTCGCGATCAGCGTGAGACTACCGTTCAATTGATCAACTACAC CAAAAGCGGTAAAAAATTTTGGAACCTGTTGCATCTTCAACCGGTACGCGACCAA AAAGGCGAGCTTCAGTACTTTATCGGCGTGCAGCTGGATGGGAGTGACCATGTG
	Amino Acid Sequence	IEKNFVITDPRLPDNPIIFASDGFLELTEYSREEILGRNARFLQGPETDQATVQKIRDAIR DQRETTVQLINYTKSGKKFWNLLHLQPVRDQKGELQYFIGVQLDGSDHV
tLOV	Nucleotide Sequence	ATCGAAAAAAACTTCGTTATCACCGACCCGaGCTTACCTGACAATCCAATCATCTT TGCCTCAGATGGTTTTCTCGAGCTAACCGAATACAGCCGTGAAGAAgTTCTGGGTC GTAATGCTCGTTTCCTGCAGGGCCCGGAAACCGACCAAGCTACGGTTCAGAAAAT TCGCGATGCAATTCGCGATCAGCGTGAGACTACCGTTCAATTGATCAACTACACC AAAAGCGGTAAAAAATTTTGGAACCTGTTGCATCTTCAACCGGTAtttGACCAAAAA GGCGAGCTTCAGTACTTTATCGGCGTGCAGCTGGATGttAGTGACCATGTG
	Amino Acid Sequence	IEKNFVITDPSLPDNPIIFASDGFLELTEYSREEVLGRNARFLQGPETDQATVQKIRDAIR DQRETTVQLINYTKSGKKFWNLLHLQPVFDQKGELQYFIGVQLDVSDHV

Table 2.5: Nucleotide and amino acid sequences of iLOV and tLOV used in this study. The four amino acid mutations in tLOV are R11S, I35V, R90F, G106V.

Chapter 3

A comprehensive deletion landscape of CRISPR-Cas9 identifies the minimal RNA guided DNA binding module

Abstract

Understanding the relationship between protein sequence and function is central to investigating biochemical mechanisms as well as engineering desired improvements. Emerging techniques have harnessed massive libraries of protein mutations to accelerate this process, typically through amino acid substitution. In contrast, while rationally constructed protein deletions have long been essential to elucidating biochemical properties, current techniques are insufficient for a comprehensive approach. Here we develop a method for constructing functional landscapes of even the largest and most complex proteins, comprehensively surveying functional deletions in the RNA-guided DNA binding protein dCas9, the foundation for powerful genome editing and modifying technologies. CRISPR proteins are highly complex with numerous distinct domains responsible for activities such as guide RNA binding, DNA recognition, DNA unwinding, specificity sensing and ultimately the cleavage of each DNA strand. We exploit the functional landscape to revert functionality and step backward in domain evolution, comprehensively minimizing dCas9 and screening for an essential function. We demonstrate the power of this technique by revealing the minimal RNA guided DNA binding module at 64% of the full CRISPR-Cas9 platform, providing many new opportunities for fusions and delivery. This exploration also uncovers evidence that the Helical II domain promotes DNA binding. These results highlight the power of comprehensive protein deletions to clearly elucidate the boundaries of a central function.

3.1 Introduction

Proteins evolve through the rearrangement of modular subunits known as domains, and most domain architectures gain complexity during evolution (Fong et al. 2007). Previously, we and others have exploited the modularity of domains to rearrange or expand the architecture of a protein, enabling new functionality. For example, the programmable DNA nuclease Cas9 can be converted into a ligand-dependent allosteric switch using advanced molecular cloning, similar to other domain insertions dictated by allostery (Oakes et al. 2016; Reynolds, McLaughlin, and Ranganathan 2011). Advanced methods now enable the construction of comprehensive protein libraries for domain insertion (Oakes et al. 2016) and domain rearrangement/circular permutation (Atkinson et al. 2018), however no such method exists for domain deletion. Domain deletions are also underrepresented in natural evolution, and when they occur are often limited to protein termini as a result of alternative start or stop codons (Björklund et al. 2005; Weiner, Beaussart, and Bornberg-Bauer 2006). Crucially, eukaryotic proteome diversity is vastly increased by alternative splicing, which tends to insert or delete protein domains (Kriventseva et al. 2003). This additional diversity is fully exploited by eukaryotic evolution, such that ~95% of multi-exon human genes are alternatively spliced (Pan et al. 2008).

Methods have been developed to mimic this natural diversity, but approaches to date have been limited in size and scope. Rationally constructed protein deletions have long been essential to elucidating functional and biochemical properties but are generally limited to a handful of truncations. Moreover, protein engineering can make use of deletions to alter enzyme substrate specificity (Simm et al. 2007), enable screens for improved activity and thermostability (Hecky and Müller 2005), or minimize protein size (D. Ma et al. 2018). Early approaches to protein deletion libraries resulted in the deletion of single amino acids using an engineered transposon, MuDel (Jones 2005; Arpino et al. 2014). Various other methods utilize direct PCR (Pisarchik, Petri, and Schmidt-Dannert 2007), random nuclease digestion (Ostermeier, Shim, and Benkovic 1999), or random in vitro transposition followed by a complicated cloning scheme (Morelli et al. 2017) to achieve deletion libraries containing a variety of lengths. These techniques are low in throughput and/or require complex molecular techniques, and in contrast to protein insertions or circular permutations where library size grows linearly with target length, deletion libraries grow as the square. Thus, to date, it has not been possible to comprehensively explore deletions in a protein of average size, e.g. 361 amino acids in eukaryotes (Brocchieri and Karlin 2005). Furthermore, a highly efficient technique could survey entire domain deletions in much larger multi domain proteins.

A simple and efficient method for building protein deletions coupled with a screen or selection would provide the ability to comprehensively query and delineate the function of domains or motifs in complex and multi-domain proteins. Such a technique could be used to identify crucial functions of complex proteins and splice variants in a manner akin to how deep mutational scanning can be used to identify the effects of single nucleotide polymorphisms on functionality (Fowler and Fields 2014; Araya and Fowler 2011). Moreover, the iterative accretion of deletions could be viewed as analogous to deconstructing the evolutionary aggregation of domains - rolling the clock back in time to an essential protein unit. Here, we introduce Minimization by Iterative Size Exclusion & Recombination (MISER). We apply MISER to the 1368 amino acid multi-domain DNA binding protein dCas9, and comprehensively assay deletions to identify the minimal RNA guided DNA binding scaffold. Furthermore, we construct synthetic versions of this scaffold by combining libraries of deletions and identify new

CRISPR Effector (CE) proteins, less than 1000 amino acids in length, that represent the minimal RNA guided DNA binding module within Cas9.

3.2 Materials and Methods

3.2a Molecular Biology

All restriction enzymes were ordered from New England Biolabs (NEB). Polymerase Chain Reaction (PCR) was performed using Q5 High-Fidelity DNA Polymerase from NEB. Ligation was performed using T4 DNA Ligase from NEB. Agarose gel extraction was performed using the Zymoclean Gel DNA Recovery kit, and PCR clean-up was performed using the DNA Clean & Concentrator, both from Zymo Research. Plasmids were isolated using the QIAprep Spin Miniprep Kit (Qiagen). All DNA-modifying procedures were performed according to the manufacturers' instructions. All molecular biology was performed in Xl1-Blue (UC Berkeley QB3 MacroLab) and standard LB (Teknova) unless otherwise stated.

3.2b MISER library construction

Two sets of 1368 oligonucleotides were designed and ordered as Oligonucleotide Library Synthesis (OLS) from Agilent Technologies. Oligonucleotides were designed to insert a six base pair (bp) recognition sequence for either the restriction enzyme NheI or SpeI between every codon in dCas9, beginning after the start codon and ending before the stop codon. Examples of the first oligonucleotide for both NheI and SpeI insertion are located in Table 3.1, including annotated design features. Internal priming sites were included in order to amplify NheI or SpeI specific oligonucleotide libraries. A modified amplification procedure was performed. In a 50 µL PCR reaction, 10 ng of template oligonucleotide library was amplified according to manufacturer's instructions, but with an extension time of only five seconds, and a total of only 15 cycles. 1.5% dimethyl sulfoxide (DMSO) was also included in the PCR reaction. These modifications were empirically determined in order to minimize undesirable higher order PCR products that were observed to be produced by amplification. These side products are likely the result of complementary oligonucleotides priming one another. Notably this phenomenon is likely inherent to amplification of a library of DNA tiled across a common sequence – in this case dCas9. PCR primers can be found in Auxiliary Supplementary Materials – Primer Sequences. 24 such reactions were typically performed in parallel and then combined, followed by concentration with with Zymo 'DNA Clean & Concentrator'. BsmbI restriction digestion was then used to remove priming ends, followed by a second concentration with Zymo 'DNA Clean & Concentrator', resulting in mature double-stranded recombineering-competent DNA.

Plasmid recombineering was performed as described (Higgins 2017), using strain EcNR2 (Addgene ID: 26931) to generate MISER libraries in plasmid pSAH060. Plasmid sequences can be found in Auxiliary Supplementary Materials – Plasmid Sequences. Briefly, mature double-stranded recombineering-competent DNA at a final volume of 50 µL of 1 µM, plus 10 ng of pSAH060, was electroporated into 1 mL of induced and washed EcNR2 using a 1 mm electroporation cuvette (BioRad GenePulser). A Harvard Apparatus ECM 630 Electroporation System was used with settings 1800 kV, 200 Ω, 25 µF. Three replicate electroporations were performed, then individually allowed to recover at 30° C for 1 hr in 1 mL of SOC (Teknova) without antibiotic. LB (Teknova) and kanamycin (Fisher) at 60 µg/mL was then added to 6 mL

final volume and grown overnight. A sample of recovered culture was diluted and plated on kanamycin to estimate the total number of transformants, typically >107. Cultures were miniprepped and combined the next day.

Plasmid recombineering is relatively inefficient, and only a fraction of recovered plasmids contained successful NheI or SpeI insertions. In order to recover completely penetrant libraries, an intermediate cloning step was performed. A PCR product (PCR primers can be found in Table 3.4) conferring resistance to chloramphenicol (Table 3.5) was cloned into both libraries of pSAH060 plasmids. This PCR product contained either flanking NheI restriction sites or SpeI restriction sites, such that only modified pSAH060 plasmids (possessing NheI or SpeI restriction sites) could obtain chloramphenicol resistance through NheI/SpeI digestion and subsequent ligation. Libraries were then cleaned and transformed into Xl1-Blue competent cells for overnight selection in chloramphenicol (Amresco) at 25 µg/mL followed by plasmid isolation the next day. Samples of recovered cultures were also plated on both kanamycin alone (native pSAH060 resistance) and chloramphenicol alone (resistance mediated by successful recombineering insertion) to estimate the fraction of modified plasmids and therefore the restriction library size. Recombineering efficiencies were observed at ~0.5% by this method, indicating restriction library sizes of ~50,000, well above the number of unique insertion sites per library (1,368). Finally, chloramphenicol resistant pSAH060 libraries were digested with either NheI or SpeI as appropriate, removing the chloramphenicol cassette. The libraries were run on an agarose gel, and the 5953 bp (5947 bp pSAH060 + 6 bp inserted restriction site) linear band corresponding to each library was gel extracted.

To construct deletion variants composed of N- and C- terminal dCas9 fragments, one µg of each library was mixed and digested with BsaI, then Zymo cleaned. The resulting DNA mixture contained equimolar free dCas9 N- and C- terminal fragments, as well as equimolar pSAH060 vector backbone. This mixture was then ligated in the presence of SpeI and NheI, 'locking' dCas9 fragments together by one of two six bp scar sites not recognized by either enzyme (Figure S1B). The ligated MISER library was transformed into XL1-Blue, grown overnight and plasmids were isolated the next day. The MISER library of dCas9 is quite large, with 936,396 possible deletions ($N(N + 1) / 2$, $N = 1368$), and all cloning steps were performed with validation that $> 10^7$ transformants were obtained.

The MISER library is theoretically composed of all possible N- and C- terminal fragments, including duplications as well as deletions. To isolate deletions in a particular size range the MISER library was digested with BsaI, to excise the dCas9 gene from the vector backbone, and run on an agarose gel. Various slices of the MISER library were individually gel extracted (Figure 3.1B), ligated into expression vector pSAH063, and transformed into E.coli strain Xl1-Blue with appropriate antibiotic selection.

3.2c Fluorescence repression assays and flow cytometry

The catalytically dead dCas9 MISER variants were used to repress the transcription of genomically encoded fluorescent reporter genes in E.coli as previously described (Oakes et al., 2014; Qi et al., 2013). Briefly, a sgRNA targeting Green Fluorescent Protein (GFP) was constitutively expressed, which results in repression of constitutively expressed GFP contingent on functional dCas9 expression from pSAH063. This repression was quantified relative to non-targeted Red Fluorescent Protein (RFP), which was expressed identically to GFP and located downstream at the same genomic locus. This assay yields robust repression detection, with at

least an order of magnitude lower GFP signal after 8 hours of growth at 37° C with 750 rpm shaking in LB media + 1 nM Isopropyl β-D-1-thiogalactopyranoside (IPTG) induction of dCas9 from pSAH063. Assays and flow cytometry were conducted in either an Infinity M1000 PRO monochromator (Tecan) or an SH800 Cell Sorter (Sony Biotechology). For GFP/RFP ratiometric measurements there was no significant difference between samples for the RFP fluorescence measurement.

3.2d Library Sequencing and Analysis

100 nucleotide single end reads were used to sequence the dCas9 Slice 4 and Slice 5 libraries. dCas9 open reading frames were amplified from pSAH064 libraries with primers SAH_356 and SAH_358. PCR products were further prepared for deep sequencing by the UC Berkeley Functional Genomics Laboratory. Sequencing was performed by the UC Berkeley Vincent J. Coates Genomics Sequencing Laboratory on an Illumina HiSeq4000. Samples were mixed at custom ratios as follows: Slice 5 Naïve Library – 10% ; Slice 5 Sorted Library – 10% ; Slice 4 Naïve Library – 40% ; Slice 4 Sorted Library – 40%.

Sequencing analysis was performed with custom MATLAB scripts available online at https://github.com/savagelab. Briefly, reads were analyzed for the novel presence of the two possible MISER scar sequences, 'GCTAGT' or 'ACTAGC'. The majority of reads were fully WT dCas9 sequence, as expected due to the fact that scar sequences can occur anywhere along dCas9. Once detected, reads containing 15 bp upstream and downstream of the scar (that exactly matched dCas9 sequence) were used to identify the location of a deletion. Sequencing statistics can be found in Table 3.2. Enrichment ratios were calculated by taking the ratio of the frequency of each variant before and after selection (Fowler 2014). To conservatively display variants only detected in one library, one artificial read was added to both datasets.

3.3 Results

MISER enables a comprehensive query of a protein deletions by 1. programmably encoding two distinct restriction enzyme sites at every codon in a target protein; 2. excising the intervening sequence using said restriction enzymes and; 3. re-ligating the remaining fragments in a cycling ligation reaction that drives towards completion (Figure 3.1A). The simple and efficient MISER system can utilize any two restriction enzymes with compatible sticky ends, here SpeI and NheI. Subsequent cleavage and ligation forms a two codon scar site not recognized by either enzyme, thereby greatly increasing efficiency and simplifying cloning (Figure 3.2). The creation of MISER libraries is simple and fully programmable-- using in vivo plasmid recombineering (Thomason et al. 2007; Higgins, Ouonkap, and Savage 2017) to generate the initial NheI and SpeI libraries, thus overcome the optimization requirements and undirected nature of transposon and nuclease-based methods. Fundamentally, the ligation of protein terminal fragments produces duplications as well as deletions, such that a MISER library is a triangular distribution, with near-WT length proteins most frequent and the largest deletions least frequent (Figure 3.1B). To empirically determine the size range of functional deletions, an agarose gel of the dCas9 MISER deletion library was sliced into six sub-libraries, independently cloned into expression vectors, and assayed for CRISPRi GFP repression via flow cytometry (Figure 3.3) (Qi et al. 2013). Sublibrary Slice 4 was the most stringent library with detectable

repression (Figure 3.1B), with functional variants becoming more frequent in slices composed of smaller deletions as expected.

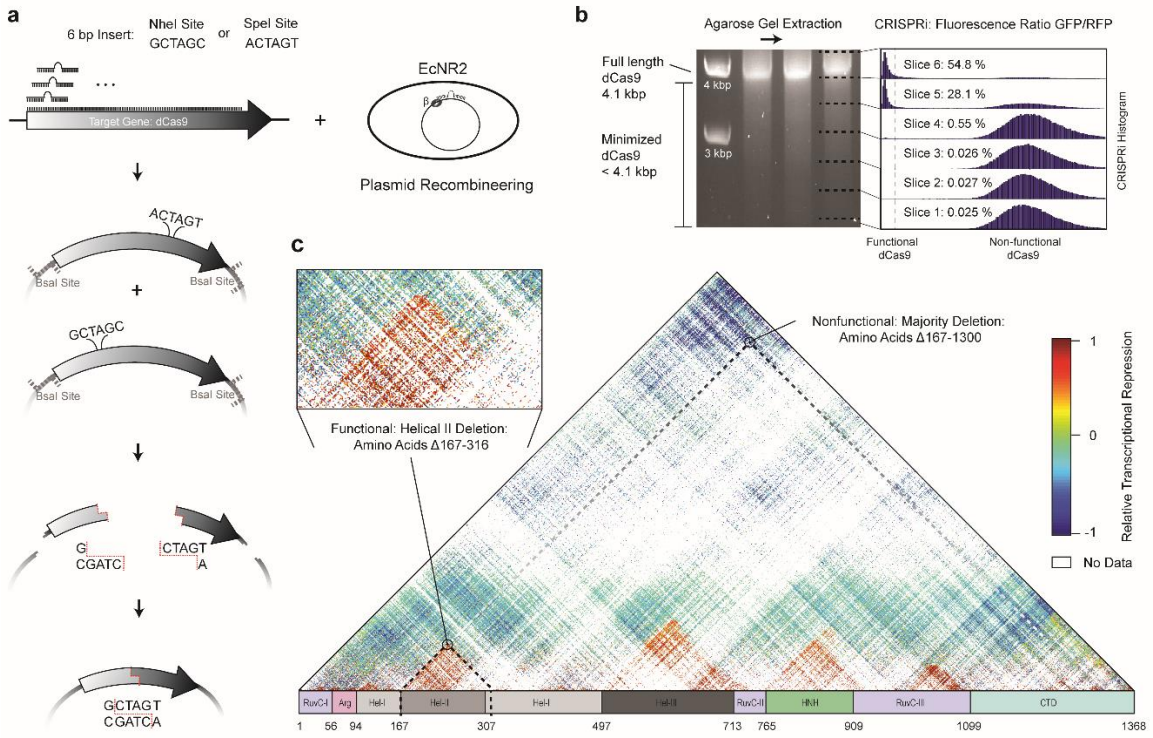
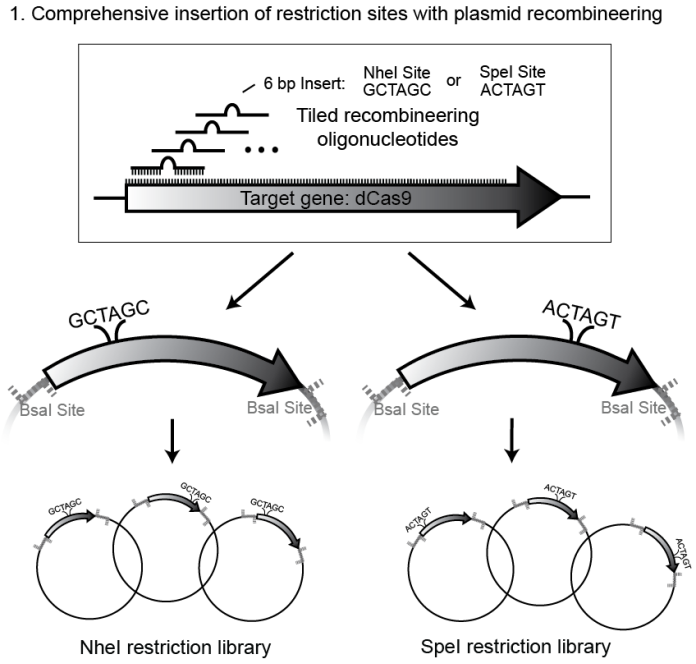


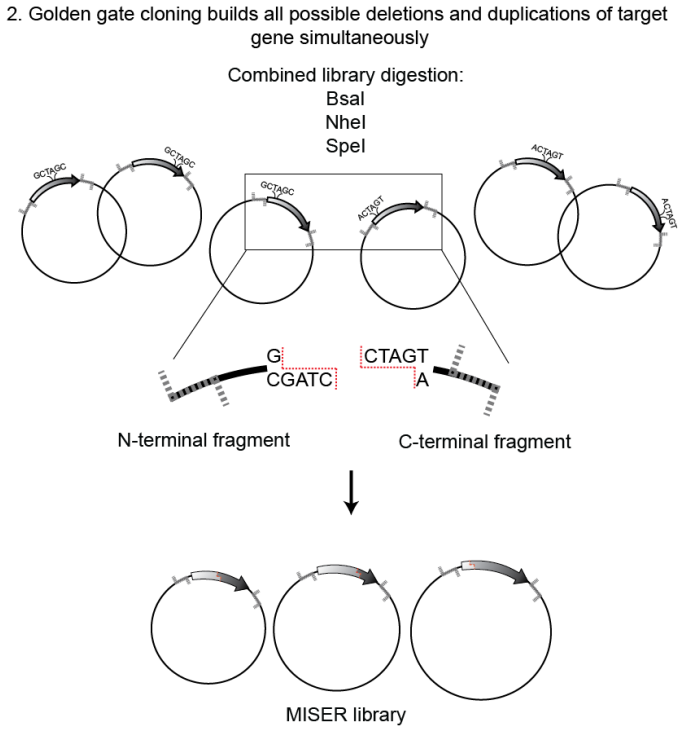
Figure 3.1: MISER produces comprehensive functional landscapes of protein deletions. (A) Cartoon of the MISER cloning scheme. Two unique restriction sites with compatible sticky ends are programmably inserted across a target gene using plasmid recombineering. Digestion and re-ligation of gene fragments produces all possible deletions. (B) The full range of dCas9 deletion sizes was separated into six differently sized sub-libraries and assayed for CRISPRi function. Slices six, five, and four were found to contain functional deletion variants. (C) Slices five and four were deep sequenced before and after a CRISPRi fluorescence assay to generate a value for relative transcriptional repression. This data is presented as a deletion landscape, where each pixel represents a particular deletion variant. Two deletion variants are explicitly annotated, one functional and one non-functional. Inset: zoom of the Helical II domain region.

Fluorescence-activated Cell Sorting (FACS) and deep sequencing of MISER variants enabled the comprehensive mapping of functional dCas9 deletions. To focus sequencing on functional variants, Slice 4 and Slice 5 were deep sequenced pre- and post- FACS sorting, and the enrichment or depletion of individual variants was quantified. Four large deletion regions were independently identified in both libraries, and the data in the libraries were highly significantly correlated (Figure 3.4). These data were normalized and combined to generate a comprehensive landscape of functional dCas9 deletions (Figure 3.1C). 80% of sequencing depth was focused on deletions from 150 to 350 amino acids in length (Slice 4), and 51.4% (115,530/224,718) of these deletions were detected. Overall this landscape includes data for 27.5% of all possible dCas9 deletions (257,737/936,396). Four large deletion regions were identified, roughly corresponding to the Helical II, Helical III, HNH, and RuvC III domains (Figure 3.5). These large deletion regions are bounded by domain topology, as observed by

a


1. Comprehensive insertion of restriction sites with plasmid recombineering
6 bp Insert:
NheI Site GCTAGC
or
SpeI Site ACTAGT
Tiled recombineering oligonucleotides
Target gene: dCas9
GCTAGC
ACTAGT
BsaI Site
BsaI Site
BsaI Site
BsaI Site
NheI restriction library
SpeI restriction library

b


2. Golden gate cloning builds all possible deletions and duplications of target gene simultaneously
Combined library digestion:
BsaI
NheI
SpeI
G
CGATC
CTAGT
A
N-terminal fragment
C-terminal fragment
MISER library

c

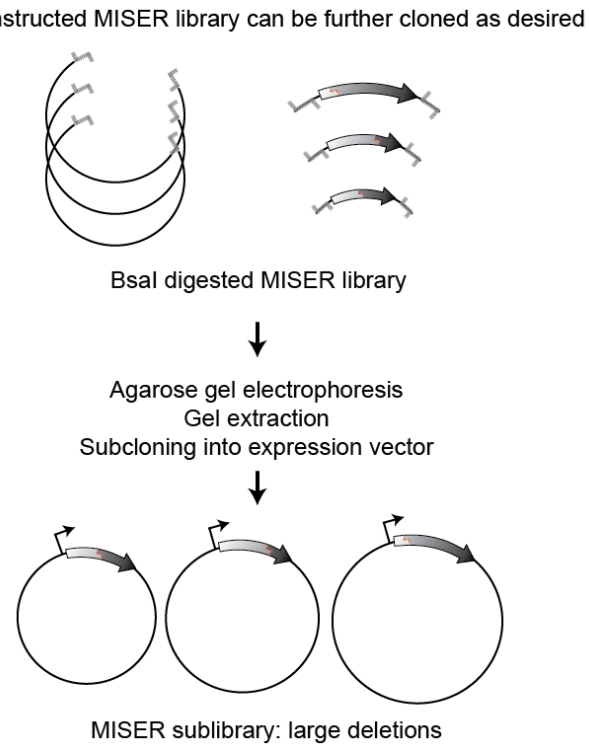
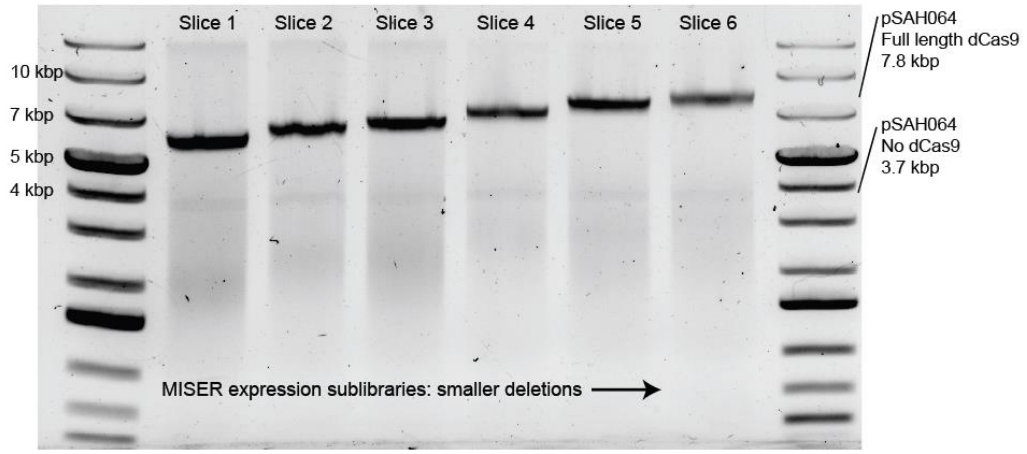


Figure 3.2: Full cloning scheme for Multiplex Iterative Size Exclusion Recombination (MISER). The method can be considered in three parts. (A) Plasmid recombineering generates two comprehensive libraries of restriction site insertions across the target gene. These restriction sites are both novel to the target plasmid and produce compatible sticky ends. Recombineering was performed similarly as in (Higgins 2017), where the target gene lacks a promotor and start codon to prevent growth biases during library construction, and is flanked by BsaI sites for later golden gate cloning (here, plasmid pSAH060) . Additionally, rather than mutagenic oligos, double stranded PCR product was used for recombineering, and another cloning step was introduced to remove unmodified plasmids. These modifications are described in Experimental Design. (B) Modified golden gate cloning generates a library of ligated N- and C- terminal fragments of the target gene, comprehensively producing protein deletion variants as well as duplication variants. An equimolar mixture of the two plasmid libraries is mixed and fully digested to produce free N- and C- terminal fragments of the target gene. This fragment mixture is then re-ligated in the presence of NheI and SpeI. Successful ligation of an N- and C-terminal fragment from differing libraries produces one of two possible 6 base-pair scar sequences. These novel scar sequences are not recognized by either NheI or SpeI, thus trapping the desired chimeric product as a final ligated vector. Because N- and C-terminal fragments are ligated randomly, these chimeric products produce both protein deletions and protein duplications. Ideally the library is both large enough and minimally biased in order to produce a large fraction of possible variants. The product of this step can be considered a MISER library of plasmid pSAH060. (C) A final cloning step moves the MISER library into a desired context – i.e. an expression plasmid, here pSAH063. Step C also allows for size-based exclusion of undesired protein variants by extraction from an agarose gel.

a



b

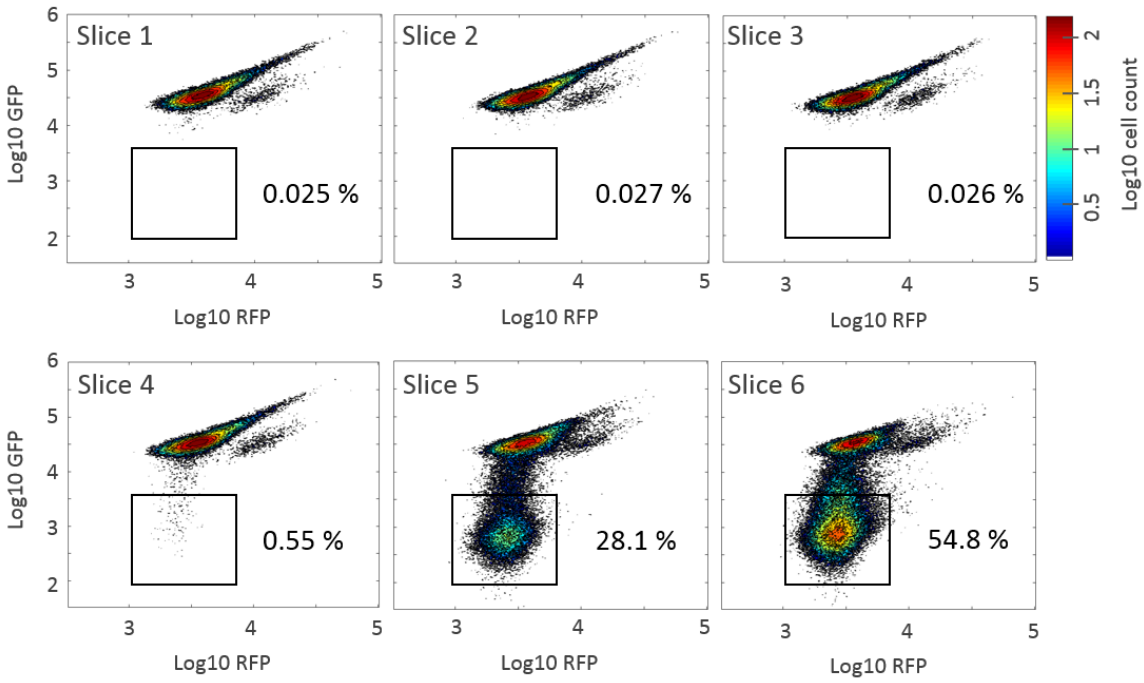
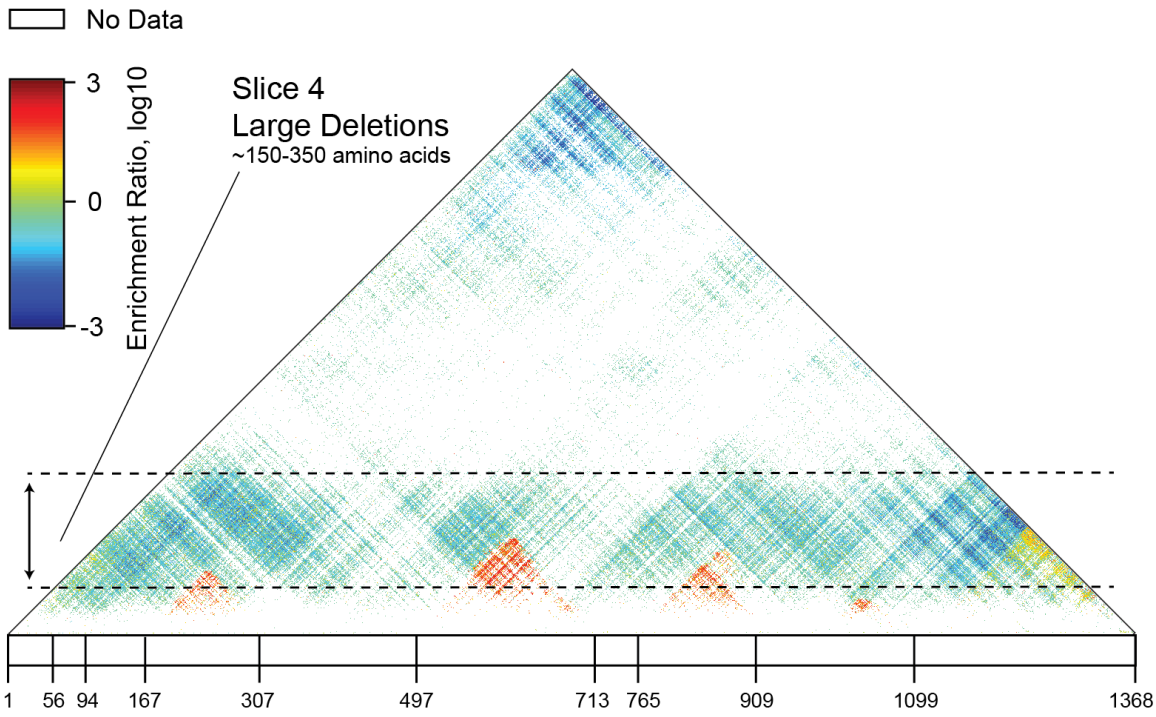
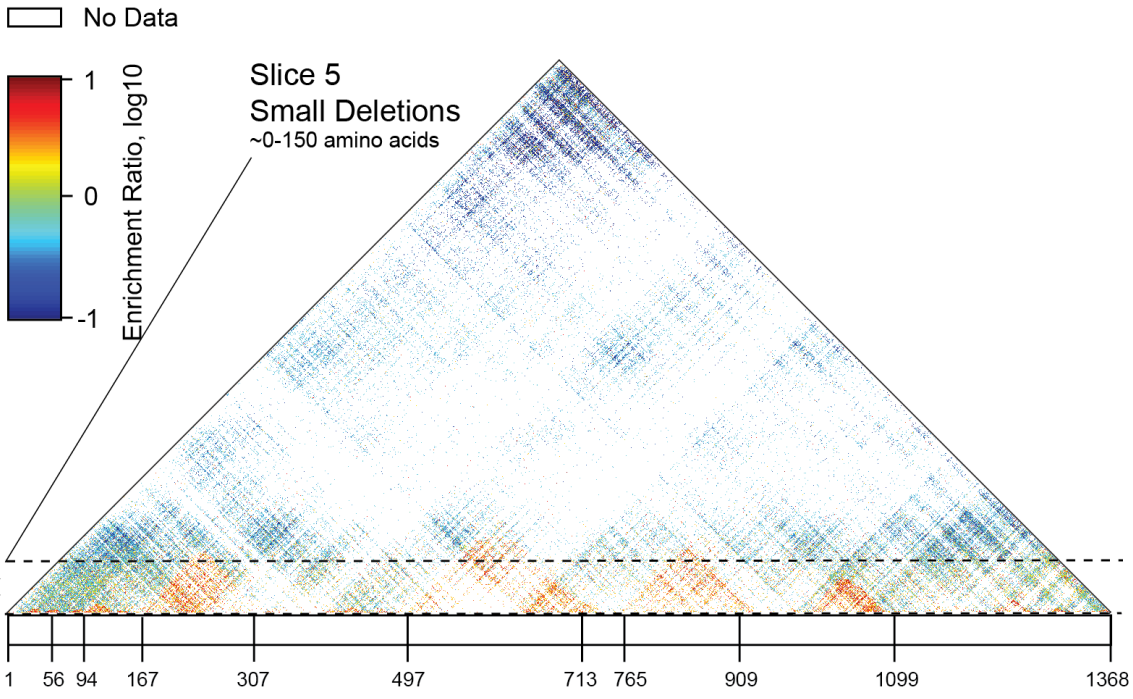


Figure 3.3: Size exclusion and flow cytometry identify the range of dCas9 deletion sizes exhibiting in vivo transcriptional repression. (A) Individual MISER sub-libraries of specific lengths can be generated by agarose size exclusion. To identify the range of functional deletion sizes, the dCas9 MISER library was size separated by agarose gel electrophoresis. The gel range encompassing dCas9 deletions was cut into six gel slices, each containing a sub-library of differently sized deletion variants. These six slices were individually gel extracted and ligated into expression vector pSAH063, generating pSAH064 plasmids with dCas9 deletions. The resulting expression sub-libraries exhibit high precision in size ranges when assayed by agarose gel electrophoresis. (B) Flow cytometry identifies Slice 4, 5, and 6 as expression sub-libraries containing functional dCas9 deletion variants. GFP repression CRISPRi was performed as described in Experimental Design. The region of phenotype defined as 'functional' is illustrated. The percent of functional hits is annotated.

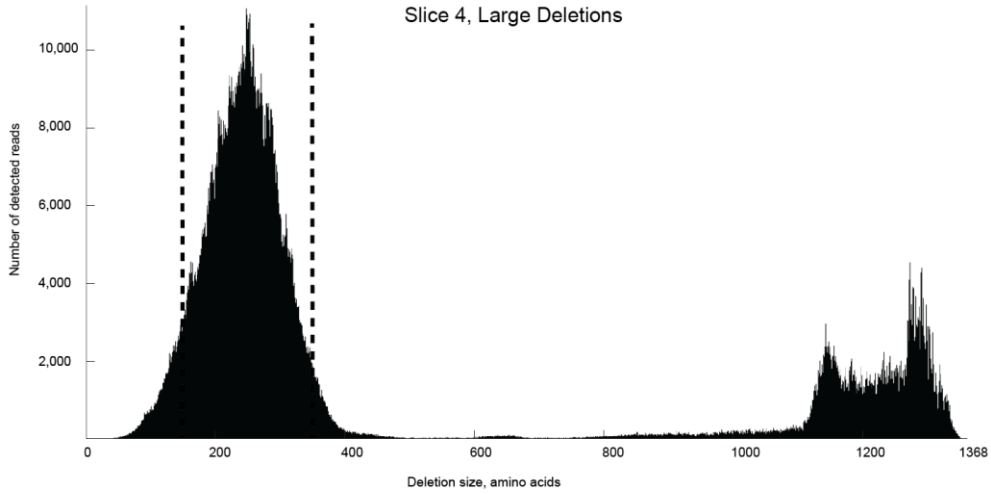
a


No Data
3
0
-3
Enrichment Ratio, log10
Slice 4
Large Deletions
~150-350 amino acids
1
56
94
167
307
497
713
765
909
1099
1368

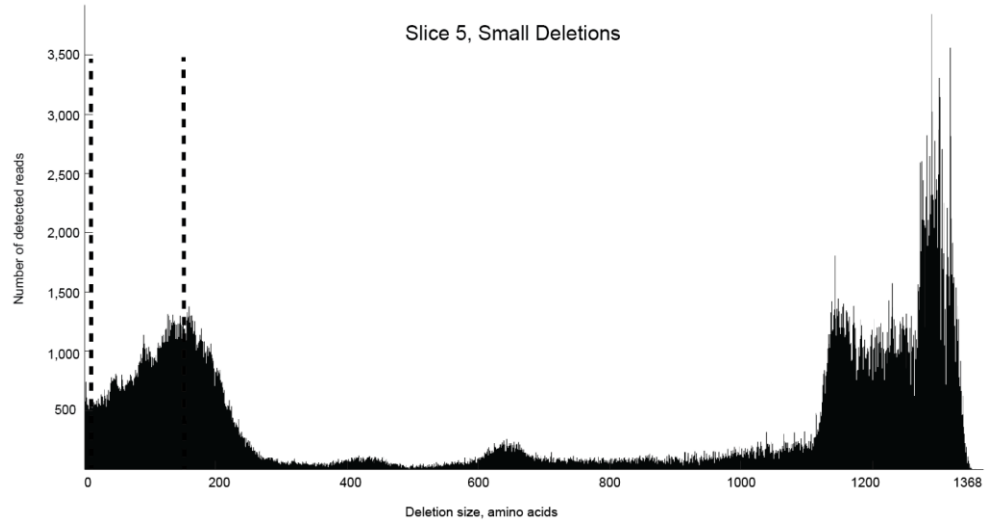
b


No Data
1
0
-1
Enrichment Ratio, log10
Slice 5
Small Deletions
~0-150 amino acids
1
56
94
167
307
497
713
765
909
1099
1368

c


Slice 4, Large Deletions
Number of detected reads
10,000
8,000
6,000
4,000
2,000
0
200
400
600
800
1000
1200
1368
Deletion size, amino acids

d


Slice 5, Small Deletions
Number of detected reads
3,500
3,000
2,500
2,000
1,500
1,000
500
0
200
400
600
800
1000
1200
1368
Deletion size, amino acids

e

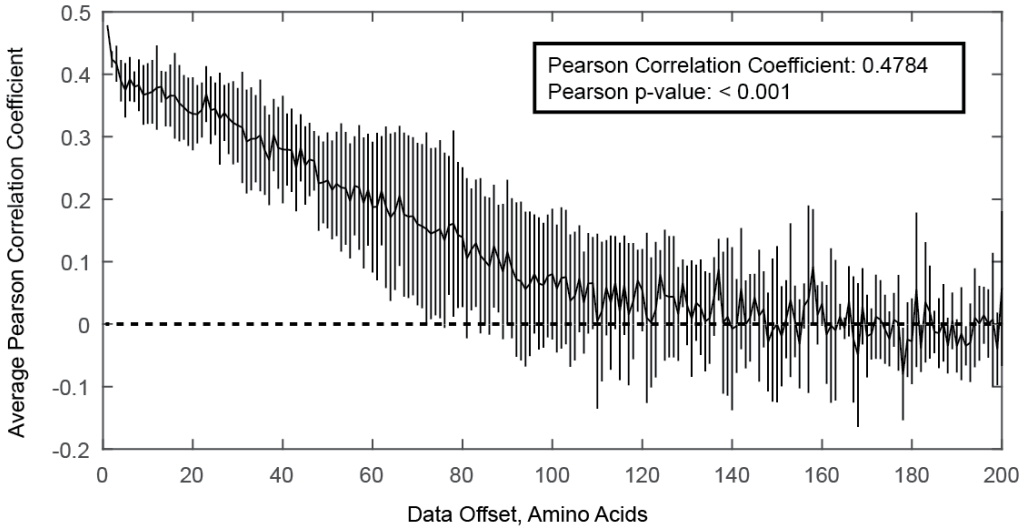

Pearson Correlation Coefficient: 0.4784
Pearson p-value: < 0.001
Average Pearson Correlation Coefficient
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
0
20
40
60
80
100
120
140
160
180
200
Data Offset, Amino Acids

Figure 3.4: Deep sequencing of the sublibraries Slice 4 and Slice 5 reveal deletion regions throughout dCas9. (A) Raw enrichment map of Slice 4 sub-library. Each pixel represents a particular deletion variant, whose start and end points are mapped to the linear amino acid sequence by lines at 45° angles, for example the dashed lines in the Helical-II domain. Domain boundaries are labeled by amino acid number. The pixel color denotes the degree of enrichment or loss following flow cytometry screening for transcriptional repression in vivo. Detailed calculations are described in the supplementary methods. Deletions corresponding to sizes within the gel slice are indicated by dashed lines. (B) Raw enrichment map of Slice 5 sub-library, as in A. Note the differing range of enrichment ratios. (C) Histogram of deletion sizes in the naïve Slice 4 library. The edges of the gel slice are indicated by dashed lines. (D) Histogram of deletion sizes in the naïve Slice 5 library. The edges of the gel slice are indicated by dashed lines. (E) Slices 4 and 5 independently replicate the same large functional deletion regions. The raw enrichment maps of Slice 4 and Slice 5 contain many of the same variants, and the Pearson correlation for these variants is highly significant ($p < 0.001$). Furthermore, this correlation is progressively lost if the two enrichment maps are shifted relative to one another. The line plots the mean of four additional Pearson correlations where the data array has been offset – either up, down, left, or right – by the indicated number of amino acids. This analysis verifies that the two enrichment maps independently identify large-scale regions of dCas9 which can be deleted, and validates the apparent visual correspondence between maps A and B. Error bars, standard deviation.

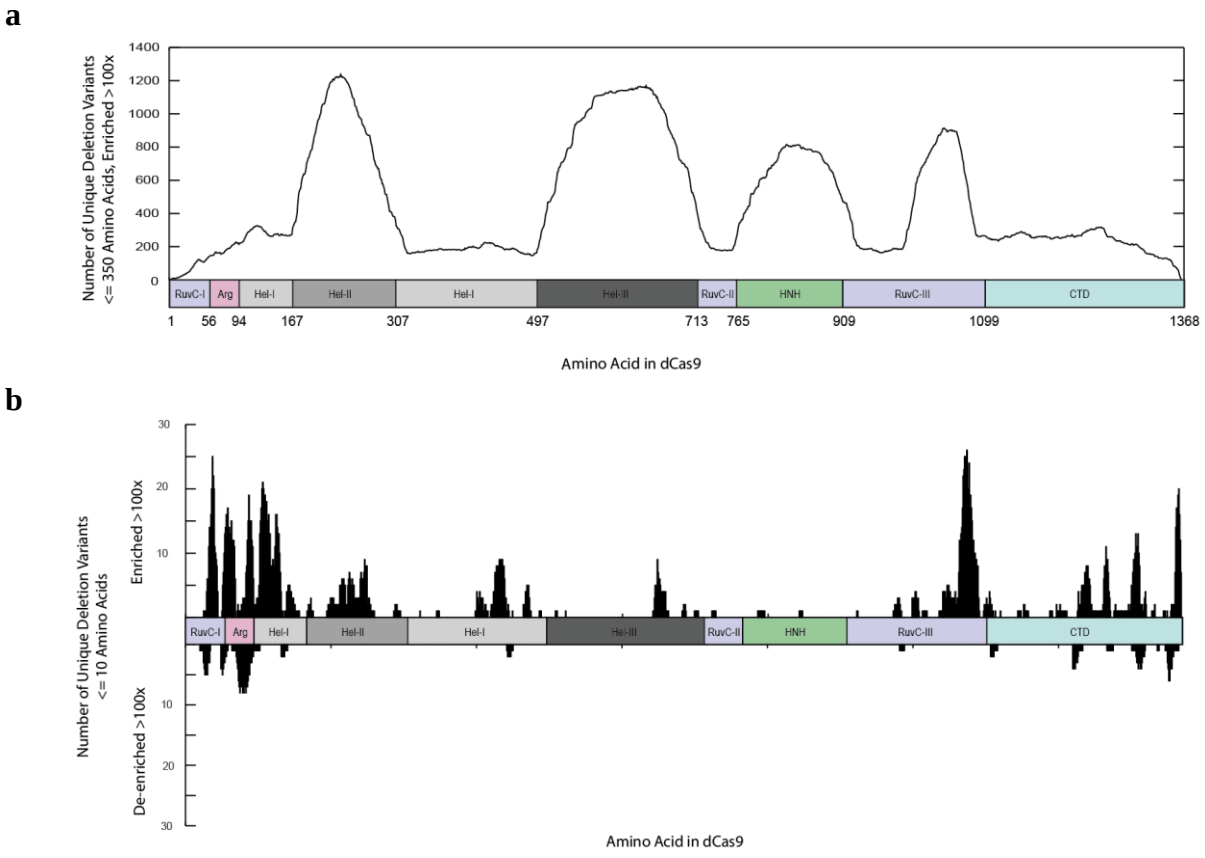


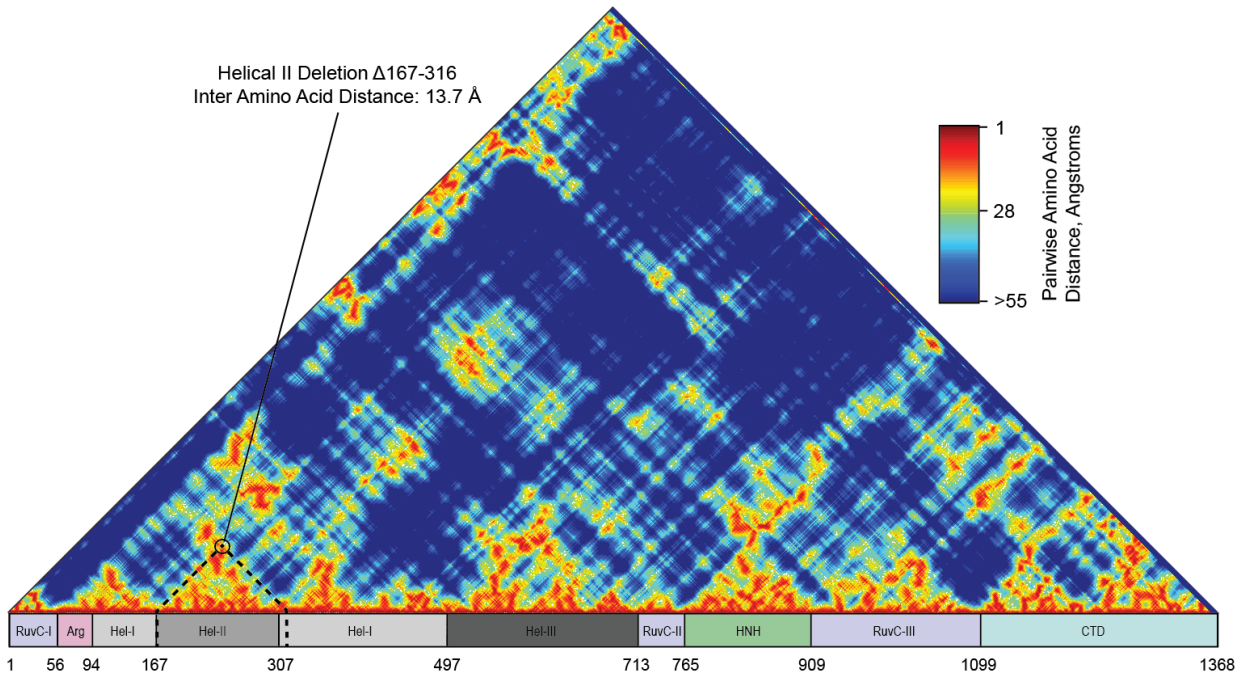
Figure 3.5: Four domains can be deleted from dCas9, and many shorter deletions are tolerated. (A) Large tolerated deletions across dCas9. Only highly enriched deletions are plotted (>100x). Deletions range from 1-350 amino acids. (B) Small functional and non-functional deletions across dCas9. Only highly enriched (>100x) or depleted (<0.01x) deletions are plotted. Deletions range from 1-10 amino acids.

correlations between functional enrichment and the three dimensional distance between deletion sites in a DNA-bound structure of Cas9 (Figure 3.6). Notably, both the correlation strength and significance are completely lost once within a region corresponding to a deletable domain. While domain-level deletions are strongly bounded, small deletions and small insertions (~10 amino acids) are tolerated in much of the structure (Figure 3.5B, Figure 3.7.), a finding that has been previously observed in other proteins (Simm et al. 2007; Pisarchik, Petri, and Schmidt-Dannert 2007). Two exceptions are the 'bridge helix' (Nishimasu et al. 2014) and the 'phosphate lock loop' (Anders et al. 2014), both known for key mechanistic roles.

To validate the function map, individual deletion variants from each of the four deletion regions were either isolated from the library (Figure 3.8) or constructed via PCR and assayed individually. Variants from each of the four deletion regions could be identified that exhibited CRISPRi nearly as effectively as full length dCas9 (Figure 3.9, Figure 3.8E). Intriguingly, many of the deletions we identified have been explored rationally over a host of reports, yielding insight into the biochemical mechanisms lost with the removal of each domain. To begin, the most obvious of the acceptable deletions, the HNH domain, is responsible for cleaving the target strand and gating cleavage by the RuvC domain. Thus it is little surprise that deletions of the HNH are tolerated in a molecule that is required to bind but not cleave DNA. In fact Sternberg et al. previously demonstrated that a HNH deleted (Δ768-919) Cas9 is competent for nearly WT levels of binding activity, but is unable to cleave (Sternberg et al. 2015). Likewise, Chen et al. previously demonstrated that the Helical III domain plays a similar role, but upstream of the HNH domain, gating the closing motion of HNH cleavage by sensing the extended duplex. Deletion of this domain (Δ497-713) also ablated cleavage activity while maintaining full binding affinity (Chen et al. 2017). Next, the Helical II domain was previously deleted because it was postulated to be unnecessary due to low conservation in other Cas9 sequences, and furthermore lacks contacts to the bound guide:target heteroduplex in the crystal structure (Nishimasu et al. 2014). Notably, unlike our deletions of HNH & Helical III, the most functional deletions of the Helical II domain were slightly but significantly less functional that dCas9 (Figure 3.9A). Finally, we also uncover a deletion set in the RuvCIII domain that has never before been seen or tested. Modeling this deletion on the crystal structure (PDB ID 59FR) reveals that this deletion removes a large set of loops, an alpha helix and two antiparallel beta sheets.

Protein domains are often modularly and progressively added during the evolution of a large protein, and as such we hypothesized that the reverse would also be possible. However, unanticipated epistatic effects between multiple domain deletions might cause any particular set of 'stacked' deletions to become non-functional. Therefore, we generated libraries of stacked deletions built upon RuvC deletion variant Δ1010-1081 as a starting point. A library of quadruple deletion variants, termed CRISPR Effectors (CE) due to their highly pared down structure compared to Cas9, were constructed as follows: individual sublibraries of deletions from Helical II, Helical III, and the HNH domains were isolated from the full MISER library (Figure 3.8). The dCas9 gene was divided into four fragments in order to PCR and combine the deletion sub-libraries via Golden Gate cloning (Figure 3.10). The resulting library, CE Library 1, was processed through the CRISPRi assay and functional variants were isolated by flow cytometry as above. A variety of highly functional CEs were obtained (Figure 3.11), although surprisingly none of them possessed a Helical II deletion. This result was curious and we therefore generated a second library, CE Library 2, in a similar manner, except we included a new sublibrary of isolated Helical II deletion variants to ensure diversity in deletions from this

a


Helical II Deletion Δ167-316
Inter Amino Acid Distance: 13.7 Å
1
28
>55
Pairwise Amino Acid Distance, Angstroms
RuvC-I
Arg
Hel-I
Hel-II
Hel-I
Hel-III
RuvC-II
HNH
RuvC-III
CTD
1
56
94
167
307
497
713
765
909
1099
1368

b

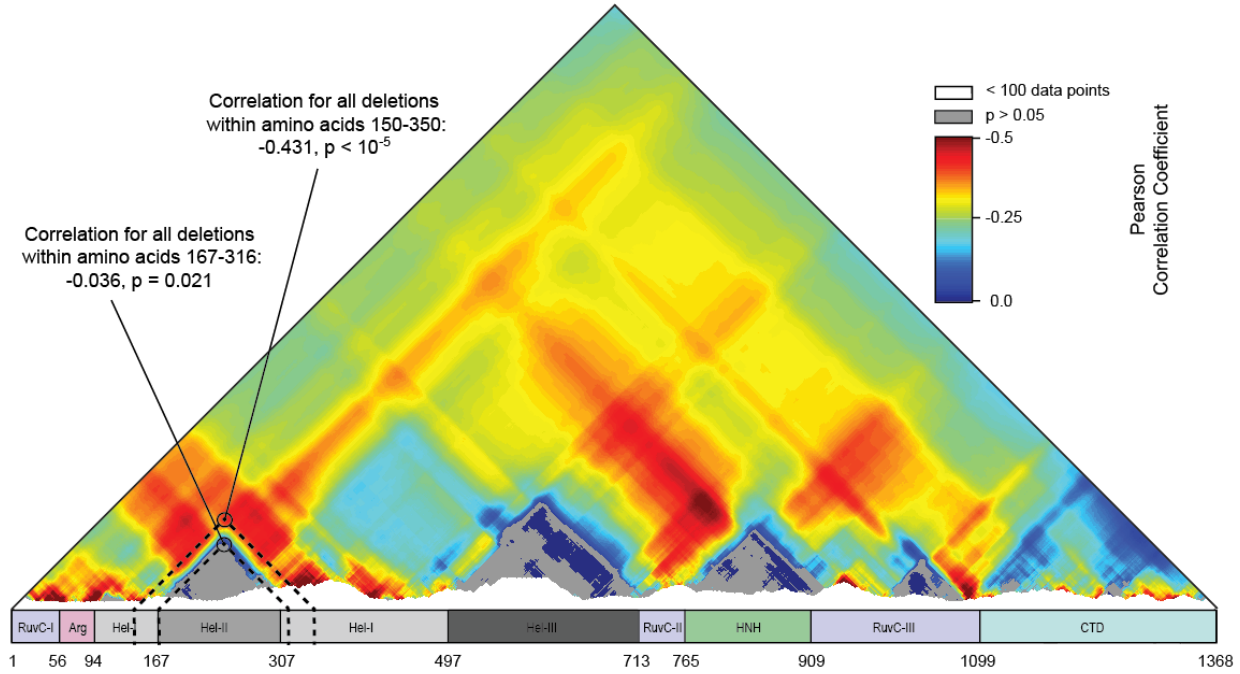

Correlation for all deletions within amino acids 150-350: -0.431, p < 10-5
Correlation for all deletions within amino acids 167-316: -0.036, p = 0.021
< 100 data points
p > 0.05
-0.5
-0.25
0.0
Pearson Correlation Coefficient
RuvC-I
Arg
Hel-I
Hel-II
Hel-I
Hel-III
RuvC-II
HNH
RuvC-III
CTD
1
56
94
167
307
497
713
765
909
1099
1368

Figure 3.6: Domain deletions are constrained by inter- but not intra- domain topology. (A) Pairwise amino acid distances of Cas9 are plotted, ranging between 1 and 55 angstroms. Each pixel represents a pair of amino acids in Cas9, corresponding to the triangular projection onto the Cas9 domain cartoon at bottom. An example Helical-II deletion is illustrated. The pixel color represents the three dimensional linear distance between the two amino acids specified. Distance data is taken from PDB ID: 5F9R. (B) Pearson correlations between pairwise amino acid distance and enrichment values. The Pearson correlation between pairwise distances and enrichment values for the entirety of dCas9 is 0.328, $p < 10^{-5}$. This indicates that overall, only a minority of the deletion variant's enrichment can be explained by distance alone. However, smaller sub-regions within dCas9 exhibit either increased or decreased correlations. All possible such sub-region correlations are plotted, with pixel locations corresponding to the region boundaries via triangular projections. Two examples are illustrated. Pixel color indicates the degree of correlation ranging between -0.5 and 0. Correlation values are only plotted if at least 100 data points are included in the region. Non-significant correlation values ($p < 0.05$) are grey. For regions containing both highly enriched and de-enriched variants, pairwise distances become significantly predictive, and explain up to half of the variation observed. In contrast, regions composed of mostly enriched variants exhibit correlations that drastically lose significance and explain very little of the observed enrichment variation. These data suggest that while domain topology constrains the edges of acceptable deletions, amino acid distance does not constrain deletions that fall entirely within a removable domain.

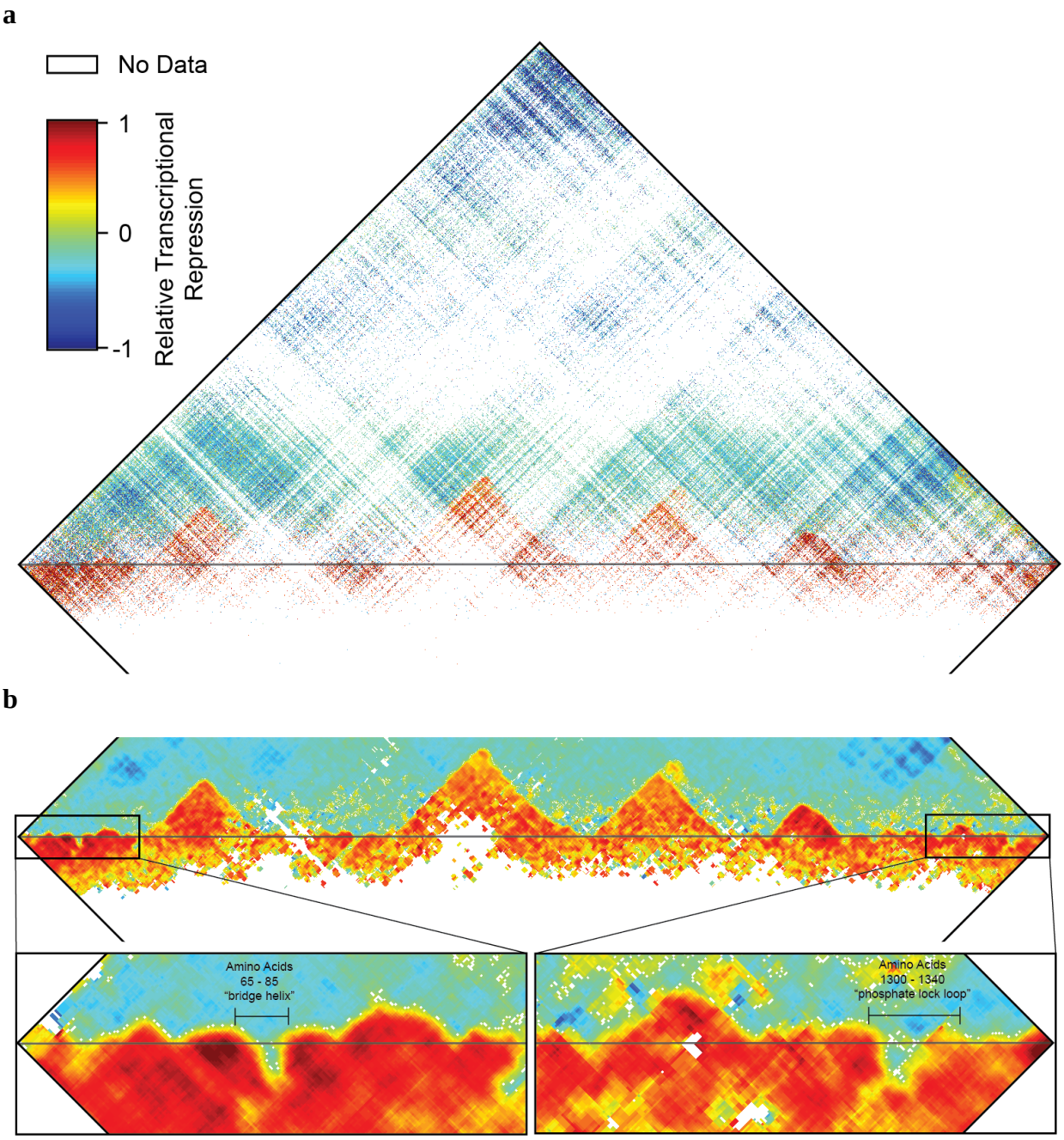
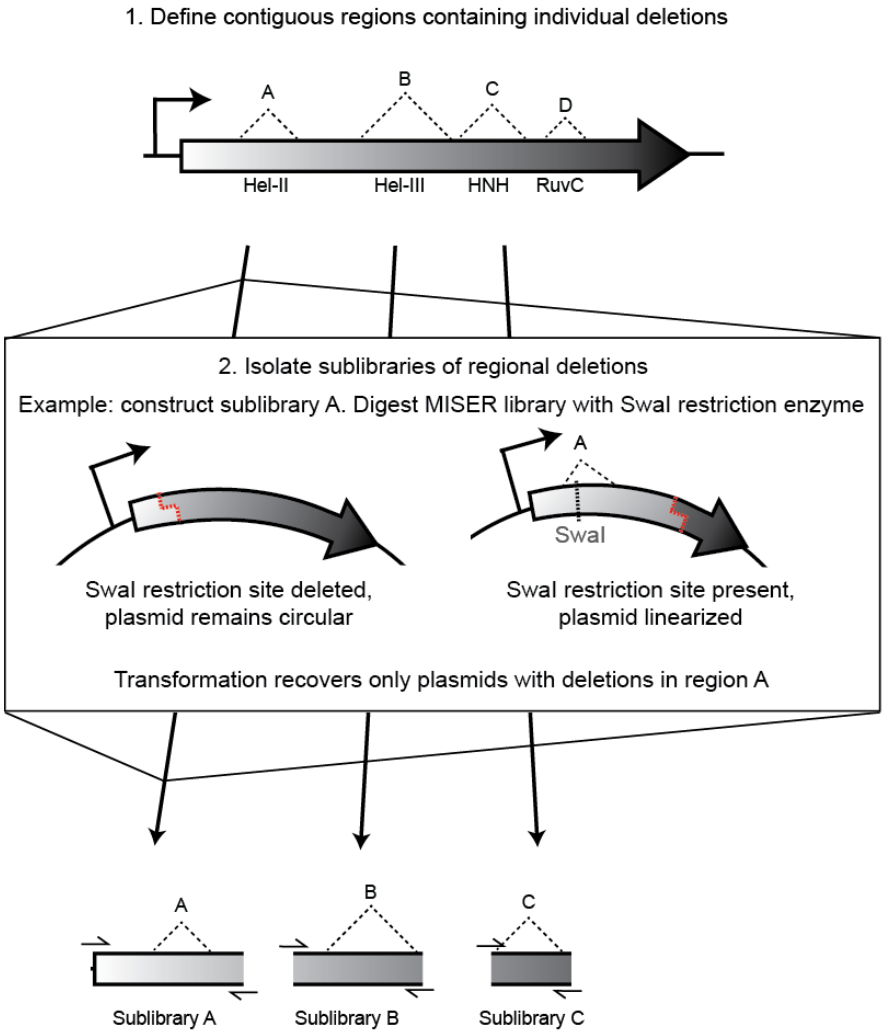
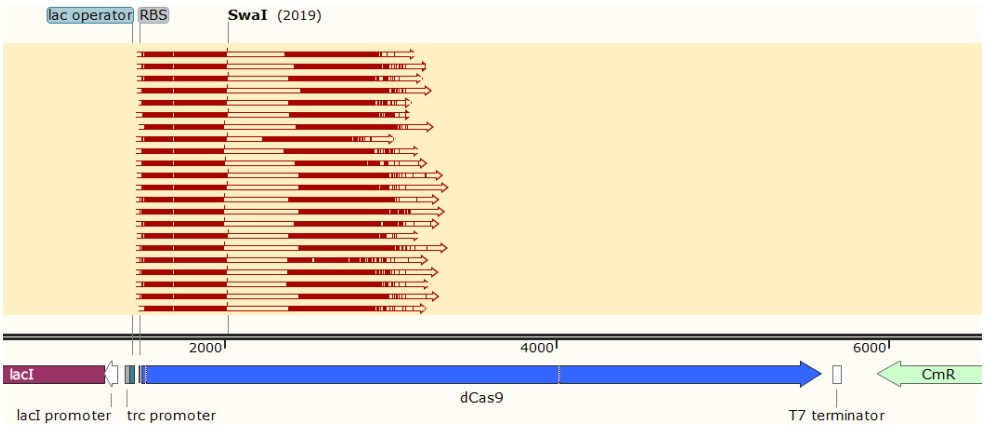


Figure 3.7: The full dCas9 MISER landscape contains insertions in addition to deletions (A) The enrichment map of Figure 3.1C is presented in its entirety, including small duplications of dCas9 sequence. The horizontal grey line corresponds to the boundary between deletions (top) and tandem duplicate insertions (bottom). Pixels reflected over the horizontal line represent the same dCas9 fragment. Note that in all cases a two amino acid MISER scar is also present (either Ala-Ser or Thr-Ser). (B) The combined enrichment map in A was interpolated to qualitatively highlight the boundaries between functional and non-functional deletions, which are not clearly visible in the raw data. Pixels were replaced by the mean enrichment value of neighboring deletions/duplications, plus itself, in a square window 10 amino acids wide. Windows with fewer than five values were left white. Insets: The N- and C-terminal regions qualitatively show particular regions where small insertions are strongly depleted, unlike in the majority of the protein. The 'bridge helix' and 'phosphate lock loop' are two examples of secondary structure which strongly disallow small insertions.

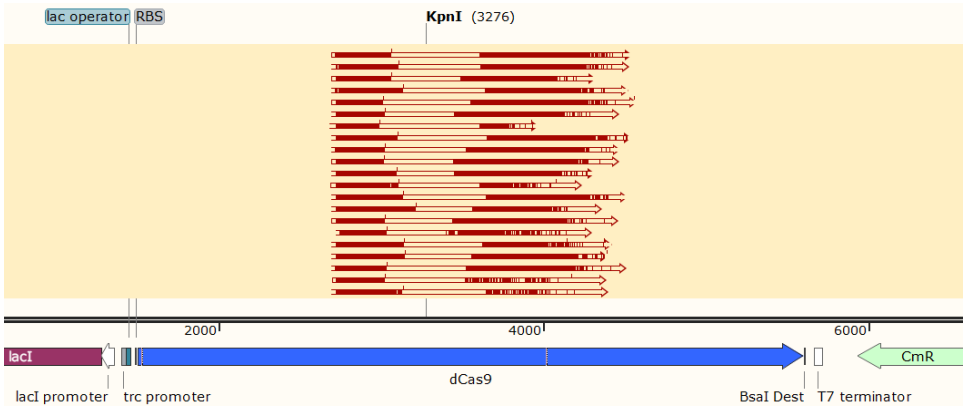
a


1. Define contiguous regions containing individual deletions
A
B
C
D
Hel-II
Hel-III
HNH
RuvC
2. Isolate sublibraries of regional deletions
Example: construct sublibrary A. Digest MISER library with SwaI restriction enzyme
A
SwaI
SwaI restriction site deleted, plasmid remains circular
SwaI restriction site present, plasmid linearized
Transformation recovers only plasmids with deletions in region A
A
B
C
Sublibrary A
Sublibrary B
Sublibrary C

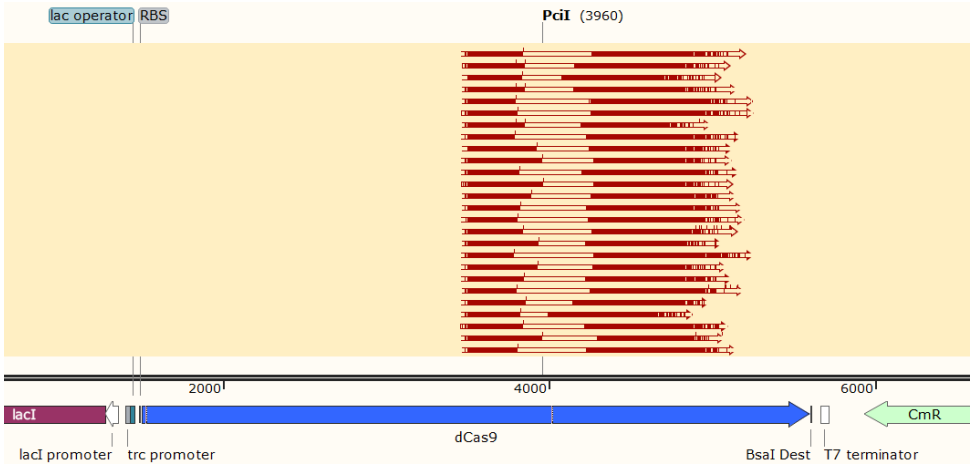
b


lac operator
RBS
SwaI (2019)
2000
4000
6000
lacI
CmR
dCas9
lacI promoter
trc promoter
T7 terminator

c



d



e

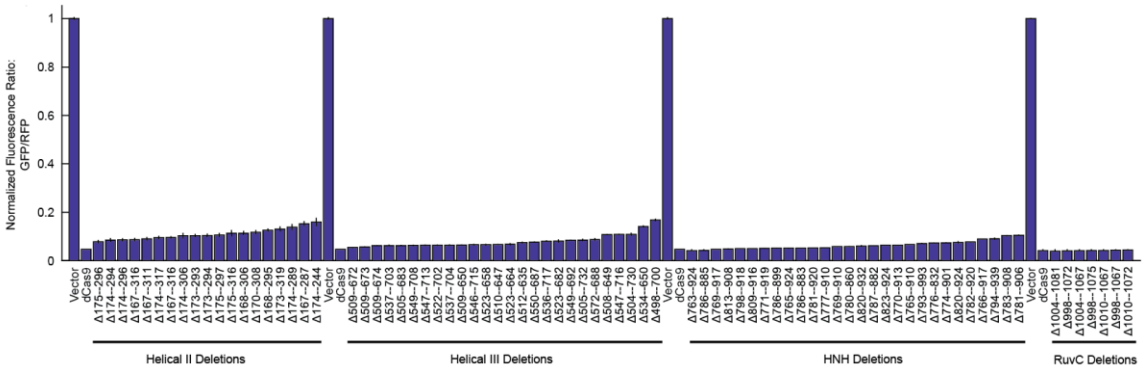


Figure 3.8: MISER sublibraries composed of specific deletions can be generated by restriction digestion. (A) Digesting a MISER library with a restriction enzyme that has exactly one site within the plasmid will linearize the majority of plasmids, while plasmids with the site deleted will remain circular. This reaction can then be transformed in order to recover a sublibrary containing deletions from a specific region. (B) The restriction enzyme SwaI was used to isolate deletions in the Helical II region. The enzyme recognition site is shown mapped to the sequence of pSAH064, the dCas9 expression plasmid, illustrating the overlap with various sequenced deletions. (C) The restriction enzyme KpnI was used to isolate deletions in the Helical III region, as in B. (D) The restriction enzyme PciI was used to isolate deletions in the HNH region, as in B. (E) Individual deletion variants were re-transformed and assayed for CRISPRi activity. RuvC deletions were cloned manually by PCR.

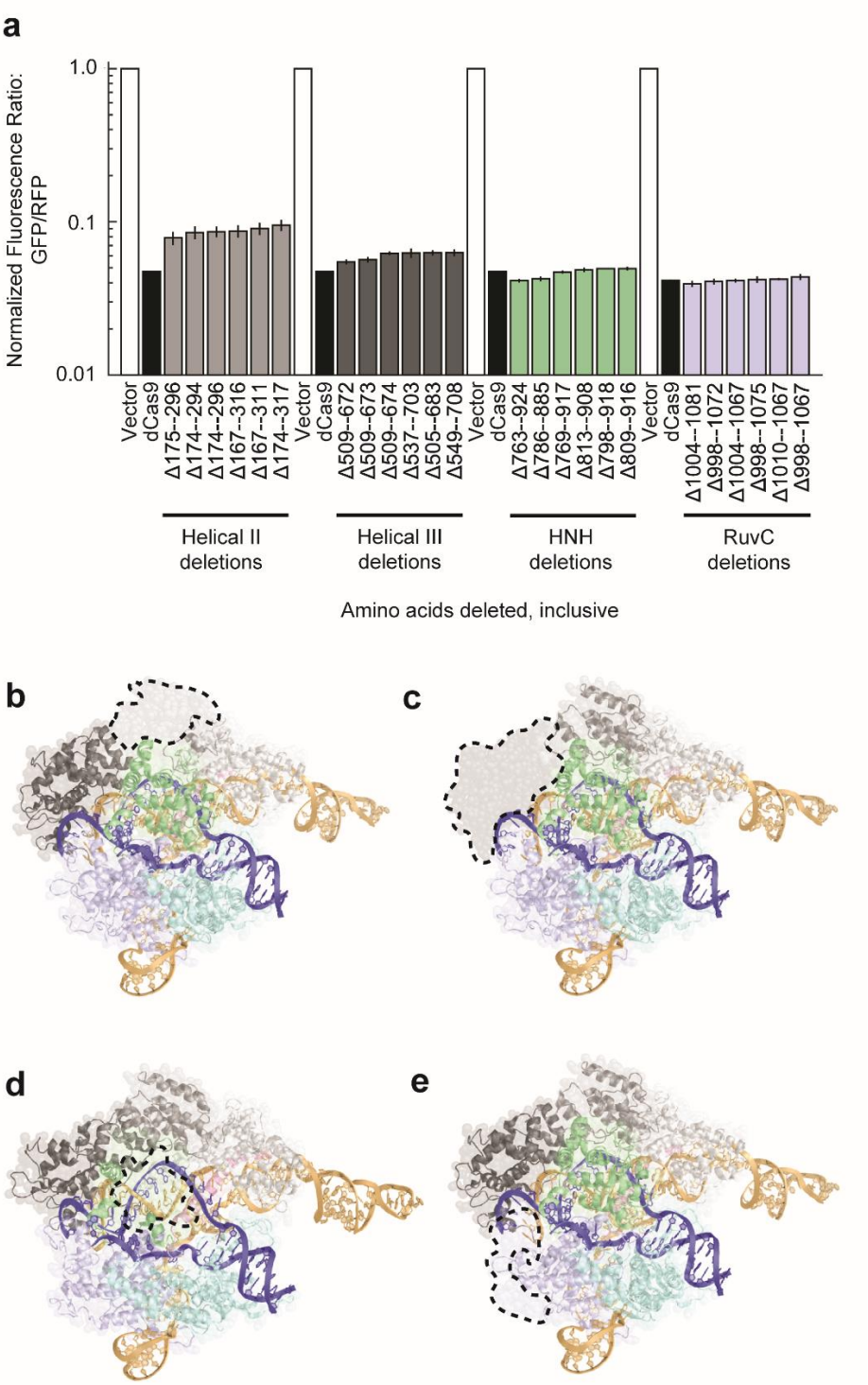
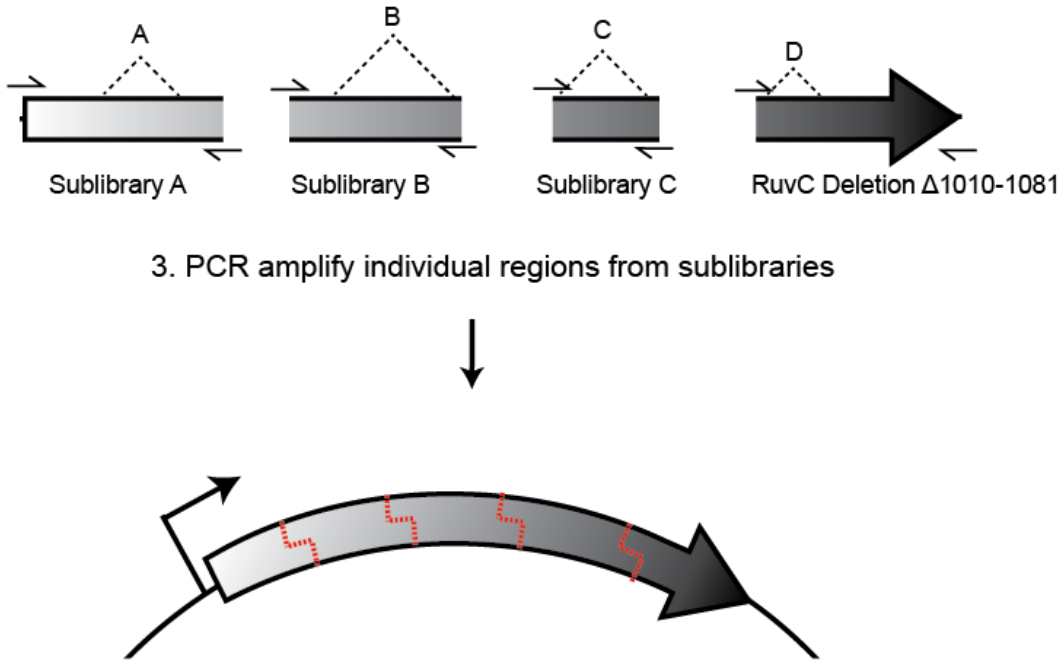


Figure 3.9: Individual deletion variants validate the MISER deletion landscape. (A) The top six functional CRISPRi deletion variants from each of the four deletion regions repress GFP nearly as well as dCas9. Notably, many of the top variants within a deletion region share a deletion start or deletion stop site. Additional variants for each region can be found in Figure 3.8E. (B). The Helical II deletion region is represented on the DNA-bound structure of Cas9 (PDB ID: 5F9R). (C) The Helical III deletion region, as in B. (D) The HNH deletion region, as in B. (E) The RuvC deletion region, as in B.

a


A
B
C
D
Sublibrary A
Sublibrary B
Sublibrary C
RuvC Deletion Δ1010-1081
3. PCR amplify individual regions from sublibraries
4. Clone fragments together and into expression plasmid
(here, Golden Gate cloning)
Library of quadruple deletions

b

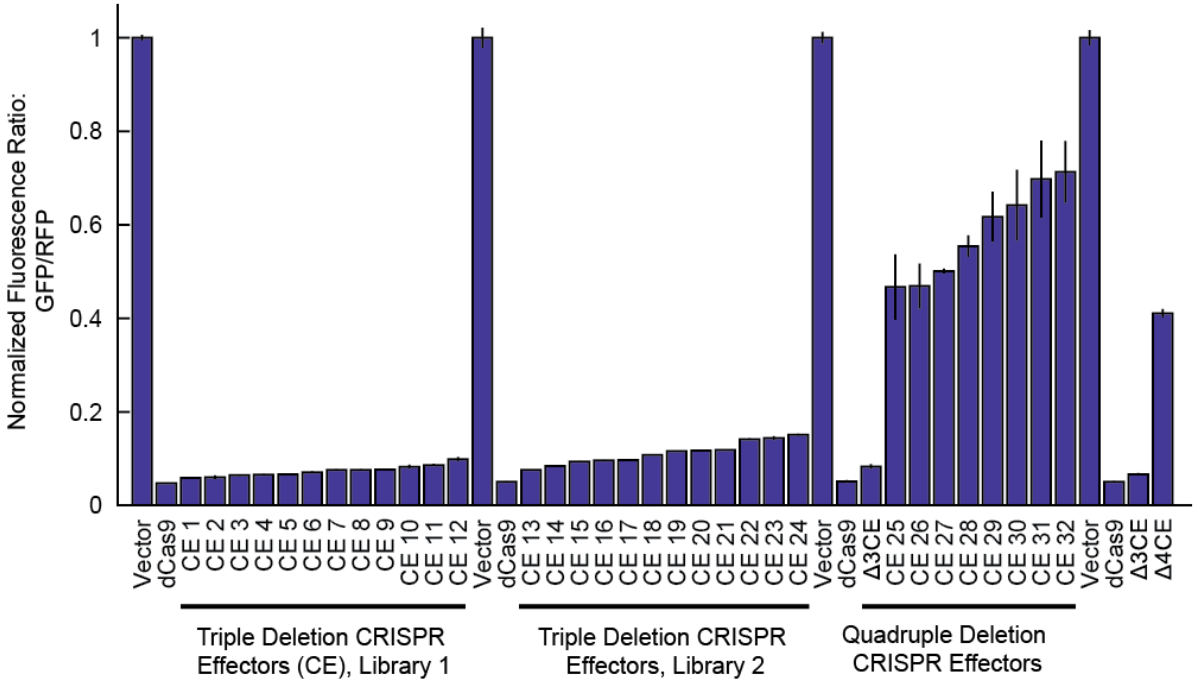

Normalized Fluorescence Ratio:
GFP/RFP
0
0.2
0.4
0.6
0.8
1
Vector
dCas9
CE 1
CE 2
CE 3
CE 4
CE 5
CE 6
CE 7
CE 8
CE 9
CE 10
CE 11
CE 12
Vector
dCas9
CE 13
CE 14
CE 15
CE 16
CE 17
CE 18
CE 19
CE 20
CE 21
CE 22
CE 23
CE 24
Vector
dCas9
Δ3CE
CE 25
CE 26
CE 27
CE 28
CE 29
CE 30
CE 31
CE 32
Vector
dCas9
Δ3CE
Δ4CE
Triple Deletion CRISPR
Effectors (CE), Library 1
Triple Deletion CRISPR
Effectors, Library 2
Quadruple Deletion
CRISPR Effectors

Figure 3.10: Golden Gate Cloning builds libraries of CRISPR Effector (CE) variants with multiple deletions. (A) One highly functional RuvC deletion variant from Region D was PCR amplified, along with Sublibraries A, B, and C. PCR primers added Golden Gate compatible sticky ends, enabling Golden Gate cloning of individual fragments to form a library of CE deletion variants, Library 1. (B) Flow cytometry was performed to isolate the most functional CE variants from Library 1. Selected sequences of CE variants can be found in Table 3.3. All highly functional CE variants were found to lack Helical-II deletions. To verify this result, a second version of Sublibrary A was created, using a different strategy to isolate Helical II deletions as follows: the full MISER library was digested with the restriction enzyme BlpI, which cuts at amino acids 227-228 (Instead of SwaI), and the resulting DNA was used directly as template for the PCR reaction (BlpI cuts pSAH064 three times and thus cannot be directly re-transformed to isolate the sublibrary). Library 2 thus contains quadruple deletion variants as in Library 1, but the sublibrary of Helical II deletions was entirely remade. Once again functional CE variants isolated by FACS lacked Helical-II deletions. The most functional variant in Library 2, CE 13, was named Δ3CE. Finally, to directly assay the effects of a Helical II deletion, the Helical-II region of Δ3CE was replaced with a library of deletions from Sublibrary A. These quadruple deletion CE variants all exhibited vastly reduced CRISPRi activity compared to Δ3CE alone. The most functional variant assayed was named Δ4CE.

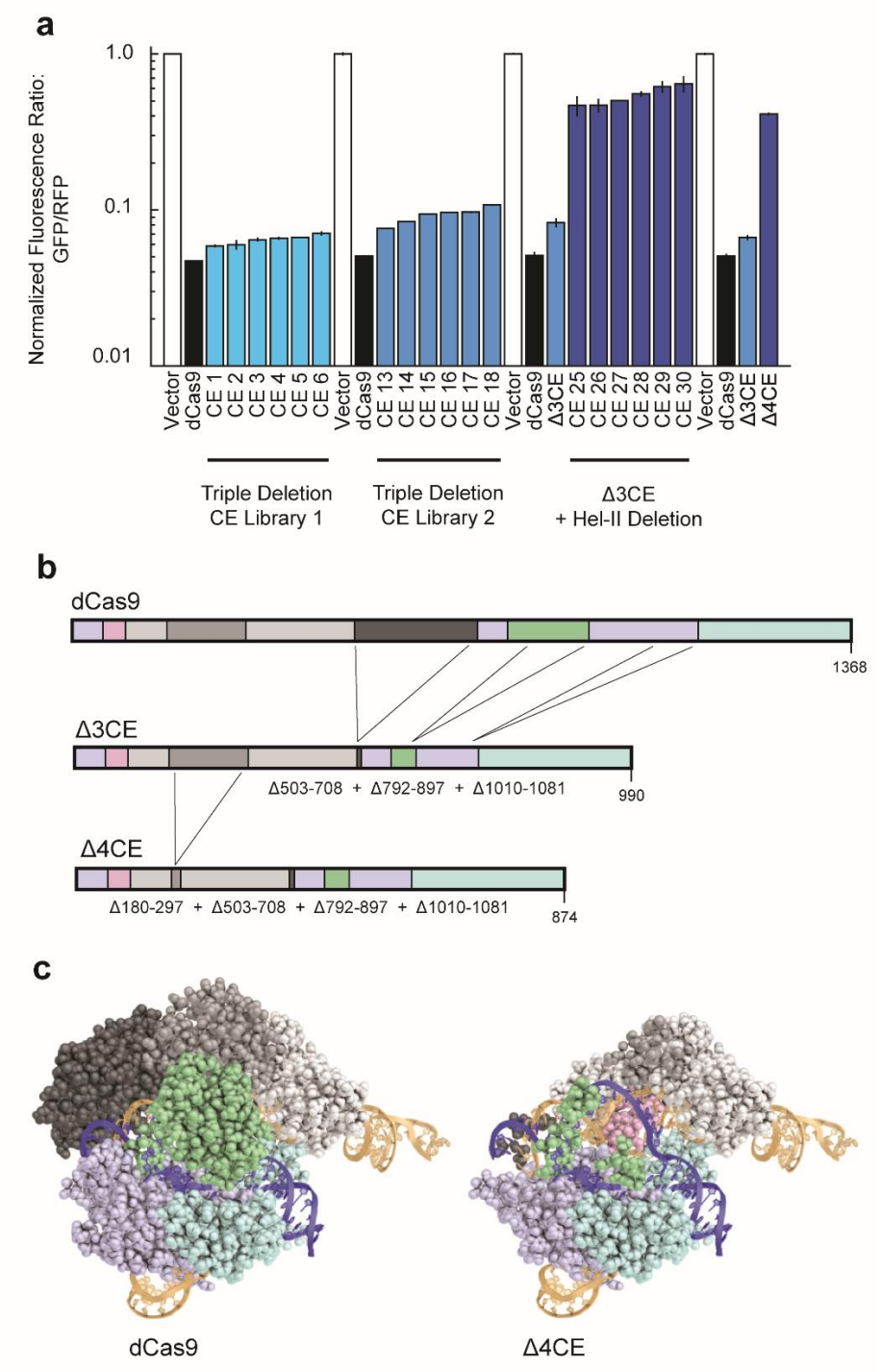


Figure 3.11: Functional deletions can be combined to identify the minimal RNA-guided DNA binding element within dCas9. (A) Combinatorial libraries of combined functional deletions yield novel minimal CRISPR Effector (CE) proteins. CE variants in Library 1 and Library 2 are composed of triple deletions in Helical III, HNH, and RuvC domains. CE 13 was named Δ3CE. CE 25-30 are variants of Δ3CE with a fourth deletion in the Helical II domain. CE 25 was named Δ4CE. (B) Cartoon domain representation of dCas9, Δ3CE, and Δ4CE. (C) The intact domains of dCas9 or Δ4CE are represented on the DNA-bound structure of Cas9 (PDB ID: 5F9R).

region. Again, the most functional CE variants isolated by FACS did not contain Helical II deletions. Finally, in an attempt to force a minimal length CE, a highly active CE variant was named Δ3CE and directly combined with a library of Helical II deletions. The resulting enforced quadruple deletion CE variants all exhibited drastic loss of function. The most active of these, Δ3CE with a further deletion of Δ180-297, was named Δ4CE. Δ3CE and Δ4CE were re-transformed and assayed once more to verify their phenotype.

3.4 Discussion

In this work we have defined Δ3CE and Δ4CE as a minimal protein scaffold required for CRISPRi, identifying the necessary competent DNA-binding elements found within dCas9. To accomplish this we have modularly removed regions responsible for nuclease regulation and activity as well as a structural region (Figure 3.11). Δ3CE is less than 1000 amino acids in length and retains near-WT GFP repression under CRISPRi, demonstrating the reversibility of evolutionary domain aggregation. In contrast, deleting Helical II to generate Δ4CE was found to drastically reduce function, an unexpected finding that recapitulates previous observations of ~50% reduction in cleavage activity in Cas9 Δ175-307 (Nishimasu et al. 2014). This previously observed functional deficiency was proposed to result from reduced expression, but we find that at even the highest levels of expression Δ4CE cannot be rescued, while Δ3CE remains highly functional even at very low levels of induction (Figure 3.12). We propose that our comprehensive investigation of dCas9 deletions has pared down protein structure in a manner that has potentially elucidated a fundamental role for the Helical II domain in Cas9's DNA binding and cleavage mechanisms. In brief, every other domain in Cas9 has an identified mechanism: RuvC is responsible for non-target strand cleavage (Jinek et al. 2012); Helical I for binding the guide RNA (Nishimasu et al. 2014); Helical III for regulating the sensing of the DNA:RNA duplex (Chen et al. 2017); the HNH domain for target strand cleavage (Jinek et al. 2012); and the CTD domain for PAM interaction and potentially DNA bubble initiation (Anders et al. 2014).

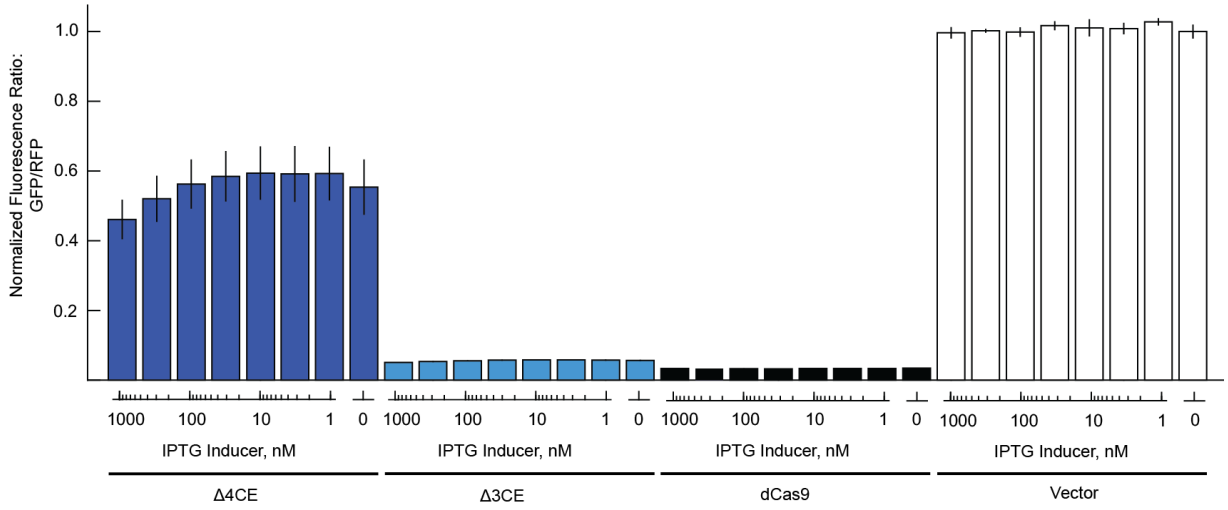


Figure 3.12: Δ4CE CRISPRi activity is not rescued by increased expression. Δ3CE and dCas9 remain highly functional even at very low levels of induction.

The Helical II domain is the least well understood mechanistically and is the least represented among Cas9 homologs, yet intriguingly SpyCas9 is better equipped to unwind dsDNA than many Cas9 homologs (E. Ma et al. 2015). A recent report investigating molecular dynamics using Förster resonance energy transfer has proposed a mechanism for the Helical II domain in stabilizing the unwinding of the NT strand and regulating the conformational transition to the on-target state, fully displacing the non-target strand (Sung et al. 2018). Here, we have shown that removal of this domain in the stacked MISER CE's severely perturbs CRISPRi repression, consistent with the general assertion that Helical II performs a key role in stabilizing DNA binding. A specific mechanism involving stabilization of the non-target strand initially might allow for the perpetuation of the fundamental 'Brownian ratchet' underlying guideRNA:targetDNA hybridization (Sternberg et al. 2014). Our data is consistent with this mechanism, but future investigations will be required to confidently classify the Helical II domain as a non-target strand DNA binding domain that stabilizes the unwound state.

We have also discovered an entirely new deletion set in the RuvC III domain that retains full CRISPRi function. Intriguingly, this deletion does not seem to overlay with a known functional domain and thus may serve as a module that further stabilizes the RuvC domain as a whole. Additionally, this deletion abuts the non-target and target strand DNA (~4-6 angstroms) and may provide a highly useful site to replace with accessory fusions such as deaminases that base edit the non-target stand and may be sterically blocked by these amino acids.

The MISER approach is programmable and comprehensive, allowing the functional, comprehensive annotation of deletion landscapes for the first time. Here we have revealed all known functional dCas9 deletions and one unknown deletion in a single experiment. This demonstration of MISERs ability to comprehensively delineate domain function highlights the power of this technique for generating useful & active engineered proteins. In contrast, rational approaches can avoid dealing with large library sizes but must instead rely on previous knowledge, which may be biased. MISER represents another tool toward fully developed methods in molecular biology that will explore comprehensive function landscapes in an unbiased manner. Understanding protein functional landscapes will fully realize the potential of protein engineering to ameliorate humanity's most important problems, such as human food security (D. Ma et al. 2018). Finally, from an evolutionary perspective, internal domain deletion is a relatively rare mutation, with domain architectures tending to aggregate over time. Here we have demonstrated that MISER libraries can be iteratively combined to reverse this process, isolating the minimal module that is specialized for a specific function. We anticipate that these comprehensive approaches to profiling protein domains will have broad application in both mechanistic biochemistry and protein engineering.

3.5 Tables

	SpeI Insertion	NheI Insertion
Recombineering Oligo: Insertion Site 1	'*AACACGTCCGTCCTAGAACT***cgtctc** atacgcaaaccgcctctccccgcgcgttggcg gtctcaatct**ATG**actagtgataagaaatac tcaataggcttagctatcggcacaaatagcgt cgg**gagacg***GCAAGCGGTACACTCAG ATCAGTGTTGAGCGTAACCAAGT*'	'*AACACGTCCGTCCTAGAACT***cgtctc** atacgcaaaccgcctctccccgcgcgttggcg gtctcaatct**ATG**gctagcgataagaaatac tcaataggcttagctatcggcacaaatagcgt cgg**gagacg***GCAAGCGGTACACTCAG ATCAGTGTTGAGCGTAACCAAGT*'

Table 3.1: Example Oligo Library Synthesis (OLS) oligonucleotides used in this study. The full list of ordered oligonucleotides is available as 'Auxiliary Supplementary Materials - Recombineering Oligonucleotides'. All oligonucleotides were ordered from Agilent Technologies, Inc. Oligos were designed to incorporate 45 and 47 bp of homology upstream or downstream of the insertion site, respectively (lowercase). Six bp were inserted between dCas9 codons, beginning after the target codon. The above example targets the start codon, 'ATG' (bold uppercase). These six bp consisted of recognition sequences for either the restriction enzyme SpeI or NheI (underlined). Flanking primer sequences allowed the amplification of the entire OLS library (italic) using primers SAH_284 and SAH_285 (Table 3.4). Specific libraries of SpeI recombineering oligonucleotides or NheI recombineering oligonucleotides were amplified using forward primer SAH_284 and either SAH_286 or SAH_287 reverse primers, respectively. After amplification, these dsDNA products can be 'matured' by cleavage with the restriction enzyme BsmbI (bold lowercase), which cleaves internally of its recognition site, thus removing all non-homologous priming sequence from the recombineering template.

	Total Reads	Deletions Sequenced	Unique Deletions	Enriched Unique Deletions	Depleted Unique Deletions
Slice 4 Naïve	132,274,232	1,923,543	192,447		
Slice 4 Sorted	140,589,968	1,960,138	25,948	19,618	6,330
Slice 5 Naïve	37,873,068	590,859	111,438		
Slice 5 Sorted	35,016,326	290,947	51,462	31,794	19,668
Total	345,753,594	4,765,487	381,295	51,412	25,998

Table 3.2: Statistics for deep sequencing of MISER libraries Slice 4 and Slice 5.

Deletion	CE 1	CE 2	CE 3	CE 4	CE 5	CE 6	Δ3CE (CE 13)	Δ4CE (CE 25)
Helical II								[180-297]
Helical III	[511-716]	[498-699]	[500-688]	[497-700]	[501-664]	[512-701]	[503-708]	[503-708]
HNH	[813-909]	[813-908]	[811-898]	[786-882]	[804-893]	[809-916]	[792-897]	[792-897]
RuvC	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]

Table 3.3: Deletions present in selected MISER variants. Indicated numbers represent the first and last amino acid deleted from the protein.

Primer Name	Primer Notes	Full Sequence
SAH_284	Recombineering oligonucleotide amplification: universal forward	AACACGTCCGTCCTAGAACT
SAH_285	Recombineering oligonucleotide amplification: universal reverse	ACTTGGTTACGCTCAACACT
SAH_286	Recombineering oligonucleotide amplification: SpeI specific reverse	GATCTGAGTGTACCGCTTGC
SAH_287	Recombineering oligonucleotide amplification: NheI specific reverse	GATCGCCTAGACAACTCCTG
SAH_292	Amplify chloramphenicol cassette forward, adds SpeI site	CACACCAACTAGTGACGTCGATATCTGG CGAAAAT
SAH_293	Amplify chloramphenicol cassette reverse, adds SpeI site	TTGTTACTAGTGCTTGGATTCTCACC
SAH_294	Amplify chloramphenicol cassette forward, adds NheI site	CACACCAGCTAGCGACGTCGATATCTGG CGAAAAT
SAH_295	Amplify chloramphenicol cassette reverse, adds NheI site	CACACCAGCTAGCGCTTGGATTCTCACC AATAAAAAACG
SAH_356	Amplify dCas9 from pSAH064, forward	GAGCGGATAACAATTCCCCTGT
SAH_358	Amplify dCas9 from pSAH064, reverse	GGCTGTGGTGATGATGGTG

Table 3.4: PCR primers used in this study. All primers were ordered from Integrated DNA Technologies, Inc.

Sequence Name	Cloning notes	Full Sequence
Chloramphenicol Selection Fragment	PCR product for SpeI or NheI recombineering selection	CACACCAGCTAGCGACGTCGATATCTGGCGAAAATGAGACGTT GATCGGCACGTAAGAGGTTCCAACTTTCACCATAATGAAATAA GATCACTACCGGGCGTATTTTTTGAGTTATCGAGATTTTCAGGA GCTAAGGAAGCTAAAATGGAGAAAAAAATCACTGGATATACCA CCGTTGATATATCCCAATGGCATCGTAAAGAACATTTTGAGGCA TTTCAGTCAGTTGCTCAATGTACCTATAACCAGACCGTTCAGCT GGATATTACGGCCTTTTTAAAGACCGTAAAGAAAAATAAGCAC AAGTTTTATCCGGCCTTTATTCACATTCTTGCCCGCCTGATGAAT GCTCATCCGGAATTTCGTATGGCAATGAAAGACGGTGAGCTGG TGATATGGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAG CAAACTGAAACGTTTTCATCGCTCTGGAGTGAATACCACGACG ATTTCCGGCAGTTTCTACACATATATTCGCAAGATGTGGCGTGT TACGGTGAAAACCTGGCCTATTTCCCTAAAGGGTTTATTGAGAA TATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCACCAGTTT TGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTT TCACCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGAT GCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCC ATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATGA GTGGCAGGGCGGGGCGTAATTTGATATCGAGCTCGCTTGGACT CCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATT TGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTGAG AATCCAAGCGCTAGCTGGTGTG

Table 3.5: Full sequence of the chloramphenicol selection fragment.

Plasmid name	Cloning notes	Full Sequence
pSAH060	Recombineering target plasmid. Lacks promoter. Flanking BsaI sites for Golden Gate cloning	TTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTT ATTGTTCATGACCATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGA AAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAAC AAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTT TCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGC CGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAA TCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAA GACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACA GCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAG AAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGT CGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTC CTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGC GGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGC CTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGA GCGAGGAAGCGGAAGAGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCGGTCT CAATCTATGGATAAGAAATACTCAATAGGCTTAGCTATCGGCACAAATAGCGTCGGATG GGCGGTGATCACTGATGAATATAAGGTTCCGTCTAAAAAGTTCAAGGTTCTGGGAAATA CAGACCGCCACAGTATCAAAAAAAATCTTATAGGGGCTCTTTTATTTGACAGTGGAGAG ACAGCGGAAGCGACTCGTCTAAAACGGACAGCTCGTAGAAGGTATACACGTCGGAAGA ATCGTATTTGTTATCTACAGGAGATTTTTTCAAATGAGATGGCGAAAGTAGATGATAGTT TCTTTCATCGACTTGAAGAGTCTTTTTTGGTGGAAGAAGACAAGAAGCATGAACGTCATC CTATTTTTGGAAATATAGTAGATGAAGTTGCTTATCATGAGAAATATCCAACTATCTATC ATCTGCGAAAAAAATTGGTAGATTCTACTGATAAAGCGGATTTGCGCTTAATCTATTTGG CCTTAGCGCATATGATTAAGTTTCGTGGTCATTTTTTGATTGAGGGAGATTTAAATCCTG ATAATAGTGATGTGGACAAACTATTTATCCAGTTGGTACAAACCTACAATCAATTATTTG AAGAAAACCCTATTAACGCAAGTGGAGTAGATGCTAAAGCGATTCTTTCTGCACGATTG AGTAAATCAAGACGATTAGAAAATCTCATTGCTCAGCTCCCCGGTGAGAAGAAAAATGG CTTATTTGGGAATCTCATTGCTTTGTCATTGGGTTTGACCCCTAATTTTAAATCAAATTTT GATTTGGCAGAAGATGCTAAATTACAGCTTTCAAAAGATACTTACGATGATGATTTAGA TAATTTATTGGCGCAAATTGGAGATCAATATGCTGATTTGTTTTTGGCAGCTAAGAATTT ATCAGATGCTATTTTACTTTCAGATATCCTAAGAGTAAATACTGAAATAACTAAGGCTCC CCTATCAGCTTCAATGATTAAACGCTACGATGAACATCATCAAGACTTGACTCTTTTAAA AGCTTTAGTTCGACAACAACTTCCAGAAAAGTATAAAGAAATCTTTTTTGATCAATCAAA AAACGGATATGCAGGTTATATTGATGGGGGAGCGAGCCAAGAAGAATTTTATAAATTTA TCAAACCAATTTTAGAAAAAATGGATGGTACTGAGGAATTATTGGTGAAACTAAATCGT GAAGATTTGCTGCGCAAGCAACGGACCTTTGACAACGGCTCTATTCCCCATCAAATTCAC TTGGGTGAGCTGCATGCTATTTTGAGAAGACAAGAAGACTTTTATCCATTTTTAAAAGAC AATCGTGAGAAGATTGAAAAAATCTTGACTTTTCGAATTCCTTATTATGTTGGTCCATTG GCGCGTGGCAATAGTCGTTTTGCATGGATGACTCGGAAGTCTGAAGAAACAATTACCCC ATGGAATTTTGAAGAAGTTGTCGATAAAGGTGCTTCAGCTCAATCATTTATTGAACGCAT GACAAACTTTGATAAAAATCTTCCAAATGAAAAAGTACTACCAAAACATAGTTTGCTTT ATGAGTATTTTACGGTTTATAACGAATTGACAAAGGTCAAATATGTTACTGAAGGAATG CGAAAACCAGCATTTCTTTCAGGTGAACAGAAGAAAGCCATTGTTGATTTACTCTTCAAA ACAAATCGAAAAGTAACCGTTAAGCAATTAAAAGAAGATTATTTCAAAAAAATAGAAT GTTTTGATAGTGTTGAAATTTCAGGAGTTGAAGATAGATTTAATGCTTCATTAGGTACCT ACCATGATTTGCTAAAAATTATTAAAGATAAAGATTTTTTGGATAATGAAGAAAATGAA GATATCTTAGAGGATATTGTTTTAACATTGACCTTATTTGAAGATAGGGAGATGATTGAG GAAAGACTTAAAACATATGCTCACCTCTTTGATGATAAGGTGATGAAACAGCTTAAACG TCGCCGTTATACTGGTTGGGGACGTTTGTCTCGAAAATTGATTAATGGTATTAGGGATAA GCAATCTGGCAAAACAATATTAGATTTTTTGAAATCAGATGGTTTTGCCAATCGCAATTT TATGCAGCTGATCCATGATGATAGTTTGACATTTAAAGAAGACATTCAAAAAGCACAAG TGTCTGGACAAGGCGATAGTTTACATGAACATATTGCAAATTTAGCTGGTAGCCCTGCTA TTAAAAAAGGTATTTTACAGACTGTAAAAGTTGTTGATGAATTGGTCAAAGTAATGGGG CGGCATAAGCCAGAAAATATCGTTATTGAAATGGCACGTGAAAATCAGACAACTCAAAA GGGCCAGAAAAATTCGCGAGAGCGTATGAAACGAATCGAAGAAGGTATCAAAGAATTA GGAAGTCAGATTCTTAAAGAGCATCCTGTTGAAAATACTCAATTGCAAAATGAAAAGCT CTATCTCTATTATCTCCAAAATGGAAGAGACATGTATGTGGACCAAGAATTAGATATTA ATCGTTTAAGTGATTATGATGTCGATGCCATTGTTCCACAAAGTTTCCTTAAAGACGATT CAATAGACAATAAGGTCTTAACGCGTTCTGATAAAAATCGTGGTAAATCGGATAACGTT CCAAGTGAAGAAGTAGTCAAAAAGATGAAAAACTATTGGAGACAACTTCTAAACGCCA AGTTAATCACTCAACGTAAGTTTGATAATTTAACGAAAGCTGAACGTGGAGGTTTGAGT GAACTTGATAAAGCTGGTTTTATCAAACGCCAATTGGTTGAAACTCGCCAAATCACTAA GCATGTGGCACAAATTTTGGATAGTCGCATGAATACTAAATACGATGAAAATGATAAAC TTATTCGAGAGGTTAAAGTGATTACCTTAAAATCTAAATTAGTTTCTGACTTCCGAAAAG

		ATTTCCAATTCTATAAAGTACGTGAGATTAACAATTACCATCATGCCCATGATGCGTATC TAAATGCCGTCGTTGGAACTGCTTTGATTAAGAAATATCCAAAACTTGAATCGGAGTTTG TCTATGGTGATTATAAAGTTTATGATGTTCGTAAAATGATTGCTAAGTCTGAGCAAGAAA TAGGCAAAGCAACCGCAAAATATTTCTTTTACTCTAATATCATGAACTTCTTCAAAACAG AAATTACACTTGCAAATGGAGAGATTCGCAAACGCCCTCTAATCGAAACTAATGGGGAA ACTGGAGAAATTGTCTGGGATAAAGGGCGAGATTTTGCCACAGTGCGCAAAGTATTGTC CATGCCCCAAGTCAATATTGTCAAGAAAACAGAAGTACAGACAGGCGGATTCTCCAAGG AGTCAATTTTACCAAAAAGAAATTCGGACAAGCTTATTGCTCGTAAAAAAGACTGGGAT CCAAAAAAATATGGTGGTTTTGATAGTCCAACGGTAGCTTATTCAGTCCTAGTGGTTGCT AAGGTGGAAAAAGGGAAATCGAAGAAGTTAAAATCCGTTAAAGAGTTACTAGGGATCA CAATTATGGAAAGAAGTTCCTTTGAAAAAAATCCGATTGACTTTTTAGAAGCTAAAGGA TATAAGGAAGTTAAAAAAGACTTAATCATTAAACTACCTAAATATAGTCTTTTTGAGTTA GAAAACGGTCGTAAACGGATGCTGGCTAGTGCCGGAGAATTACAAAAAGGAAATGAGC TGGCTCTGCCAAGCAAATATGTGAATTTTTTATATTTAGCTAGTCATTATGAAAAGTTGA AGGGTAGTCCAGAAGATAACGAACAAAAACAATTGTTTGTGGAGCAGCATAAGCATTAT TTAGATGAGATTATTGAGCAAATCAGTGAATTTTCTAAGCGTGTTATTTTAGCAGATGCC AATTTAGATAAAGTTCTTAGTGCATATAACAAACATAGAGACAAACCAATACGTGAACA AGCAGAAAATATTATTCATTTATTTACGTTGACGAATCTTGGAGCTCCCGCTGCTTTTAA ATATTTTGATACAACAATTGATCGTAAACGATATACGTCTACAAAAGAAGTTTTAGATGC CACTCTTATCCATCAATCCATCACTGGTCTTTATGAAACACGCATTGATTTGAGTCAGCT AGGAGGTGACTAATAGGTGAGACCTCCGCTTACAGACAAGCTGTGACCGTCACCGGGAG CTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCTTAGAAAAACTCATCGAGC ATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGC CGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTG GTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTC AAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAAT GGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCA TCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGGCG AAATACGCGATCACTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGC AGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACC TGGAATGCTGTTTTGCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACG GATAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCA TCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCG CATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAG CCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAG ACGT
pSAH063	MISER expression plasmid. Golden Gate compatible with pSAH060	CAATAAACCCTTTAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTTGCGAA TATATGTGTAGAAACTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCGATGAAAACGT TTCAGTTTGCTCATGGAAAACGGTGTAACAAGGGTGAACACTATCCCATATCACCAGCT CACCGTCTTTCATTGCCATACGGAACTCCGGATGAGCATTCATCAGGCGGGCAAGAATG TGAATAAAGGCCGGATAAAACTTGTGCTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTA ATATCCAGCTGAACGGTCTGGTTATAGGTACATTGAGCAACTGACTGAAATGCCTCAAA ATGTTCTTTACGATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTTTTTTCTCC ATTTTAGCTTCCTTAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGGTAGTGAT CTTATTTCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCAACGTCTCATTTTCGCC AAAAGTTGGCCCAGGGCTTCCCGGTATCAACAGGGACACCAGGATTTATTTATTCTGCG AAGTGATCTTCCGTCACAGGTATTTATTCGGCGCAAAGTGCGTCGGGTGATGCTGCCAAC TTACTGATTTAGTGTATGATGGTGTTTTTGAGGTGCTCCAGTGGCTTCTGTTTCTATCAGC TGTCCCTCCTGTTCAGCTACTGACGGGGTGGTGCGTAACGGCAAAAGCACCGCCGGACA TCAGCTCTAGCGGAGTGTATACTGGCTTACTATGTTGGCACTGATGAGGGTGTCAGTGAA GTGCTTCATGTGGCAGGAGAAAAAAGGCTGCACCGGTGCGTCAGCAGAATATGTGATAC AGGATATATTCCGCTTCCTCGCTCACTGACTCGCTACGCTCGGTCGTTCGACTGCGGCGA GCGGAAATGGCTTACGAACGGGGCGGAGATTTCCTGGAAGATGCCAGGAAGATACTTA ACAGGGAAGTGAGAGGGCCGCGGCAAAGCCGTTTTTCCATAGGCTCCGCCCCCCTGACA AGCATCACGAAATCTGACGCTCAAATCAGTGGTGGCGAAACCCGACAGGACTATAAAG ATACCAGGCGTTTCCCCTGGCGGCTCCCTCGTGCGCTCTCCTGTTCCTGCCTTTCGGTTTA CCGGTGTCATTCCGCTGTTATGGCCGCGTTTGTCTCATTCCACGCCTGACACTCAGTTCCG GGTAGGCAGTTCGCTCCAAGCTGGACTGTATGCACGAACCCCCCGTTCAGTCCGACCGC TGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGAAAGACATGCAAAAGCACC ACTGGCAGCAGCCACTGGTAATTGATTTAGAGGAGTTAGTCTTGAAGTCATGCGCCGGT TAAGGCTAAACTGAAAGGACAAGTTTTGGTGACTGCGCTCCTCCAAGCCAGTTACCTCG GTTCAAAGAGTTGGTAGCTCAGAGAACCTTCGAAAAACCGCCCTGCAAGGCGGTTTTTT CGTTTTCAGAGCAAGAGATTACGCGCAGACCAAAACGATCTCAAGAAGATCATCTTATT AATCAGATAAAATATTTCTAGATTTCAGTGCAATTTATCTCTTCAAATGTAGCACCTGAA GTCAGCCCCATACGATATAAGTTGTAATTCTCATGTTAGTCATGCCCCGCGCCCACCGGA AGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGA GTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTG TCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGG GCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCG

		CCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAA TCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTAT CCCACTACCGAGATGTCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGC GCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCA GCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCC GAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAG ATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGT CTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAA TGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGA AGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACC ACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGC GTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTT GTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCC GCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAG ACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAAT TGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTG TCCGGGATCTCGACGCTCTCCCTTATGCGACTCCGATCCCGCAAATTTGACAATTAATCA TCCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTCCCCTGTAGAAATAATTT TGTTTAACTTTAATAAGGAGATAATCTTGAGACCTGGTGTGCACACCAGGTCTCATAGGG GCAGCAGCCATCACCATCATCACCACAGCCAGGATCCTAGGCTGCTGCCACCGCTGAGC AATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAC CTCAGGCATTTGAGAAGCACACGGTCACACTGCTTCCGGTAGTCAATAAACCGGTAAAC CAGCAATAGACATAAGCGGCTATTTAACGACCCTGCCCTGAACCGACGACCGGGTCGAA TTTGCTTTCGAATTTCTGCCATTCATCCGCTTATTATCACTTATTCAGGCGTAGCACCAGG CGTTTAAGGGCACCAATAACTGCCTTAAAAAAATTACGCCCCGCCCTGCCACTCATCGC AGTACTGTTGTAATTCATTAAGCATTCTGCCGACATGGAAGCCATCACAGACGGCATGA TGAACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATA GTGAAAACGGGGGCGAAGAAGTTGTCCATATTGGCCACGTTTAAATCAAAACTGGTGAA ACTCACCCAGGGATTGGCTGAGACGAAAAACATATTCT
pSAH064	dCas9 expression plasmid	CTAGATTTCAGTGCAATTTATCTCTTCAAATGTAGCACCTGAAGTCAGCCCCATACGATA TAAGTTGTAATTCTCATGTTAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTG AAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATT AATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTA ATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTT CTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAG TTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGG TTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATGT CCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGA TCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGT TGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTG CGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGG GCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTC GCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCA AGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATC CAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCG CTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTT GATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTG GAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTT GGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAAC GTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTG CGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGC GCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGC TCTCCCTTATGCGACTCCGATCCCGCAAATTTGACAATTAATCATCCGGCTCGTATAATG TGTGGAATTGTGAGCGGATAACAATTCCCCTGTAGAAATAATTTTGTTTAACTTTAATAA GGAGATAATCTATGGATAAGAAATACTCAATAGGCTTAGCTATCGGCACAAATAGCGTC GGATGGGCGGTGATCACTGATGAATATAAGGTTCCGTCTAAAAAGTTCAAGGTTCTGGG AAATACAGACCGCCACAGTATCAAAAAAAATCTTATAGGGGCTCTTTTATTTGACAGTG GAGAGACAGCGGAAGCGACTCGTCTAAAACGGACAGCTCGTAGAAGGTATACACGTCG GAAGAATCGTATTTGTTATCTACAGGAGATTTTTTCAAATGAGATGGCGAAAGTAGATG ATAGTTTCTTTCATCGACTTGAAGAGTCTTTTTTGGTGGAAGAAGACAAGAAGCATGAAC GTCATCCTATTTTTGGAAATATAGTAGATGAAGTTGCTTATCATGAGAAATATCCAACTA TCTATCATCTGCGAAAAAAATTGGTAGATTCTACTGATAAAGCGGATTTGCGCTTAATCT ATTTGGCCTTAGCGCATATGATTAAGTTTCGTGGTCATTTTTTGATTGAGGGAGATTTAA ATCCTGATAATAGTGATGTGGACAAACTATTTATCCAGTTGGTACAAACCTACAATCAAT TATTTGAAGAAAACCCTATTAACGCAAGTGGAGTAGATGCTAAAGCGATTCTTTCTGCA CGATTGAGTAAATCAAGACGATTAGAAAATCTCATTGCTCAGCTCCCCGGTGAGAAGAA AAATGGCTTATTTGGGAATCTCATTGCTTTGTCATTGGGTTTGACCCCTAATTTTAAATCA AATTTTGATTTGGCAGAAGATGCTAAATTACAGCTTTCAAAAGATACTTACGATGATGAT

		TTAGATAATTTATTGGCGCAAATTGGAGATCAATATGCTGATTTGTTTTTGGCAGCTAAG AATTTATCAGATGCTATTTTACTTTCAGATATCCTAAGAGTAAATACTGAAATAACTAAG GCTCCCCTATCAGCTTCAATGATTAAACGCTACGATGAACATCATCAAGACTTGACTCTT TTAAAAGCTTTAGTTCGACAACAACTTCCAGAAAAGTATAAAGAAATCTTTTTTGATCAA TCAAAAAACGGATATGCAGGTTATATTGATGGGGGAGCGAGCCAAGAAGAATTTTATAA ATTTATCAAACCAATTTTAGAAAAAATGGATGGTACTGAGGAATTATTGGTGAAACTAA ATCGTGAAGATTTGCTGCGCAAGCAACGGACCTTTGACAACGGCTCTATTCCCCATCAA ATTCACTTGGGTGAGCTGCATGCTATTTTGAGAAGACAAGAAGACTTTTATCCATTTTTA AAAGACAATCGTGAGAAGATTGAAAAAATCTTGACTTTTCGAATTCCTTATTATGTTGGT CCATTGGCGCGTGGCAATAGTCGTTTTGCATGGATGACTCGGAAGTCTGAAGAAACAAT TACCCCATGGAATTTTGAAGAAGTTGTCGATAAAGGTGCTTCAGCTCAATCATTTATTGA ACGCATGACAAACTTTGATAAAAATCTTCCAAATGAAAAAGTACTACCAAAACATAGTT TGCTTTATGAGTATTTTACGGTTTATAACGAATTGACAAAGGTCAAATATGTTACTGAAG GAATGCGAAAACCAGCATTTCTTTCAGGTGAACAGAAGAAAGCCATTGTTGATTTACTC TTCAAAACAAATCGAAAAGTAACCGTTAAGCAATTAAAAGAAGATTATTTCAAAAAAAT AGAATGTTTTGATAGTGTTGAAATTTCAGGAGTTGAAGATAGATTTAATGCTTCATTAGG TACCTACCATGATTTGCTAAAAATTATTAAAGATAAAGATTTTTTGGATAATGAAGAAA ATGAAGATATCTTAGAGGATATTGTTTTAACATTGACCTTATTTGAAGATAGGGAGATGA TTGAGGAAAGACTTAAAACATATGCTCACCTCTTTGATGATAAGGTGATGAAACAGCTT AAACGTCGCCGTTATACTGGTTGGGGACGTTTGTCTCGAAAATTGATTAATGGTATTAGG GATAAGCAATCTGGCAAAACAATATTAGATTTTTTGAAATCAGATGGTTTTGCCAATCGC AATTTTATGCAGCTGATCCATGATGATAGTTTGACATTTAAAGAAGACATTCAAAAAGC ACAAGTGTCTGGACAAGGCGATAGTTTACATGAACATATTGCAAATTTAGCTGGTAGCC CTGCTATTAAAAAAGGTATTTTACAGACTGTAAAAGTTGTTGATGAATTGGTCAAAGTA ATGGGGCGGCATAAGCCAGAAAATATCGTTATTGAAATGGCACGTGAAAATCAGACAA CTCAAAAGGGCCAGAAAAATTCGCGAGAGCGTATGAAACGAATCGAAGAAGGTATCAA AGAATTAGGAAGTCAGATTCTTAAAGAGCATCCTGTTGAAAATACTCAATTGCAAAATG AAAAGCTCTATCTCTATTATCTCCAAAATGGAAGAGACATGTATGTGGACCAAGAATTA GATATTAATCGTTTAAGTGATTATGATGTCGATGCCATTGTTCCACAAAGTTTCCTTAAA GACGATTCAATAGACAATAAGGTCTTAACGCGTTCTGATAAAAATCGTGGTAAATCGGA TAACGTTCCAAGTGAAGAAGTAGTCAAAAAGATGAAAAACTATTGGAGACAACTTCTAA ACGCCAAGTTAATCACTCAACGTAAGTTTGATAATTTAACGAAAGCTGAACGTGGAGGT TTGAGTGAACTTGATAAAGCTGGTTTTATCAAACGCCAATTGGTTGAAACTCGCCAAATC ACTAAGCATGTGGCACAAATTTTGGATAGTCGCATGAATACTAAATACGATGAAAATGA TAAACTTATTCGAGAGGTTAAAGTGATTACCTTAAAATCTAAATTAGTTTCTGACTTCCG AAAAGATTTCCAATTCTATAAAGTACGTGAGATTAACAATTACCATCATGCCCATGATGC GTATCTAAATGCCGTCGTTGGAACTGCTTTGATTAAGAAATATCCAAAACTTGAATCGGA GTTTGTCTATGGTGATTATAAAGTTTATGATGTTCGTAAAATGATTGCTAAGTCTGAGCA AGAAATAGGCAAAGCAACCGCAAAATATTTCTTTTACTCTAATATCATGAACTTCTTCAA AACAGAAATTACACTTGCAAATGGAGAGATTCGCAAACGCCCTCTAATCGAAACTAATG GGGAAACTGGAGAAATTGTCTGGGATAAAGGGCGAGATTTTGCCACAGTGCGCAAAGT ATTGTCCATGCCCCAAGTCAATATTGTCAAGAAAACAGAAGTACAGACAGGCGGATTCT CCAAGGAGTCAATTTTACCAAAAAGAAATTCGGACAAGCTTATTGCTCGTAAAAAAGAC TGGGATCCAAAAAAATATGGTGGTTTTGATAGTCCAACGGTAGCTTATTCAGTCCTAGTG GTTGCTAAGGTGGAAAAAGGGAAATCGAAGAAGTTAAAATCCGTTAAAGAGTTACTAG GGATCACAATTATGGAAAGAAGTTCCTTTGAAAAAAATCCGATTGACTTTTTAGAAGCT AAAGGATATAAGGAAGTTAAAAAAGACTTAATCATTAAACTACCTAAATATAGTCTTTT TGAGTTAGAAAACGGTCGTAAACGGATGCTGGCTAGTGCCGGAGAATTACAAAAAGGA AATGAGCTGGCTCTGCCAAGCAAATATGTGAATTTTTTATATTTAGCTAGTCATTATGAA AAGTTGAAGGGTAGTCCAGAAGATAACGAACAAAAACAATTGTTTGTGGAGCAGCATA AGCATTATTTAGATGAGATTATTGAGCAAATCAGTGAATTTTCTAAGCGTGTTATTTTAG CAGATGCCAATTTAGATAAAGTTCTTAGTGCATATAACAAACATAGAGACAAACCAATA CGTGAACAAGCAGAAAATATTATTCATTTATTTACGTTGACGAATCTTGGAGCTCCCGCT GCTTTTAAATATTTTGATACAACAATTGATCGTAAACGATATACGTCTACAAAAGAAGTT TTAGATGCCACTCTTATCCATCAATCCATCACTGGTCTTTATGAAACACGCATTGATTTG AGTCAGCTAGGAGGTGACTAATAGGGGCAGCAGCCATCACCATCATCACCACAGCCAGG CTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTG AGGGGTTTTTTGCTGAAACCTCAGGCATTTGAGAAGCACACGGTCACACTGCTTCCGGTA GTCAATAAACCGGTAAACCAGCAATAGACATAAGCGGCTATTTAACGACCCTGCCCTGA ACCGACGACCGGGTCGAATTTGCTTTCGAATTTCTGCCATTCATCCGCTTATTATCACTTA TTCAGGCGTAGCACCAGGCGTTTAAGGGCACCAATAACTGCCTTAAAAAAATTACGCCC CGCCCTGCCACTCATCGCAGTACTGTTGTAATTCATTAAGCATTCTGCCGACATGGAAGC CATCACAGACGGCATGATGAACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCTTGC GTATAATATTTGCCCATAGTGAAAACGGGGGCGAAGAAGTTGTCCATATTGGCCACGTT TAAATCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAAAAACATATTCTCAA TAAACCCTTTAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTTGCGAATAT ATGTGTAGAAACTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCGATGAAAACGTTTC AGTTTGCTCATGGAAAACGGTGTAACAAGGGTGAACACTATCCCATATCACCAGCTCAC CGTCTTTCATTGCCATACGGAACTCCGGATGAGCATTCATCAGGCGGGCAAGAATGTGA

		ATAAAGGCCGGATAAAACTTGTGCTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTAATA TCCAGCTGAACGGTCTGGTTATAGGTACATTGAGCAACTGACTGAAATGCCTCAAAATG TTCTTTACGATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTTTTTTCTCCATT TTAGCTTCCTTAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGGTAGTGATCTT ATTTCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCAACGTCTCATTTTCGCCAA AAGTTGGCCCAGGGCTTCCCGGTATCAACAGGGACACCAGGATTTATTTATTCTGCGAA GTGATCTTCCGTCACAGGTATTTATTCGGCGCAAAGTGCGTCGGGTGATGCTGCCAACTT ACTGATTTAGTGTATGATGGTGTTTTTGAGGTGCTCCAGTGGCTTCTGTTTCTATCAGCTG TCCCTCCTGTTCAGCTACTGACGGGGTGGTGCGTAACGGCAAAAGCACCGCCGGACATC AGCTCTAGCGGAGTGTATACTGGCTTACTATGTTGGCACTGATGAGGGTGTCAGTGAAG TGCTTCATGTGGCAGGAGAAAAAAGGCTGCACCGGTGCGTCAGCAGAATATGTGATACA GGATATATTCCGCTTCCTCGCTCACTGACTCGCTACGCTCGGTCGTTCGACTGCGGCGAG CGGAAATGGCTTACGAACGGGGCGGAGATTTCCTGGAAGATGCCAGGAAGATACTTAAC AGGGAAGTGAGAGGGCCGCGGCAAAGCCGTTTTTCCATAGGCTCCGCCCCCCTGACAAG CATCACGAAATCTGACGCTCAAATCAGTGGTGGCGAAACCCGACAGGACTATAAAGATA CCAGGCGTTTCCCCTGGCGGCTCCCTCGTGCGCTCTCCTGTTCCTGCCTTTCGGTTTACCG GTGTCATTCCGCTGTTATGGCCGCGTTTGTCTCATTCCACGCCTGACACTCAGTTCCGGGT AGGCAGTTCGCTCCAAGCTGGACTGTATGCACGAACCCCCCGTTCAGTCCGACCGCTGC GCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGAAAGACATGCAAAAGCACCACT GGCAGCAGCCACTGGTAATTGATTTAGAGGAGTTAGTCTTGAAGTCATGCGCCGGTTAA GGCTAAACTGAAAGGACAAGTTTTGGTGACTGCGCTCCTCCAAGCCAGTTACCTCGGTTC AAAGAGTTGGTAGCTCAGAGAACCTTCGAAAAACCGCCCTGCAAGGCGGTTTTTTCGTT TTCAGAGCAAGAGATTACGCGCAGACCAAAACGATCTCAAGAAGATCATCTTATTAATC AGATAAAATATTT

Table 3.6: Plasmid sequences used in this study

Chapter 4

Conclusion

† The work presented in this chapter has previously been published in the following review article: Higgins, S.A., and Savage, D.F. (2018). Protein Science by DNA Sequencing: How Advances in Molecular Biology Are Accelerating Biochemistry. Biochemistry. 57 (1), 38-46.

4.1 Summary

As experimental throughput has increased from point mutation (Flavell et al. 1975) to alanine scanning (Cunningham and Wells 1989) and finally to scans of protein variants encompassing all possible single mutations (Gold et al. 2006), comprehensive fitness landscapes are now being generated to explore residues important for protein stability, function, and structure (Fowler and Fields 2014; Rocklin et al. 2017). These massive datasets promise to greatly advance many fields, from fundamental protein biochemistry to disease prediction and protein engineering.

A key observation from these studies is the fact that proteins are robust to a large number of different single mutations (Bershtein et al. 2006). Multiple mutations, however, often display negative epistasis in accordance with a threshold robustness model (Tokuriki and Tawfik 2009) of protein stability, whereby the overall protein structure can tolerate a threshold loss of free energy over which negative fitness effects become exponentially severe (Olson, Wu, and Sun 2014). Epistatic interactions are exceedingly difficult to predict, posing a barrier to personalized medicine (Shendure and Fields 2016; Weile et al. 2017; Starita et al. 2015; Majithia et al. 2016). At the same time, protein engineering has long sought to develop new tools in areas such as molecular recognition and catalyst engineering (Mandecki 1998; Schmid et al. 2001; Fox et al. 2007; Wrenbeck, Faber, and Whitehead 2017), where the highest fitness variants are often several mutations away from the starting sequence. Fundamentally, surveying protein fitness landscapes requires mutagenesis methods that are both high throughput enough to be comprehensive and programmable in order to define the specific type of landscape to be examined. Techniques based on PCR require *in vitro* reactions involving specialized protocols and reagents. This has complicated efforts to rapidly and reliably produce desired protein libraries.

Here we have demonstrated that plasmid recombineering (PR) is a simple and robust *in vivo* method for constructing comprehensive libraries of protein mutations. The method requires a single day and is performed in a single tube. The resulting libraries can comprehensively evaluate amino acid substitutions, as shown here for iLOV, with moderate efficiency even without selection. As in previous reports of genomic recombineering (H. Wang et al. 2009), this efficiency could be improved to nearly three quarters after five rounds of PR. The libraries also exhibited moderate bias related to the mechanism of oligonucleotide recombineering, but this bias could be ameliorated by altering ratios of oligonucleotides such that the comprehensive mutagenesis library was effectively uniform for screening purposes. It will be important to understand this mechanism in order to predict efficiencies rather than rely on empirical data as used here. Regardless of these modifications, the reagent cost and experimental effort remain low – standard 60 bp oligonucleotides and simple cycles of electroporation, growth, and plasmid isolation.

The other key aspect to ideal mutagenesis methods is programmability. Here, we demonstrate the PR is capable of evaluating much more focused fitness landscapes. The experimental design is nearly identical to a comprehensive case, with the exception that care might be taken to use recombineering oligonucleotides that minimize the possibility for ‘overwriting’ each other’s mutation. This can be accomplished either by shifting the exact location of the homology arms, as was done here, or by including multiple mutations in a single oligonucleotide. To demonstrate programmability, we first comprehensively explored the iLOV single mutation sequence-space for thermostability. We chose 25 fitness enhancing mutations

and demonstrated the utility of PR for advanced protein engineering by multiplexing many different single and double mutations at discontinuous sites across iLOV in a second library. The ability to select and easily mutate numerous specific and non-contiguous locations across a protein is highly useful for a variety of techniques that utilize experimental or phylogenetic data to computationally predict and enhance enzymes (Heinzelman et al. 2009), explore epistatic interactions (Olson, Wu, and Sun 2014), or even scan SNPs in human proteins for disease prediction (Majithia et al. 2016). The best of the initial mutations were found to improve the thermal stability of iLOV up to 10° C increase in Tm. We found that programmably focused combinations of these mutations could yield doubly improved variants, up to 20° C. In summary, we found that PR was suited for the construction of both comprehensive and targeted libraries, and that the simplicity of the protocol led to rapid and reliable screening experiments. In particular, PR is suitable for cycles of iterative design, construction, and sampling of genetic libraries, requiring no specialized reagents or protocols.

Beyond simple amino acid substitutions, protein topology is also well-established as a key mechanism by which large, complex multi-domain proteins evolve highly specialized functions (Fong et al. 2007). While rationally constructed protein deletions have long been essential to elucidating biochemical properties, previous techniques have been insufficient for a comprehensive approach. Here we have developed a method for constructing fitness landscapes for even the largest and most complex proteins, building upon the PR platform. We comprehensively surveyed functional deletion landscape of the RNA-guided DNA binding protein dCas9, the foundation for powerful genome editing and modifying technologies (Qi et al., 2013; Oakes et al., 2014). CRISPR proteins are highly complex with numerous distinct domains responsible for activities such as guide RNA binding, DNA recognition, DNA unwinding, specificity sensing and ultimately the cleavage of each DNA strand (Jinek et al. 2012; Nishimasu et al. 2014; Anders et al. 2014; Chen et al. 2017). We interpret this landscape in the context of known deletion function, uncovering both (1) a previously unknown domain region that is dispensable for DNA binding as well as (2) a potential role in DNA binding for another domain thought to be dispensable for this function. Furthermore, we exploited this fitness landscape to revert functionality and step backward in domain evolution, comprehensively minimizing dCas9 and screening for an essential function. We have demonstrated the power of this technique by revealing the minimal RNA guided DNA binding module at 64% of the full CRISPR-Cas9 platform, providing many new opportunities for fusions and delivery. These results highlight the power of comprehensive protein deletions to clearly elucidate the boundaries of a central function.

4.2 Applications and future directions

4.2a Fundamental Protein Characteristics

Comprehensive maps are now being generated to explore residues important for protein stability, function, and structure (Fowler and Fields 2014). Protein stability in particular has benefited from insights gained specifically due to the availability of large scale data (Rocklin et al. 2017). As introduced above, early studies that sought to explore large fractions of protein sequence space unexpectedly found that proteins are robust to a large number of different single mutations (Bershtein et al. 2006). Multiple deleterious mutations, however, were found to produce even worse fitness than the sum of individual mutations (i.e., display negative epistasis).

This observation has since been incorporated into a general theory of protein evolvability known as the threshold robustness model (Tokuriki and Tawfik 2009). Briefly, a wild-type protein will tend to have a margin of stability such that the vast majority of the population ensemble will exist in a functional conformation. In some cases, this margin allows the protein to absorb a small functional instability caused by a deleterious mutation, somewhat independently of the specific amino acid change, resulting in little loss of function. Multiple such deleterious mutations, however, will result in the margin being exhausted and a rapid decrease in fitness/function is observed. Deep mutational scans continue to produce data compatible with this theory (Olson, Wu, and Sun 2014).

Sequence determinants of protein function also stand to gain from large datasets. For example, determining the substrate specificity of an enzyme is a common question. A recent comprehensive mutational analysis of AmiE for three amide substrates revealed that substrate specificity was not solely encoded by residues near the active site (Wrenbeck, Azouz, and Whitehead 2017). Additionally, substrate-specific beneficial mutations could not be predicted based on known fitness towards another substrate. These results agree with other deep mutational scans suggesting that fitness landscapes exhibit substantial substrate dependence (Melnikov et al. 2014).

Large mutation-function datasets can also enable enhanced protein structural prediction. It may be possible to experimentally identify co-varying residues which are nearby in the protein structure (Fowler and Fields 2014), and this information has been shown to enhance structural models produced by prediction software such as Rosetta (Kim et al. 2014). It is thus likely that the synthesis of computation and large datasets will improve the predictability of protein modifications. De novo protein design has already successfully generated a variety of stable structures, although challenges remain with solubility, oligomerization, and the design of specific functions or dynamics (Huang, Boyken, and Baker 2016).

4.2b Disease Prediction

Mutations with unknown functional effect, also known as "variants of uncertain significance," pose a barrier to the development of personalized medicine (Shendure and Fields 2016). Despite thousands of known common variants that reproducibly associate with disease, the causal variant is rarely understood. Deep mutational scans present one path toward predicting the functional consequences of these variants, with some proposals even envisioning an understanding of the comprehensive human SNP-ome (Weile et al. 2017).

Deep mutational scanning experiments have begun to outperform other predictive approaches in some cases. For example, a scan of 2000 substitutions in the BRCA1 RING domain enabled a quantitative map of effects on E3 ubiquitin ligase activity and BARD1 RING domain binding activity (Starita et al. 2015). These functional scores were used to generate a predictive model for BRCA1 variants' capacity for tumor suppression through homology-directed DNA repair. This work culminated in the creation of a mutational map of BRCA1 more accurate than those produced using computational tools. Similarly, a library of all possible single amino acid mutations has been created and analyzed for peroxisome proliferator-activated receptor gamma in order to annotate variants of unknown consequence for type 2 diabetes (Majithia et al. 2016). Additional prediction maps are likely to follow for other genes involved in disease provided an appropriate functional assay can be developed.

Finally, the exploration of protein sequence-space can potentially predict evolutionary outcomes if combined with a physiological model. For example, mutations to proteins involved in antibiotic resistance that produce even small changes to a Km (less than twofold) are sufficient to yield highly successful adaptive mutants, and these results can be integrated into a model relating in vitro protein properties to bacterial growth rates (Walkiewicz et al. 2012). Additionally, allele frequency measurements in continuous culture experiments can reveal hierarchy and order in genetic changes, identifying new targets for pathogen drug design – particularly for antibiotic resistance (Miller et al. 2013).

4.2c Protein Engineering

Deep mutational scanning experiments have achieved success in molecular recognition and enzyme engineering, and a recent review highlights several of these studies (Wrenbeck, Faber, and Whitehead 2017). Successful engineering is often linked to the ease and quality of the assay. Binding assays, in particular, possess many ideal properties including high throughput, low cost, and ease of experimentation due to direct linkage between protein function (binding) and its sequence via DNA sequence. A variety of linkage mechanisms have been developed, including yeast display (Cherf and Cochran 2015), phage display (Wu et al. 2016), and ribosome display (Hanes and Plückthun 1997). Improving enzyme catalysis (e.g. rate, specificity, stability) represents another important goal for protein engineering, due to the promise of biocatalysis in chemical applications (Schmid et al. 2001). Enzyme assays are more complicated to develop than those for binding, but recent studies have found success in physiologically linking enzyme function to growth, as discussed above. Screens can be employed for enzymes with no connection to physiology, but at the cost of reduced bandwidth.

In contrast to comprehensive mutagenesis experiments, directed evolution can be pursued when the goal is simply to generate an improved protein rather than exhaustively evaluate sequence-function relationships. Fundamentally, exponential growth in combinatorial sequence space (Mandecki 1998) limits comprehensive saturation mutagenesis libraries to only a few simultaneous mutations (practically speaking, no more than two). At the same time, the highest fitness variants are often several mutations away from the starting sequence. Directed evolution makes use of iterative rounds of mutation and selection to explore sequence space incrementally further from the wild-type sequence (P. a Romero and Arnold 2009). Like above, DNA sequencing can be used guide this process. One such approach used statistical analysis of protein sequence activity relationships (ProSAR) to identify a 4,000-fold improvement in the performance of an industrially relevant biocatalyst, where the final improved variants possessed 35+ mutations (Fox et al. 2007).

In contrast to statistically-guided directed evolution, an alternative approach known as continuous evolution seeks to conduct the cycles of mutagenesis and selection in vivo such that the process becomes automated within the cell. These systems potentially provide the highest possible throughput so as to maximize the likelihood of discovering rare, interesting variants. As discussed above, several techniques have been developed to localize mutation to a single target locus, including an orthogonal error-prone polymerase (Fabret et al. 2000), a targeting glycosylase (Finney-Manchester and Maheshri 2013), an error-prone Ty1 reverse transcriptase (Crook et al. 2016), PACE (Esvelt, Carlson, and Liu 2011), or a cytidine deaminase-Cas9 fusion (Komor et al. 2016). PACE is one of the highest throughput directed evolution methods yet developed, achieving hundreds of rounds of protein evolution on the week timescale (Esvelt,

Carlson, and Liu 2011). PACE was first used to rapidly evolve T7 RNA polymerase variants capable of binding distinct promoters or nucleotides, and has since been improved for easier tuning of selection stringency and negative selection against unwanted activities (Carlson et al. 2014). The drawbacks to continuous evolution systems include the need for careful customization and validation, but more importantly these systems currently lack specificity in library creation and do not cover the full codon mutational space (Kitzman et al. 2015). Innovations in in vivo mutagenesis techniques will be needed to overcome this limitation.

Understanding the sequence determinants for desired protein characteristics will enable directed evolution of protein libraries guided and/or designed by computational methods. For example, the SCHEMA computational algorithm identifies fragments of proteins that can be recombined with minimal disruption to protein structure (Voigt et al. 2002), and properties of recombined enzymes such as stability can be predicted using models that sample these chimeras (Li et al. 2007). Additional work has developed models that investigate the 'recombinational landscape' as a whole, finding enrichment of functional sequences and additive properties of independent sequence elements (P. A. Romero and Arnold 2012). Fruitful areas of sequence-space can also be identified through iterative machine learning and massively parallel experimentation.

Computational prediction can also be used to narrow the search space required for experimental validation. The Baker group has developed an assay to quantify stability for entire libraries of proteins, consisting of proteolytic cleavage of a fluorescent tag from the surface of yeast (Rocklin et al. 2017). This assay allowed the assignment of a stability score to each of thousands of designed miniproteins. Various topologies were targeted for design using approaches developed previously (Koga et al. 2012) and thousands of diverse sequence variants were generated for testing. Analysis comparing the most and least stable designed proteins indicated that total buried nonpolar surface area was a key determinant of stability. Adjusting design parameters for this finding resulted in a much larger fraction of successful designs. These findings demonstrate that specific library characteristics can be tuned to search areas of sequence space that are more likely to produce improved variants. Moreover, analysis of high-throughput data can uncover specific relationships between sequence and function, findings useful in protein engineering and protein biochemistry alike.

4.3 Outlook

Together, amino acid substitution and topological mutation (encompassing deletions, insertions, and circular permutations) comprise the fundamental units of protein mutation. This work has served to develop simple and robust methods, which remain programmable and comprehensive, for both substitution and topological mutagenesis. In the process, novel proteins have been developed for both iLOV and dCas9, but more importantly these mutagenesis methods can be applied to any DNA sequence. A central theme in this work has been massively parallel approaches in light of insufficient predictive power. As computational protein design becomes more robust, our predictive power may become sufficient to ignore certain substitutions or topologies in pursuit of desired proteins. In this case, programmability will become even more important so as to avoid fruitless combinations that take up valuable experimental throughput. Ultimately, understanding the general principles of protein sequence-function landscapes - enabled by massively parallel experimentation - will allow computational methods to synergize with programmable mutagenesis and vastly improve the search for novel fitness variants.

5.1 Bibliography

Agresti, Jeremy J, Eugene Antipov, Adam R Abate, Keunho Ahn, Amy C Rowat, Jean-Christophe Baret, Manuel Marquez, Alexander M Klibanov, Andrew D Griffiths, and David A Weitz. 2010. "Ultrahigh-Throughput Screening in Drop-Based Microfluidics for Directed Evolution." Proceedings of the National Academy of Sciences of the United States of America 107 (9): 4004–9. doi:10.1073/pnas.0910781107.

Anders, Carolin, Ole Niewoehner, Alessia Duerst, and Martin Jinek. 2014. "Structural Basis of PAM-Dependent Target DNA Recognition by the Cas9 Endonuclease." Nature 513 (7519): 569–73. doi.org/10.1038/nature13579.

Araya, Carlos L., and Douglas M. Fowler. 2011. "Deep Mutational Scanning:assessing Protein Function on a Massive Scale." Trends in Biotechnology 29 (9): 435–42. doi:10.1016/j.tibtech.2011.04.003.Deep.

Araya, Carlos L., and Douglas M. Fowler. 2011. "Deep Mutational Scanning: Assessing Protein Function on a Massive Scale." Trends in Biotechnology 29 (9): 435–42. doi.org/10.1016/j.tibtech.2011.04.003.

Aroul-Selvam, R., Tim Hubbard, and Rajkumar Sasidharan. 2004. "Domain Insertions in Protein Structures." Journal of Molecular Biology 338 (4): 633–41. doi:10.1016/j.jmb.2004.03.039.

Arpino, James a J, Samuel C. Reddington, Lisa M. Halliwell, Pierre J. Rizkallah, and D. Dafydd Jones. 2014. "Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure." Structure 22 (6). The Authors: 889–98. doi:10.1016/j.str.2014.03.014.

Arpino, James A. J., Samuel C. Reddington, Lisa M. Halliwell, Pierre J. Rizkallah, and D. Dafydd Jones. 2014. "Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure." Structure 22 (6): 889–98. doi.org/10.1016/j.str.2014.03.014.

Atkinson, Joshua T., Alicia M. Jones, Quan Zhou, and Jonathan J. Silberg. 2018. "Circular Permutation Profiling by Deep Sequencing Libraries Created Using Transposon Mutagenesis." Nucleic Acids Research 46 (13): e76. doi.org/10.1093/nar/gky255.

Belsare, Ketaki Deepak, Mary C. Andorfer, Frida Cardenas, Julia R. Chael, Hyun June Park, and Jared C. Lewis. 2016. "A Simple Combinatorial Codon Mutagenesis Method for Targeted Protein Engineering." ACS Synthetic Biology 6: 416–20. doi:10.1021/acssynbio.6b00297.

Bershtein, Shimon, Michal Segal, Roy Bekerman, Nobuhiko Tokuriki, and Dan S Tawfik. 2006. “Robustness-Epistasis Link Shapes the Fitness Landscape of a Randomly Drifting Protein.” Nature 444 (December): 929–32. doi:10.1038/nature05385.

Bhattacharyya, Roby P, Attila Reményi, Brian J Yeh, and Wendell A Lim. 2006. “Domains, Motifs, and Scaffolds: The Role of Modular Interactions in the Evolution and Wiring of Cell Signaling Circuits.” Annual Review of Biochemistry 75: 655–80. doi:10.1146/annurev.biochem.75.103004.142710.

Björklund, Asa K., Diana Ekman, Sara Light, Johannes Frey-Skött, and Arne Elofsson. 2005. “Domain Rearrangements in Protein Evolution.” Journal of Molecular Biology 353 (4): 911–23. doi.org/10.1016/j.jmb.2005.08.067.

Brocchieri, Luciano, and Samuel Karlin. 2005. “Protein Length in Eukaryotic and Prokaryotic Proteomes.” Nucleic Acids Research 33 (10): 3390–3400. doi.org/10.1093/nar/gki615.

Callis, Patrik R., and Tiqing Liu. 2006. “Short Range Photoinduced Electron Transfer in Proteins: QM-MM Simulations of Tryptophan and Flavin Fluorescence Quenching in Proteins.” Chemical Physics 326 (1): 230–39. doi:10.1016/j.chemphys.2006.01.039.

Carlson, J C, A H Badran, D A Guggiana-Nilo, and D R Liu. 2014. “Negative Selection and Stringency Modulation in Phage-Assisted Continuous Evolution.” Nat Chem Biol 10 (3): 216–22. doi:10.1038/nchembio.1453.

Carr, Peter A, and George M Church. 2009. “Genome Engineering.” Nat. Biotechnol. 27 (12). Nature Publishing Group: 1151–62. doi:10.1038/nbt.1590.

Chapman, Sean, Christine Faulkner, Eirini Kaiserli, Carlos Garcia-Mata, Eugene I Savenkov, Alison G Roberts, Karl J Oparka, and John M Christie. 2008. “The Photoreversible Fluorescent Protein iLOV Outperforms GFP as a Reporter of Plant Virus Infection.” Proceedings of the National Academy of Sciences of the United States of America 105: 20038–43. doi:10.1073/pnas.0807551105.

Cheltsov, Anton V., Michael J. Barber, and Gloria C. Ferreira. 2001. “Circular Permutation of 5-Aminolevulinate Synthase: Mapping the Polypeptide Chain to Its Function.” Journal of Biological Chemistry 276 (22): 19141–49. doi:10.1074/jbc.M100329200.

Chen, Janice S., Yavuz S. Dagdas, Benjamin P. Kleinstiver, Moira M. Welch, Alexander A. Sousa, Lucas B. Harrington, Samuel H. Sternberg, J. Keith Joung, Ahmet Yildiz, and Jennifer A. Doudna. 2017. “Enhanced Proofreading Governs CRISPR-Cas9 Targeting Accuracy.” Nature 550 (7676): 407–10. doi.org/10.1038/nature24268.

Cherf, Gerald M, and Jennifer R Cochran. 2015. “Applications of Yeast Surface Display for Protein Engineering.” Methods Mol Biol. 1319: 155–75. doi:10.1007/978-1-4939-2748-7.

Christie, John M., Kenichi Hitomi, Andrew S. Arvai, Kimberly a. Hartfield, Marcel Mettlen, Ashley J. Pratt, John a. Tainer, and Elizabeth D. Getzoff. 2012. "Structural Tuning of the Fluorescent Protein iLOV for Improved Photostability." Journal of Biological Chemistry 287 (26): 22295–304. doi:10.1074/jbc.M111.318881.

Copeland, N G, N a Jenkins, and D L Court. 2001. "Recombineering: A Powerful New Tool for Mouse Functional Genomics." Nature Reviews. Genetics 2 (10): 769–79. doi:10.1038/35093556.

Copeland, N G, N a Jenkins, and D L Court. 2001. "Recombineering: A Powerful New Tool for Mouse Functional Genomics." Nature Reviews. Genetics 2 (10): 769–79. doi:10.1038/35093556.

Crook, Nathan, Joseph Abatemarco, Jie Sun, James M. Wagner, Alexander Schmitz, and Hal S. Alper. 2016. "In Vivo Continuous Evolution of Genes and Pathways in Yeast." Nature Communications 7: 13051. doi:10.1038/ncomms13051.

Cunningham, B. C., and J. A. Wells. 1989. "High-Resolution Epitope Mapping of hGH-Receptor Interactions by Alanine-Scanning Mutagenesis." Science 244: 1081–85.

Edwards, Wayne R., Abigail J. Williams, Josephine L. Morris, Amy J. Baldwin, Rudolf K. Allemann, and D. Dafydd Jones. 2010. "Regulation of β-Lactamase Activity by Remote Binding of Heme: Functional Coupling of Unrelated Proteins through Domain Insertion." Biochemistry 49 (31): 6541–49. doi:10.1021/bi100793y.

Ellefson, Jared W, Adam J Meyer, Randall a Hughes, Joe R Cannon, Jennifer S Brodbelt, and Andrew D Ellington. 2014. "Directed Evolution of Genetic Parts and Circuits by Compartmentalized Partnered Replication." Nature Biotechnology 32 (1). Nature Publishing Group: 97–101. doi:10.1038/nbt.2714.

Engler, Carola, Romy Kandzia, and Sylvestre Marillonnet. 2008. "A One Pot, One Step, Precision Cloning Method with High Throughput Capability." PloS One 3 (11): e3647. doi:10.1371/journal.pone.0003647.

Esvelt, Kevin M, and Harris H Wang. 2013. "Genome-Scale Engineering for Systems and Synthetic Biology." Molecular Systems Biology 9 (641). Nature Publishing Group: 641. doi:10.1038/msb.2012.66.

Esvelt, Kevin M, Jacob C Carlson, and David R Liu. 2011. "A System for the Continuous Directed Evolution of Biomolecules." Nature 472 (7344). Nature Publishing Group: 499–503. doi:10.1038/nature09929.

Fabret, C, S Poncet, S Danielsen, T V Borchert, S D Ehrlich, and L Jannière. 2000. "Efficient Gene Targeted Random Mutagenesis in Genetically Stable Escherichia Coli Strains." Nucleic Acids Research 28 (21): E95.

Finney-Manchester, Shawn P., and Narendra Maheshri. 2013. “Harnessing Mutagenic Homologous Recombination for Targeted Mutagenesis in Vivo by TaGTEAM.” Nucleic Acids Research 41 (9): 1–10. doi:10.1093/nar/gkt150.

Firnberg, Elad, and Marc Ostermeier. 2012. “PFunkel: Efficient, Expansive, User-Defined Mutagenesis.” PLoS ONE 7 (12): 1–10. doi:10.1371/journal.pone.0052031.

Firnberg, Elad, and Marc Ostermeier. 2012. “PFunkel: Efficient, Expansive, User-Defined Mutagenesis.” PLoS ONE 7 (12): e52031. doi:10.1371/journal.pone.0052031.

Firnberg, Elad, Jason W. Labonte, Jeffrey J. Gray, and Marc Ostermeier. 2014. “A Comprehensive, High-Resolution Map of a Gene’s Fitness Landscape.” Molecular Biology and Evolution 31 (6): 1581–92. doi:10.1093/molbev/msu081.

Firnberg, Elad, Jason W. Labonte, Jeffrey J. Gray, and Marc Ostermeier. 2014. “A Comprehensive, High-Resolution Map of a Gene’s Fitness Landscape.” Molecular Biology and Evolution 31 (6): 1581–92. doi:10.1093/molbev/msu081.

Flavell, R A, D L Sabo, E F Bandle, and C Weissmann. 1975. “Site-Directed Mutagenesis: Effect of an Extracistronic Mutation on the in Vitro Propagation of Bacteriophage Qbeta RNA.” Proceedings of the National Academy of Sciences of the United States of America 72 (1): 367–71. doi:10.1073/pnas.72.1.367.

Fong, Jessica H., Lewis Y. Geer, Anna R. Panchenko, and Stephen H. Bryant. 2007. “Modeling the Evolution of Protein Domain Architectures Using Maximum Parsimony.” Journal of Molecular Biology 366 (1): 307–15. doi.org/10.1016/j.jmb.2006.11.017.

Fowler, Douglas M, and Stanley Fields. 2014. “Deep Mutational Scanning: A New Style of Protein Science.” Nature Methods 11 (8): 801–7. doi:10.1038/nmeth.3027.

Fowler, Douglas M, Carlos L Araya, Sarel J Fleishman, Elizabeth H Kellogg, Jason J Stephany, David Baker, and Stanley Fields. 2010. “High-Resolution Mapping of Protein Sequence-Function Relationships.” Nature Methods 7 (9). Nature Publishing Group: 741–46. doi:10.1038/nmeth.1492.

Fowler, Douglas M., and Stanley Fields. 2014. “Deep Mutational Scanning: A New Style of Protein Science.” Nature Methods 11 (8): 801–7. doi.org/10.1038/nmeth.3027.

Fowler, Douglas M., Carlos L. Araya, Wayne Gerard, and Stanley Fields. 2011. “Enrich: Software for Analysis of Protein Function by Enrichment and Depletion of Variants.” Bioinformatics 27 (24): 3430–31. doi:10.1093/bioinformatics/btr577.

Fox, Richard J, S Christopher Davis, Emily C Mundorff, Lisa M Newman, Vesna Gavrilovic, Steven K Ma, Loleta M Chung, et al. 2007. “Improving Catalytic Function by ProSAR-Driven Enzyme Evolution.” Nature Biotechnology 25 (3): 338–44. doi:10.1038/nbt1286.

Garst, Andrew D, Marcelo C Bassalo, Gur Pines, Sean a Lynch, Andrea L Halweg-Edwards, Rongming Liu, Liya Liang, et al. 2016. "Genome-Wide Mapping of Mutations at Single-Nucleotide Resolution for Protein, Metabolic and Genome Engineering." Nature Biotechnology 35 (1). Nature Publishing Group: 48–55. doi:10.1038/nbt.3718.

Gilbert, W. 1987. "The Exon Theory of Genes." Cold Spring Harbor Symposia on Quantitative Biology LII: 901–5.

Gold, Matthew G., Birgitte Lygren, Pawel Dokurno, Naoto Hoshi, George McConnachie, Kjetil Taskén, Cathrine R. Carlson, John D. Scott, and David Barford. 2006. "Molecular Basis of AKAP Specificity for PKA Regulatory Subunits." Molecular Cell 24 (3): 383–95. doi:10.1016/j.molcel.2006.09.006.

Goodwin, Sara, John D Mcpherson, and W Richard Mccombie. 2016. "Coming of Age : Ten Years of next- Generation Sequencing Technologies." Nature Publishing Group 17 (6). Nature Publishing Group: 333–51. doi:10.1038/nrg.2016.49.

Graf, R., and H. K. Schachman. 1996. "Random Circular Permutation of Genes and Expressed Polypeptide Chains: Application of the Method to the Catalytic Chains of Aspartate Transcarbamoylase." Proceedings of the National Academy of Sciences 93 (21): 11591–96. doi:10.1073/pnas.93.21.11591.

Guntas, Gurkan, and Marc Ostermeier. 2004. "Creation of an Allosteric Enzyme by Domain Insertion." Journal of Molecular Biology 336 (1): 263–73. doi:10.1016/j.jmb.2003.12.016.

Guntas, Gurkan, Thomas J Mansell, Jin Ryoun Kim, and Marc Ostermeier. 2005. "Directed Evolution of Protein Switches and Their Application to the Creation of Ligand-Binding Proteins." Proceedings of the National Academy of Sciences of the United States of America 102 (32): 11224–29. doi:10.1073/pnas.0502673102.

Hanes, J, and A Plückthun. 1997. "In Vitro Selection and Evolution of Functional Proteins by Using Ribosome Display." Proceedings of the National Academy of Sciences of the United States of America 94 (10): 4937–42. doi:10.1073/pnas.94.10.4937.

Hassanpour, Neda, Ehsan Ullah, Mona Yousofshahi, Nikhil U. Nair, and Soha Hassoun. 2017. "Selection Finder (SelFi): A Computational Metabolic Engineering Tool to Enable Directed Evolution of Enzymes." Metabolic Engineering Communications 4 (October 2016). Elsevier B.V.: 37–47. doi:10.1016/j.meteno.2017.02.003.

Hecky, Jochen, and Kristian M. Müller. 2005. "Structural Perturbation and Compensation by Directed Evolution at Physiological Temperature Leads to Thermostabilization of Beta-Lactamase." Biochemistry 44 (38): 12640–54. doi.org/10.1021/bi0501885.

Heim, R, a B Cubitt, and R Y Tsien. 1995. "Improved Green Fluorescence." Nature 373 (6516): 663–64. doi:10.1038/373663b0.

Heinzelman, Pete, Christopher D Snow, Indira Wu, Catherine Nguyen, Alan Villalobos, Sridhar Govindarajan, Jeremy Minshull, and Frances H Arnold. 2009. “A Family of Thermostable Fungal Cellulases Created by Structure-Guided Recombination.” Proceedings of the National Academy of Sciences of the United States of America 106 (14): 5610–15. doi:10.1073/pnas.0901417106.

Higgins, Sean A., Sorel V. Y. Ouonkap, and David F. Savage. 2017. “Rapid and Programmable Protein Mutagenesis Using Plasmid Recombineering.” ACS Synthetic Biology. doi:10.1021/acssynbio.7b00112.

Higgins, Sean A., Sorel V. Y. Ouonkap, and David F. Savage. 2017. “Rapid and Programmable Protein Mutagenesis Using Plasmid Recombineering.” ACS Synthetic Biology 6 (10): 1825–33. doi.org/10.1021/acssynbio.7b00112.

Huang, Po-Ssu, Scott E. Boyken, and David Baker. 2016. “The Coming of Age of de Novo Protein Design.” Nature 537 (7620): 320–27. doi:10.1038/nature19946.

Hutchison, C. A., S. Phillips, M. H. Edgell, S. Gillam, P. Jahnke, and M. Smith. 1978. “Mutagenesis at a Specific Position in a DNA Sequence.” Journal of Biological Chemistry 253 (18): 6551–60.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. “A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity.” Science 337 (6096). science.sciencemag.org/content/337/6096/816.

Jones, Alicia M., Manan M. Mehta, Emily E. Thomas, Joshua T. Atkinson, Thomas H. Segall-Shapiro, Shirley Liu, and Jonathan J. Silberg. 2016. “The Structure of a Thermophilic Kinase Shapes Fitness upon Random Circular Permutation.” ACS Synthetic Biology 5 (5): 415–25. doi:10.1021/acssynbio.5b00305.

Jones, D. D. 2005. “Triplet Nucleotide Removal at Random Positions in a Target Gene: The Tolerance of TEM-1 -Lactamase to an Amino Acid Deletion.” Nucleic Acids Research 33 (9): e80–e80. doi.org/10.1093/nar/gni077.

Jones, D. Dafydd. 2005. “Triplet Nucleotide Removal at Random Positions in a Target Gene: The Tolerance of TEM-1 ??-Lactamase to an Amino Acid Deletion.” Nucleic Acids Research 33 (9): 1–8. doi:10.1093/nar/gni077.

Kim, David E, Frank Dimaio, Ray Yu-ruei Wang, Yifan Song, and David Baker. 2014. “One Contact for Every Twelve Residues Allows Robust and Accurate Topology-Level Protein Structure Modeling.” Proteins 82 (2): 208–18. doi:10.1002/prot.24374.One.

Kitzman, Jacob O, Lea M Starita, Russell S Lo, Stanley Fields, and Jay Shendure. 2015. “Massively Parallel Single-Amino-Acid Mutagenesis.” Nature Methods 12 (3): 203–6. doi:10.1038/nmeth.3223.

Koga, Nobuyasu, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione, and David Baker. 2012. “Principles for Designing Ideal Protein Structures.” Nature 491 (7423). Nature Publishing Group: 222–27. doi:10.1038/nature11600.

Komor, Alexis C., Yongjoo B. Kim, Michael S. Packer, John A. Zuris, and David R. Liu. 2016. “Programmable Editing of a Target Base in Genomic DNA without Double-Stranded DNA Cleavage.” Nature 533 (7603). Nature Publishing Group: 420–24. doi:10.1038/nature17946.

Kosuri, Sriram, Nikolai Eroshenko, Emily Leproust, Michael Super, Jeffrey Way, Jin Billy Li, and George M Church. 2011. “A Scalable Gene Synthesis Platform Using High-Fidelity DNA Microchips.” Nat Biotechnol 28 (12): 1295–99. doi:10.1038/nbt.1716.A.

Kriventseva, Evgenia V., Ina Koch, Rolf Apweiler, Martin Vingron, Peer Bork, Mikhail S. Gelfand, and Shamil Sunyaev. 2003. “Increase of Functional Diversity by Alternative Splicing.” Trends in Genetics: TIG 19 (3): 124–28. doi.org/10.1016/S0168-9525(03)00023-4.

Lander, E S, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” Nature 409: 860–921. doi:10.1038/35057062.

LeProust, Emily M., Bill J. Peck, Konstantin Spirin, Heather Brummel McCuen, Bridget Moore, Eugeni Namsaraev, and Marvin H. Caruthers. 2010. “Synthesis of High-Quality Libraries of Long (150mer) Oligonucleotides by a Novel Depurination Controlled Process.” Nucleic Acids Research 38 (8): 2522–40. doi:10.1093/nar/gkq163.

Li, Yougen, D Allan Drummond, Andrew M Sawayama, Christopher D Snow, Jesse D Bloom, and Frances H Arnold. 2007. “A Diverse Family of Thermostable Cytochrome P450s Created by Recombination of Stabilizing Fragments.” Nature Biotechnology 25 (9): 1051–56. doi:10.1038/nbt1333.

Lim, Sung In, Byung Eun Min, and Gyoo Yeol Jung. 2008. “Lagging Strand-Biased Initiation of Red Recombination by Linear Double-Stranded DNAs.” Journal of Molecular Biology 384 (5). Elsevier Ltd: 1098–1105. doi:10.1016/j.jmb.2008.10.047.

Loose, Matthew W. 2017. “The Potential Impact of Nanopore Sequencing on Human Genetics.” Human Molecular Genetics 26 (July): 202–7. doi:10.1093/hmg/ddx287.

Ma, Dacheng, Shuguang Peng, Weiren Huang, Zhiming Cai, and Zhen Xie. 2018. “Rational Design of Mini-Cas9 for Transcriptional Activation.” ACS Synthetic Biology 7 (4): 978–85. doi.org/10.1021/acssynbio.7b00404.

Ma, Enbo, Lucas B. Harrington, Mitchell R. O’Connell, Kaihong Zhou, and Jennifer A. Doudna. 2015. “Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes.” Molecular Cell 60 (3): 398–407. doi.org/10.1016/j.molcel.2015.10.030.

Majithia, Amit R, Ben Tsuda, Maura Agostini, Keerthana Gnanapradeepan, Robert Rice, Gina Peloso, Kashyap A Patel, et al. 2016. “Prospective Functional Classification of All Possible Missense Variants in PPARG.” Nature Genetics 48 (12): 1570–75. doi:10.1038/ng.3700.Prospective.

Majithia, Amit R, Ben Tsuda, Maura Agostini, Keerthana Gnanapradeepan, Robert Rice, Gina Peloso, Kashyap A Patel, et al. 2016. “Prospective Functional Classification of All Possible Missense Variants in PPARG.” Nature Genetics 48 (12): 1570–75. doi:10.1038/ng.3700.

Mandecki, Wlodek. 1998. “The Game of Chess and Searches in Protein Sequence Space.” Trends in Biotechnology 16 (5): 200–202. doi:10.1016/S0167-7799(98)01188-3.

Mardis, Elaine R. 2013. “Next-Generation Sequencing Platforms.” Annu. Rev. Anal. Chem 6: 287–303. doi:10.1146/annurev-anchem-062012-092628.

Matuszewski, Sebastian, Marcel E. Hildebrandt, Ana-Hermina Ghenu, Jeffrey D. Jensen, and Claudia Bank. 2016. “A Statistical Guide to the Design of Deep Mutational Scanning Experiments.” Genetics 204 (1): 77–87. doi:10.1534/genetics.XXX.XXXXXX.

Mehta, Manan M., Shirley Liu, and Jonathan J. Silberg. 2012. “A Transposase Strategy for Creating Libraries of Circularly Permuted Proteins.” Nucleic Acids Research 40 (9): 1–8. doi:10.1093/nar/gks060.

Melnikov, Alexandre, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S. Mikkelsen. 2014. “Comprehensive Mutational Scanning of a Kinase in Vivo Reveals Substrate-Dependent Fitness Landscapes.” Nucleic Acids Research 42 (14): 1–8. doi:10.1093/nar/gku511.

Melnikov, Alexandre, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S. Mikkelsen. 2014. “Comprehensive Mutational Scanning of a Kinase in Vivo Reveals Substrate-Dependent Fitness Landscapes.” Nucleic Acids Research 42 (14): e112. doi:10.1093/nar/gku511.

Miller, Corwin, Jiayi Kong, Truc T. Tran, Cesar A. Arias, Gerda Saxer, and Yousif Shamoo. 2013. “Adaptation of Enterococcus Faecalis to Daptomycin Reveals an Ordered Progression to Resistance.” Antimicrobial Agents and Chemotherapy 57 (11): 5373–83. doi:10.1128/AAC.01473-13.

Morelli, Aleardo, Yari Cabezas, Lauren J. Mills, and Burckhard Seelig. 2017. “Extensive Libraries of Gene Truncation Variants Generated by in Vitro Transposition.” Nucleic Acids Research 45 (10): gkx030. doi:10.1093/nar/gkx030.

Morelli, Aleardo, Yari Cabezas, Lauren J. Mills, and Burckhard Seelig. 2017. “Extensive Libraries of Gene Truncation Variants Generated by in Vitro Transposition.” Nucleic Acids Research 45 (10): e78. doi.org/10.1093/nar/gkx030.

Morgan, Stacy Anne, Dana C. Nadler, Rayka Yokoo, and David F. Savage. 2016. “Biofuel Metabolic Engineering with Biosensors.” Current Opinion in Chemical Biology 35: 150–58. doi:10.1016/j.cbpa.2016.09.020.

Mosberg, J. a., M. J. Lajoie, and G. M. Church. 2010. “Lambda Red Recombineering in Escherichia Coli Occurs through a Fully Single-Stranded Intermediate.” Genetics 186 (3): 791–99. doi:10.1534/genetics.110.120782.

Mukherjee, Arnab, Kevin B. Weyant, Utsav Agrawal, Joshua Walker, Isaac K O Cann, and Charles M. Schroeder. 2015. “Engineering and Characterization of New LOV-Based Fluorescent Proteins from Chlamydomonas Reinhardtii and Vaucheria Frigida.” ACS Synthetic Biology 4 (4): 371–77. doi:10.1021/sb500237x.

Nadler, Dana C, Stacy-Anne Morgan, Avi Flamholz, Kaitlyn E Kortright, and David F Savage. 2016. “Rapid Construction of Metabolite Biosensors Using Domain-Insertion Profiling.” Nature Communications 7. Nature Publishing Group: 12266. doi:10.1038/ncomms12266.

Nadler, Dana C, Stacy-Anne Morgan, Avi Flamholz, Kaitlyn E Kortright, and David F Savage. 2016. “Rapid Construction of Metabolite Biosensors Using Domain-Insertion Profiling.” Nature Communications 7. Nature Publishing Group: 12266. doi:10.1038/ncomms12266.

Nishimasu, Hiroshi, F. Ann Ran, Patrick D. Hsu, Silvana Konermann, Soraya I. Shehata, Naoshi Dohmae, Ryuichiro Ishitani, Feng Zhang, and Osamu Nureki. 2014. “Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA.” Cell 156 (5): 935–49. doi.org/10.1016/j.cell.2014.02.001.

Nyerges, Ákos, Bálint Csörgő, István Nagy, Balázs Bálint, Péter Bihari, Viktória Lázár, Gábor Apjok, et al. 2016. “A Highly Precise and Portable Genome Engineering Method Allows Comparison of Mutational Effects across Bacterial Species.” Proceedings of the National Academy of Sciences 113 (9): 2502–7. doi:10.1073/pnas.1520040113.

Oakes, Benjamin L, Dana C Nadler, Avi Flamholz, Christof Fellmann, Brett T Staahl, Jennifer A Doudna, and David F Savage. 2016. “Profiling of Engineering Hotspots Identifies an Allosteric CRISPR-Cas9 Switch.” Nature Biotechnology 34 (6). Nature Publishing Group: 646–51. doi:10.1038/nbt.3528.

Oakes, Benjamin L., Dana C. Nadler, and David F. Savage. 2014. “Protein Engineering of Cas9 for Enhanced Function.” Methods in Enzymology 546: 491–511. doi.org/10.1016/B978-0-12-801185-0.00024-6.

Oakes, Benjamin L., Dana C. Nadler, Avi Flamholz, Christof Fellmann, Brett T. Staahl, Jennifer A. Doudna, and David F. Savage. 2016. “Profiling of Engineering Hotspots Identifies an Allosteric CRISPR-Cas9 Switch.” Nature Biotechnology 34 (6): 646–51. doi.org/10.1038/nbt.3528.

Okada, Satoshi, Kazuhisa Ota, and Takashi Ito. 2009. “Circular Permutation of Ligand-Binding Module Improves Dynamic Range of Genetically Encoded FRET-Based Nanosensor.” Protein Science 18 (12): 2518–27. doi:10.1002/pro.266.

Olson, C. A., N. C. Wu, and R. Sun. 2014. “A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain.” Curr. Biol. 24: 2643–51. doi:10.1038/jid.2014.371.

Olson, C. Anders, Nicholas C. Wu Wu, and Ren Sun. 2014. “A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain.” Curr. Biol. 24 (22): 2643–51. doi:10.1038/jid.2014.371.

Ostermeier, M., J. H. Shim, and S. J. Benkovic. 1999. “A Combinatorial Approach to Hybrid Enzymes Independent of DNA Homology.” Nature Biotechnology 17 (12): 1205–9. doi.org/10.1038/70754.

Osuna, Joel, Alejandra Pérez-Blancas, and Xavier Soberón. 2002. “Improving a Circularly Permuted TEM-1 Beta-Lactamase by Directed Evolution.” Protein Engineering 15 (6): 463–70. doi:10.1093/protein/15.6.463.

Packer, Michael S, and David R Liu. 2015. “Methods for the Directed Evolution of Proteins.” Nature Reviews. Genetics 16 (7). Nature Publishing Group: 379–94. doi:10.1038/nrg3927.

Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” Nature Genetics 40 (12): 1413–15. doi.org/10.1038/ng.259.

Persikov, Anton V., Elizabeth F. Rowland, Benjamin L. Oakes, Mona Singh, and Marcus B. Noyes. 2014. “Deep Sequencing of Large Library Selections Allows Computational Discovery of Diverse Sets of Zinc Fingers That Bind Common Targets.” Nucleic Acids Research 42 (3): 1497–1508. doi:10.1093/nar/gkt1034.

Pierre, Brennal, Vandan Shah, Jenny Xiao, and Jin Ryoun Kim. 2015. “Construction of a Random Circular Permutation Library Using an Engineered Transposon.” Analytical Biochemistry 474. Elsevier Inc.: 16–24. doi:10.1016/j.ab.2014.12.011.

Pisarchik, Alexander, Ralf Petri, and Claudia Schmidt-Dannert. 2007. “Probing the Structural Plasticity of an Archaeal Primordial Cobaltochelatase CbiX(S).” Protein Engineering, Design & Selection: PEDS 20 (6): 257–65. doi.org/10.1093/protein/gzm018.

Qi, Lei S., Matthew H. Larson, Luke A. Gilbert, Jennifer A. Doudna, Jonathan S. Weissman, Adam P. Arkin, and Wendell A. Lim. 2013. “Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression.” Cell 152 (5): 1173–83. doi.org/10.1016/j.cell.2013.02.022.

Qian, Zhen, and Stefan Lutz. 2005. “Improving the Catalytic Activity of Candida Antarctica Lipase B by Circular Permutation.” Journal of the American Chemical Society 127 (39): 13466–67. doi:10.1021/ja053932h.

Reitinger, Stephan, Ying Yu, Jacqueline Wicki, Martin Ludwiczek, Igor D’Angelo, Simon Baturin, Mark Okon, et al. 2010. “Circular Permutation of Bacillus Circulam Xylanase: A Kinetic and Structural Study.” Biochemistry 49 (11): 2464–74. doi:10.1021/bi100036f.

Reynolds, Kimberly A., Richard N. McLaughlin, and Rama Ranganathan. 2011. “Hot Spots for Allosteric Regulation on Protein Surfaces.” Cell 147 (7): 1564–75. doi.org/10.1016/j.cell.2011.10.049.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” Genomics, Proteomics and Bioinformatics 13 (5). Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China: 278–89. doi:10.1016/j.gpb.2015.08.002.

Rocklin, Gabriel J., Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, et al. 2017. “Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing.” Science 357 (6347): 168–75. doi:10.1126/science.aan0693.

Rogers, Jameson K., and George M. Church. 2016. “Genetically Encoded Sensors Enable Real-Time Observation of Metabolite Production.” Proceedings of the National Academy of Sciences 113 (9): 2388–93. doi:10.1073/pnas.1600375113.

Romero, Philip a, and Frances H Arnold. 2009. “Exploring Protein Fitness Landscapes by Directed Evolution.” Nature Reviews. Molecular Cell Biology 10. Nature Publishing Group: 866–76. doi:10.1038/nrm2805.

Romero, Philip A., and Frances H. Arnold. 2012. “Random Field Model Reveals Structure of the Protein Recombinational Landscape.” PLoS Computational Biology 8 (10). doi:10.1371/journal.pcbi.1002713.

Rye, Peter T., and William A. LaMarr. 2015. “Measurement of Glycolysis Reactants by High-Throughput Solid Phase Extraction with Tandem Mass Spectrometry: Characterization of Pyrophosphate-Dependent Phosphofructokinase as a Case Study.” Analytical Biochemistry 482. Elsevier Inc.: 40–47. doi:10.1016/j.ab.2015.03.029.

Schmid, A, J S Dordick, B Hauer, A Kiener, M Wubbolts, and B Witholt. 2001. “Industrial Biocatalysis Today and Tomorrow” 409: 258–268.

Shah, Vandan, and Jin Ryoun Kim. 2016. “Transposon for Protein Engineering.” Mobile Genetic Elements 6 (6). Taylor & Francis: e1239601. doi:10.1080/2159256X.2016.1239601.

Shendure, Jay, and Stanley Fields. 2016. “Massively Parallel Genetics.” Genetics 203 (2): 617–19. doi:10.1534/genetics.115.180562.

Simm, Alan M., Amy J. Baldwin, Kathy Busse, and D. Dafydd Jones. 2007. “Investigating Protein Structural Plasticity by Surveying the Consequence of an Amino Acid Deletion from TEM-1 Beta-Lactamase.” FEBS Letters 581 (21): 3904–8. doi.org/10.1016/j.febslet.2007.07.018.

Starita, Lea M., David L. Young, Muhtadi Islam, Jacob O. Kitzman, Justin Gullingsrud, Ronald J. Hause, Douglas M. Fowler, Jeffrey D. Parvin, Jay Shendure, and Stanley Fields. 2015. “Massively Parallel Functional Analysis of BRCA1 RING Domain Variants.” Genetics 200 (2): 413–22. doi:10.1534/genetics.115.175802.

Steinberg, B., and M. Ostermeier. 2016. “Environmental Changes Bridge Evolutionary Valleys.” Science Advances 2 (1): e1500921–e1500921. doi:10.1126/sciadv.1500921.

Sternberg, Samuel H., Benjamin LaFrance, Matias Kaplan, and Jennifer A. Doudna. 2015. “Conformational Control of DNA Target Cleavage by CRISPR–Cas9.” Nature 527 (7576): 110–13. doi.org/10.1038/nature15544.

Sternberg, Samuel H., Sy Redding, Martin Jinek, Eric C. Greene, and Jennifer A. Doudna. 2014. “DNA Interrogation by the CRISPR RNA-Guided Endonuclease Cas9.” Nature 507 (7490): 62–67. doi.org/10.1038/nature13011.

Sung, Keewon, Jinho Park, Younggyu Kim, Nam Ki Lee, and Seong Keun Kim. 2018. “Target Specificity of Cas9 Nuclease via DNA Rearrangement Regulated by the REC2 Domain.” Journal of the American Chemical Society 140 (25): 7778–81. doi.org/10.1021/jacs.8b03102.

Thomason, L. C., N. Constantino, D. V. Shaw, and D. L. Court. 2007. “Multicopy Plasmid Modification with Phage λ Red Recombineering.” Plasmid 58 (5): 148–58. doi:10.1038/jid.2014.371.

Thomason, Lynn C., Nina Costantino, Dana V. Shaw, and Donald L. Court. 2007. “Multicopy Plasmid Modification with Phage Lambda Red Recombineering.” Plasmid 58 (2): 148–58. doi.org/10.1016/j.plasmid.2007.03.001.

Tokuriki, Nobuhiko, and Dan S. Tawfik. 2009. “Stability Effects of Mutations and Protein Evolvability.” Current Opinion in Structural Biology 19: 596–604. doi:10.1016/j.sbi.2009.08.003.

Tokuriki, Nobuhiko, and Dan S. Tawfik. 2009. “Stability Effects of Mutations and Protein Evolvability.” Current Opinion in Structural Biology 19: 596–604. doi:10.1016/j.sbi.2009.08.003.

Tullman, Jennifer, Nathan Nicholes, Matt R. Dumont, Lucas F. Ribeiro, and Marc Ostermeier. 2016. “Enzymatic Protein Switches Built from Paralogous Input Domains.” Biotechnology and Bioengineering 113 (4): 852–58. doi:10.1002/bit.25852.

Turrientes, M. C., F. Baquero, B. R. Levin, J. L. Martinez, Aida Ripoll, Jose Maria Gonzalez-Alba, Raquel Tobes, et al. 2013. “Normal Mutation Rate Variants Arise in a Mutator (Mut S) Escherichia Coli Population.” PLoS ONE 8 (9): e72963. doi:10.1371/journal.pone.0072963.

Vanderporten, Erica, Lauren Frick, Rebecca Turincio, Peter Thana, William Lamarr, and Yichin Liu. 2013. “Label-Free High-Throughput Assays to Screen and Characterize Novel Lactate Dehydrogenase Inhibitors.” Analytical Biochemistry 441. Elsevier Inc.: 115–22. doi:10.1016/j.ab.2013.07.003.

Voigt, Christopher A., Carlos Martinez, Zhen-Gang Wang, Stephen L. Mayo, and Frances H. Arnold. 2002. “Protein Building Blocks Preserved by Recombination.” Nature Structural Biology 9 (7): 553–58. doi:10.1038/nsb805.

Walkiewicz, Katarzyna, Andres S Benitez, Christine Sun, Colin Bacorn, Gerda Saxer, Andres S Benitez Cardenas, Christine Sun, Colin Bacorn, Gerda Saxer, and Yousif Shamoo. 2012. “Small Changes in Enzyme Function Can Lead to Surprisingly Large Fi Tness Effects during Adaptive Evolution of Antibiotic Resistance.” Proceedings of the National Academy of Sciences 109 (52): 21408–13. doi:10.1073/pnas.1209335110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1209335110.

Wang, Harris H, Farren J Isaacs, Peter a Carr, Zachary Z Sun, George Xu, Craig R Forest, and George M Church. 2009. “Programming Cells by Multiplex Genome Engineering and Accelerated Evolution.” Nature 460 (7257). Nature Publishing Group: 894–98. doi:10.1038/nature08187.

Wang, HH, and GM Church. 2011. “Multiplexed Genome Engineering and Genotyping Methods Applications for Synthetic Biology and Metabolic Engineering.” Methods Enzymol 498 (January). Elsevier Inc.: 409–26. doi:10.1016/B978-0-12-385120-8.00018-8.

Warner, Joseph R, Philippa J Reeder, Anis Karimpour-Fard, Lauren B a Woodruff, and Ryan T Gill. 2010. “Rapid Profiling of a Microbial Genome Using Mixtures of Barcoded Oligonucleotides.” Nature Biotechnology 28 (8). Nature Publishing Group: 856–62. doi:10.1038/nbt.1653.

Weile, Jochen, Song Sun, Atina G. Cote, Jennifer Knapp, Marta Verby, Joseph C. Mellor, Yingzhou Wu, et al. 2017. “Expanding the Atlas of Functional Missense Variation for Human Genes.” bioRxiv 1. http://www.biorxiv.org/content/early/2017/07/27/166595?%3Fcollection=.

Weiner, January, 3rd, Francois Beaussart, and Erich Bornberg-Bauer. 2006. “Domain Deletions and Substitutions in the Modular Protein Evolution.” The FEBS Journal 273 (9): 2037–47. doi.org/10.1111/j.1742-4658.2006.05220.x.

Wrenbeck, Emily E, Justin R Klesmith, James A Stapleton, Adebola Adeniran, Keith E J Tyo, and Timothy A Whitehead. 2016. “Plasmid-Based One-Pot Saturation Mutagenesis.” Nature Methods 13 (11): 928–30. doi:10.1038/nmeth.4029.

Wrenbeck, Emily E, Justin R Klesmith, James A Stapleton, Adebola Adeniran, Keith E J Tyo, and Timothy A Whitehead. 2016. “Plasmid-Based One-Pot Saturation Mutagenesis.” Nature Methods 13 (11): 928–30. doi:10.1038/nmeth.4029.

Wrenbeck, Emily E, Laura R Azouz, and Timothy A Whitehead. 2017. “Single-Mutation Fitness Landscapes for an Enzyme on Multiple Substrates Reveal Specificity Is Globally Encoded.” Nature Communications 8. Nature Publishing Group: 15695. doi:10.1038/ncomms15695.

Wrenbeck, Emily E., Matthew S. Faber, and Timothy A. Whitehead. 2017. “Deep Sequencing Methods for Protein Engineering and Design.” Current Opinion in Structural Biology 45. Elsevier Ltd: 36–44. doi:10.1016/j.sbi.2016.11.001.

Wu, Chien-Hsun, I-Ju Liu, Ruei-Min Lu, and Han-Chung Wu. 2016. “Advancement and Applications of Peptide Phage Display Technology in Biomedical Science.” Journal of Biomedical Science 23 (1). Journal of Biomedical Science: 8. doi:10.1186/s12929-016-0223-x.

Yang, Guangyu, Jamie R Rich, Michel Gilbert, Warren W Wakarchuk, Yan Feng, Stephen G Withers, Biological Sciences, Sussex Drive, and S Structures. 2010. “Fluorescence Activated Cell Sorting as a General Ultra High-Throughput Screening Method for Directed Evolution of Glycosyltransferases,” no. 9: 1–31.

Zheng, Xiang, Xin-Hui Xing, and Chong Zhang. 2017. “Targeted Mutagenesis: A Sniper-like Diversity Generator in Microbial Engineering.” Synthetic and Systems Biotechnology 2. Elsevier Ltd: 75–86. doi:10.1016/j.synbio.2017.07.001.