# HIGH-RESOLUTION SEQUENCE-FUNCTION MAPPING OF PROTEIN-PROTEIN INTERACTIONS FOR CONFORMATIONAL EPITOPE MAPPING

By

Caitlin Adele Stein

A DISSERTATION

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

Chemical Engineering – Doctor of Philosophy

2016

## ABSTRACT

HIGH-RESOLUTION SEQUENCE-FUNCTION MAPPING OF PROTEIN-PROTEIN INTERACTIONS FOR CONFORMATIONAL EPITOPE MAPPING

By

Caitlin Adele Stein

Protein-protein interactions are essential for biological signaling, including the adaptive immune system, membrane transport and cell metabolism. A protein's sequence defines its function; however, the relationship between a protein's sequence and its function is not a well-understood problem. Recent advances in DNA sequencing technologies have allowed the development of independent high-throughput methods to couple a protein's sequence to its function. These methods analyze the effect of individual mutations on a protein's fitness. However, the methods lack standardization leading to many different experimental setups and data presentations. In this dissertation we present a validated and standardized method to determine the sequence-function relationships of protein-protein interactions. A series of equations was developed to model and optimize experimental conditions and to expand the accessibility of the technique. The method was further used to characterize the effect on binding affinity of all single-point mutations for two protein-protein interactions involved in biomass degradation. Finally, we have utilized this method to introduce a novel platform technology for rapid determination of fine conformational epitopes. This technology involves deep sequencing of yeast displayed antigen libraries and analytical equations to identify epitope positions. We show the methods effectiveness by determining critical (and previously unknown) neutralizing epitopes for pertussis toxin and a breast cancer target. We further show the implications of this method for structural-based vaccine design.

## ACKNOWLEDGEMENTS

I would never have been able to complete graduate school and my dissertation with out the guidance of my advisor, help from my friends and support from my family.

I would like to express my deepest appreciation to my advisor, Professor Tim Whitehead, who took allowed me to join his lab five years ago. Without your excellent guidance, encouragement, and patience I would never have been able to reach this milestone. Thank you for providing an excellent research environment where I have developed the skills to become an independent thinker and researcher.

To the past and present members of the lab that have worked side-by-side with me, thank you. Dr. Jim Stapleton, thank you for your mentorship and for always keeping me positive. Your willingness to help troubleshoot experiments, brainstorm ideas, interpret results and proof read my papers have been invaluable. Justin Klesmith, Carolyn Haarmeyer, Emily Wrenbeck and Matthew Faber I am proud to call you my colleagues and friends. You all have been incredibly supportive and helpful, I could not have asked for better lab mates.

I am grateful for the collaborations I have been able to have with Professor Jennifer Maynard's Lab at the University of Texas, Austin and with Professor Christina Chan's Lab here at Michigan State University. These collaborations provided exciting opportunities to develop and apply our methods to real world problems.

Finally, to my graduate school friends, especially Cory Sarks and Phillip Angart. Thank you for listening to my frustrations and for all your help getting through our classes.

# PREFACE

Antibodies are a leading class of therapeutics with an excess of $80 billion in sales for 2015. With about 4 new antibody therapies being approved every year the antibody market is expected to reach $125 billion in world-wide sales by 2020 [1]. In addition to therapeutic uses antibodies are also used in diagnostic and research applications.

Antibodies attach directly to antigens at specific epitopes to prevent pathogens from damaging or entering healthy cells. Conformational epitopes consist of discontinuous stretches of amino acids that upon folding become adjacent. Broadly neutralizing antibodies target conformational epitopes for a variety of well-known pathogens such as influenza and HIV. Knowledge of the conformational epitope aids in the understanding of the structural basis of the protein-protein interaction and may lead to improved vaccine designs and neutralizing antibody therapies for less studied pathogens. Recent technology advances have resulted in the rapid isolation of neutralizing antibodies for Ebola and Dengue viruses but structural vaccine design advances awaits the development of a high-throughput method to map the fine conformation epitopes. In this thesis I present three major contributions to the field (1.) a high-throughput standardized protocol for probing the sequence-function relationships of protein-protein interactions; (2.) the ability to quantify the energetic contributions of protein mutations from deep sequencing data sets and (3.) utilizing sequence-function relationships in the context of a standardized epitope mapping protocol.

**Chapter 1** presents an introduction to protein-protein interactions and current methods of probing sequence-function relationships. We sought to develop a standardized protocol using deep mutational scanning for resolving the sequence determinants of function for full-length proteins,

given its demonstrated utility and growing popularity as a tool to understand and optimize protein function. In **Chapter 2**, we present experimental methods for mutant library creation, functional selections, and sequencing library preparation. We also develop equations that allow quantitative comparisons across different populations in growth-based selections and fluorescence activated cell sorting (FACS). These advances enable optimal selection criteria to be determined for these versatile selection techniques. This work has been published in *PLoS One*.

In **Chapter 3**, we use our deep mutational scanning pipeline to evaluate the effect of binding affinity to dockerin for nearly all-possible single point mutants on the type I cohesin domain for both *Clostridium thermocellum* and *Clostridium cellulolyticum* species, giving a comprehensive picture of one side of the affinity landscape for these protein-protein complexes. These advances allow for generation of deep mutational scanning benchmark sets where the change in binding affinity can be quantitatively evaluated for each point mutant in a given protein sequence. This work has been submitted for publication in *PROTEINS.*

**Chapter 4** shows an application of the deep mutational scanning method for rapid conformational epitope mapping. We develop analytical equations to identify epitope positions, and show the method effectiveness by mapping the fine epitope for different antibodies targeting TNF, Pertussis Toxin, and the cancer target TROP2. This work has been published in *The Journal of Biological Chemistry*.

During the development of antibody therapeutics many candidate antibodies are screened. These antibodies will bind to many different epitopes, but not all will be neutralizing, or elicit the correct immune response. Knowledge of both neutralizing and non-neutralizing epitopes would help with the design of antigens for vaccines. In **Chapter 5**, we show a follow-up study done to show the applications of the conformational epitope mapping method by mapping the

conformational epitopes for a panel of antibodies against the Pertussis Toxin as well as discuss future improvements and applications of the method.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1

## 1. Introduction

Protein-protein interactions are essential for biological signaling, including the adaptive immune system, membrane transport, and cell metabolism among many other functions. The ability to design protein-protein interactions could lead to more successful vaccines and antibody therapies. However, programming specific functions into proteins is difficult as proteins are only marginally stable and protein structure-function relationships are not well understood. A better understanding of sequence-function relationships of protein-protein interactions would facilitate antibody-epitope mapping, rapid antibody-antigen maturation, and fine-tuning of computationally designed proteins [2-5].

### 1.1 Background

#### 1.1.1 Probing sequence-function relationships of protein-protein interactions

Protein-protein interactions are characterized by the combination of many weak interactions like hydrogen bonds, Van der Waals forces, electrostatic interactions and hydrophobic contacts from individual amino acids contribute to a binding affinity. Traditional methods for probing sequence-function relationships for protein-protein interactions are laborious and inefficient. One of the first methods of exploring these relationships is alanine scanning [2]. In this approach, individual residues are mutated to alanine and the resulting activity evaluated. Simply, if a residue is mutated to alanine, removing its functional group without introducing conformational flexibility, and there is no change in binding affinity, the residue is not involved in the protein-protein interaction. However, if upon mutation to alanine, there is a decrease in binding affinity, the residue is important, or a 'hot spot'[6], in the protein-protein interaction. Alanine scanning has been used in many applications including epitope mapping of human-growth hormone and its receptor [2, 7]. However, this method requires the construction, expression and

characterization of each mutant protein separately, thus limiting the method to testing tens or hundreds of protein mutants.

The ability to parallelize the expression and evaluation of each protein variant was introduced using phage display [8]. With phage display, researchers could construct a library of phage, with each phage displaying a different mutant protein. The library could then be subject to a screen to isolate functional mutants. Using phage display, Pal et al. provided a comprehensive look at all of the structural and functional effects of all possible mutations across a large protein-protein interface, demonstrating a comprehensive and quantitative mapping of a protein's energy landscape[9]. Protein libraries were created using saturation mutagenesis at multiple interface positions and screened to enrich the population in mutants with enhanced function at the expense of those with impaired function. Individual clones from the selection were sequenced to determine the occurrence of each amino acid residue. However, their method requires extensive sequencing and many selection experiments, thus limiting the method to selected residues at the interface.

With the improved screening throughput using phage display, many other methods using combinatorial libraries such as saturation mutagenesis or limited representative amino acid libraries such as Look-Through Mutagenesis (LTM) have been developed to probe the sequence-function space of proteins [10, 11]. These methods have been instrumental and powerful in antibody and metabolic engineering. However, like alanine scanning they are laborious and limited in scope and resolution, providing limited information about sequence-function relationships.

### 1.1.2 Deep mutational scanning

Next generation sequencing technologies allows millions of sequencing reads to be acquired inexpensively and in parallel. Linking this new technology with functional protein screens, entire libraries can be sequenced before and after a selection giving measurements of

function. The combination of functional protein library screens linking genotype to phenotype with the sequencing capabilities of next generation sequencing has been deemed 'deep mutational scanning' [12, 13]. Deep mutational scanning has provided a high-throughput method to comprehensively map the energetic landscapes of protein-protein interactions [12, 13] (**Figure 1**). In deep mutational scanning, a single site-saturation mutagenesis library representing all possible single residue amino acid substitutions is made. A high-throughput assay that couples the protein's genotype to phenotype is used to screen the library. For binding proteins, cell-based assays such as phage- or yeast-display systems are typically used. The protein library is put through this functional selection, enriching the library in beneficial mutations. In a key step, deep sequencing is used to quantify the frequency of each mutant in the library before and after selection, resulting in an enrichment ratio. The ability to sequence millions of sequences in a library using next generation sequencing, allows quantification of thousands of protein variants in a single experiment. Deep mutational scanning has been independently applied to many protein engineering applications including affinity maturation, specificity switches and protein stability among others[4, 14-18].

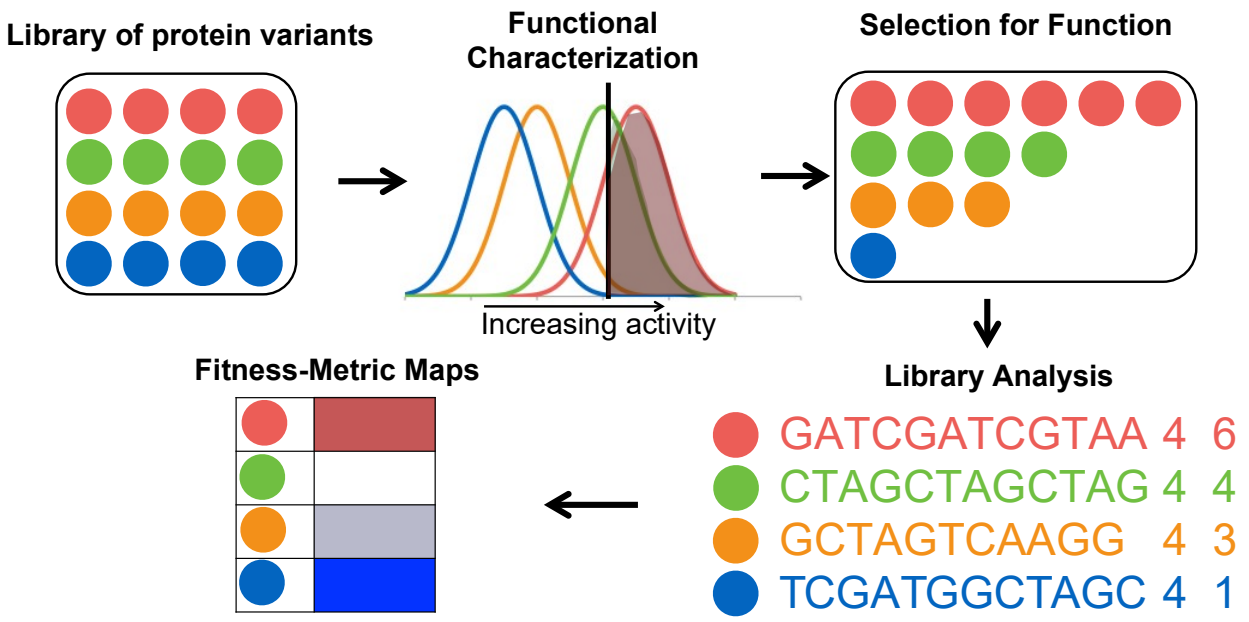



**Figure 1- Schematic of deep mutational scanning.**
A single site saturation mutagenesis library is created. The library is discriminated using a functional selection. After selection mutants with enhanced function are present in the population more frequently then mutants with decreased function. The library is sequenced before and after selection and the number of times a single mutant appears is counted. The enrichment ratio is determined based on the frequency of each mutant in the population. Enrichment ratios are used as a proxy to measure the fitness of each mutation. (adapted from Ayara et al.[19]).

### 1.1.3 Quantifying Deep Mutational Scanning

The ability to screen thousands of protein variants using deep mutational scanning has enabled researchers to determine the fitness of mutants by giving a score to each mutant based on its frequency in the population before and after a selection. This score, or measure of activity is known as an enrichment ratio. The enrichment ratio is calculated by taking the log2 ratio of the frequency of a mutant before and after selection. An enrichment ratio of zero represents a neutral mutation, a positive enrichment ratio a beneficial mutation and a negative enrichment ratio a deleterious mutation. The enrichment ratio is usefulness is limited as it can only be used as a proxy for affinity relative to the wild-type protein. McLaughlin et al. have reported data suggesting that there is a correlation of enrichment ratios with binding affinities, but the exact relationship is

unknown [20]. These high-throughput deep mutational scanning characterization methods would be more useful if they could deliver accurate estimates of affinity.

There have been many attempts to quantify the effects of mutations mostly resulting in rank ordering of binding characteristics. Kinney et al. and Sharon et al. designed experiments that allowed the analysis of frequency distributions rather than enrichment ratios [21, 22]. Reich et al. furthered their methods introducing their SORTCERY which combines cell sorting, and deep sequencing of protein libraries that semi-quantitatively determines the binding characteristics for large peptide libraries by producing a rank ordered list of 1000 peptide liginds[23]. These methods however, require multi-bin sorts and high sequencing depth while only providing semi-quantitative results. There remains to be a method that reconstructs quantitative energetics from deep sequencing data sets.

**1.2 Conformational epitope mapping**

The epitope is the part of an antigen recognized by the immune system by an antibody. Epitopes can be linear or conformational. A linear epitope consists of a continuous stretch of the amino acid sequence while conformational epitopes consist of discontinuous amino acid sequences that upon folding are adjacent. Identification of the fine conformational epitope targeted by an antibody can give a basis for intellectual property protection, lead to improved therapies or give a better understanding mechanism of protection [4, 24-31].

The gold standard in epitope mapping is co-crystallography, which provides an unambiguous high-resolution epitope. High-quality crystals and structures however can require large amounts of purified protein as well as considerable effort and training often requiring the preparation of many antigen variants in order to find one compatible with crystallization [25]. Methods linking hydrogen-deuterium exchange and mass spectrometry also identify

conformational epitopes to about 5 amino acid resolution but require large amounts of purified protein, specialized training and rigorous controls as well to obtain quality results [32, 33]. Recently, cryoelectron microscopy has emerged as an alternative to obtain conformational epitope maps requiring small amounts of sample and no time-limiting steps (cite). However, without specialized training epitope maps obtained are low-resolution, not providing information on specific residue interactions. Other conformational epitope mapping methods that have been developed include epitope binning which provide low-resolution information about relative locations of epitopes based off of competitive binding experiments in a high-throughput manner.

In addition to conformational epitope mapping methods, many linear epitope mapping methods have been developed as they easily lend themselves to high-throughput platforms. In these methods, libraries of short peptide sequences are immobilized using phage- or yeast-display technologies or on chips. Antigens are allowed to bind to the libraries and the peptide sequences that bind to the antigen are identified. These high-throughput methods provide high-resolution linear epitopes however, since a majority of epitopes are conformational the applicability of linear epitope methods is limited [34].

A high-throughput method to identify high-resolution conformational epitopes has yet to be identified. Weiss and colleagues introduced the use of alanine scanning to map a proteins functional epitope by determining the alanine/wild-type ratio of DNA sequences at each mutated position. The ratio was used to calculate the effect on the change in free energy each alanine mutation had. Mutations with a change in free energy grater than 1.0 kcal/mol were considered to be part of the epitope[35]. In addition to alanine scanning epitopes have also been mapped by identifying escape mutants for dengue virus, hemagglutinin and hepatitis C[36-38]. While alanine scanning and other mutation based methods provide detailed information about conformational

epitopes, their throughput is limited as each individual mutation is separately characterized. With the introduction next generation sequencing technologies again, larger comprehensive and representative libraries have been constructed and incorporated with display-based methods for epitope mapping[39-41]. However these methods use many sorts[39] or identify only partial epitopes[40, 41].

# CHAPTER 2

## 2. High-Resolution Sequence-Function Mapping of Full-Length Proteins

### 2.1 Abstract

Comprehensive sequence-function mapping involves detailing the fitness contribution of every possible single mutation to a gene by comparing the abundance of each library variant before and after selection for the phenotype of interest. Deep sequencing of library DNA allows frequency reconstruction for tens of thousands of variants in a single experiment, yet short read lengths of current sequencers makes it challenging to probe genes encoding full-length proteins. Here we extend the scope of sequence-function maps to entire protein sequences with a modular, universal sequence tiling method. We demonstrate the approach with both growth-based selections and FACS screening, offer parameters and best practices that simplify design of experiments, and present analytical solutions to normalize data across independent selections. Using this protocol, sequence-function maps covering full sequences can be obtained in four to six weeks. Best practices introduced in this manuscript are fully compatible with, and complementary to, other recently published sequence-function mapping protocols.

### 2.2 Introduction

The amino acid sequence of a protein defines its function, yet our understanding of the contribution of each amino acid to overall activity remains incomplete. As a result, current computational and experimental methods of designing functional proteins have success rates significantly less than 10% [1]. Random directed evolution approaches provide activity improvements, but require high throughputs because about 98% of amino acid substitutions are either deleterious or neutral with respect to the desired function or specific fold [2]. Traditional methods for probing sequence-function relationships, such as alanine scanning and site-saturation

mutagenesis, are laborious and inefficient [3-6]. A systematic method to survey the sequence-function space of large proteins would facilitate enzymatic efficiency improvements, antibody-epitope mapping, rapid antibody-antigen maturation, and fine-tuning of computationally designed proteins [7-10].

About a decade ago, Pal et al. introduced a quantitative scanning method to map the energetic landscapes of protein-protein interactions [11]. Libraries were created using saturation mutagenesis at multiple positions and screened with phage display to enrich the population in mutants with enhanced function at the expense of those with impaired function. The complete library was sequenced before and after selection, and comparisons of these frequencies gave a measure of activity for each variant. More recently, Fowler et al. used a similar framework to develop deep mutational scanning [12, 13]. In a key step, deep sequencing is used to quantify the frequency of each mutant in the library before and after selection, and the resulting enrichment ratio provides a fitness metric. The ability to sequence millions of sequences in a library allows quantification of thousands of protein variants in a single experiment. Independently, Hietpas et al. developed a similar technique termed EMPIRIC, which they applied to measure fitness effects of point mutations of regions of genes in yeast [14, 15]. Since the introduction of deep mutational scanning, similar methods have been applied to characterize protein-ligand interactions and chaperone protein function [11, 16]. In a recent report demonstrating the power of the approach, Firnberg et al. produced a comprehensive map of nearly all possible single mutations to a full-length protein, TEM-1β-Lactamase [17]. By combining comprehensive single-site mutagenesis with selection through antibiotic resistance they were able to assess the fitness of 5,760 different mutant protein sequences in a single experiment.

Deep mutational scanning methods were extended to protein engineering applications by Whitehead et al., who applied the deep mutational scanning technique to enhance the affinity and specificity of two designed influenza inhibitors [9]. Deep mutational scanning has since been applied in many different areas of protein engineering including specificity switches and protein stability [18, 19].

Given the demonstrated utility and growing popularity of deep mutational scanning as a tool to understand and optimize protein function, we sought to develop a standardized protocol for resolving the sequence determinants of function for full-length proteins. In this contribution, we develop and validate experimental methods for mutant library creation, functional selections, and sequencing library preparation. We derive equations that allow direct, quantitative comparisons across different populations in growth-based selections and fluorescence activated cell sorting (FACS), enabling optimal selection criteria to be determined for these versatile selection techniques. We introduce a gene tiling technique which splits a long gene sequence into several independent libraries, each of which contain a mutated region short enough to be covered with a paired-end read [20]. This approach, combined with the equations developed herein, allow for the unambiguous reconstruction of the sequence-function determinants of full-length proteins. Key considerations for each step in the process are discussed.

## 2.3 Materials and Methods

### 2.3.1 Strains

E. coli strains used in this study: XL1-Blue (Agilent, Santa Clara, CA) recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F' proAB lacI1$^{q}$ZΔM15 Tn10 (Tet$^{r}$)]; Tuner (Novagen, Billerica, MA) F- ompT hsdS$_B$ (r$_B$- m$_B$-) gal dcm lacY1; K12 CJ236 (NEB) FΔ(HindIII)::cat (Tra+ Pil+ CamR)/ ung-1 relA1 thi-1 spoT1 mcrA

### 2.3.2 Plasmids

The plasmid pJK_proJK1_LGK was created by inserting a codon-optimized gene encoding levoglucosan kinase (LGK) (Genscript, Piscataway, NJ) with LEHHHHHH as 95 the C-terminal tag into a pJK-series plasmid [20] using flanking NdeI/XhoI restriction sites. The plasmid pJK_proJK1_kanR_LGK was created by switching the ampR with a kanR resistance cassette using Gibson cloning [21]. Full sequences of both plasmids are given in Kowalsky et al.[22] pJK_eGFP-series plasmids are from a previous study and are listed in Bienick et al.[23] .

### 2.3.3 Pfunkel Mutagenesis

Single-site saturation mutagenesis primers containing an NNN degenerate codon were designed in one of two ways: (1.) the online QuikChange Primer Design module (Agilent, Santa Clara, CA); or (2.) primer-design software as detailed in Firnberg et al. [24, 25]. Mutagenic libraries were generated from a ssDNA template using the Pfunkel method for comprehensive codon mutagenesis [25]. A separate Pfunkel reaction was performed for each tile region. Protocols were performed as published except the reaction cycling conditions were 95°C for 2 min, followed by 15 cycles of 95°C for 30 sec, 55°C for 45 sec, and 68°C for 15 min. Following the nuclease step the reaction was concentrated using the Zymo Clean and Concentrate kit (Zymo Research, Irvine, CA) and eluted in 6 µL of nanopure water. The entire volume was mixed with 40 µL of electrocompetent XL1-Blue cells (Agilent, Santa Clara, CA). Cells were transformed by electroporation at 1200 V in a 1 mm electroporation cuvette (Eppendorf, Hauppage, NY) with an Eppendorf Eporator. Transformed cells were grown overnight at 37°C on LB agar supplemented with appropriate antibiotic on Nalgene BioAssay plates (245mm x 245mm x 25mm, Sigma Aldrich, St. Louis, MO). Library plasmid DNA was recovered by scraping the BioAssay plate with

5 mL LB, centrifuging the solution to recover the cell pellet, and performing a plasmid midiprep (Qiagen, Valencia, CA) on the cell pellet.

Cells were also plated in serial dilutions from $10^{-1}$ to $10^{-6}$ to assess transformation efficiency. Transformation efficiencies ranged from $3x10^5$-$1x10^6$ cfu/transformation.

### 2.3.4 Secondary Transformations

*E. coli* Tuner (Novagen, Billercia, MA) was prepared to be electrocompetent by standard means [26]. Plasmids pJK_proJK1_LGK and pJK_proJK1_kanR_LGK were mixed at a mass ratio of 100:1 respectively. 5-40 ng of the mixed plasmid DNA was transformed into 40 µL of culture by electroporation at 1200 V in 0.1 cm cuvettes (Eppendorf, Hauppage, NY). The reaction was split and plated on ampicillin, kanamycin and ampicillin/kanamycin resistant plates in serial dilutions from $10^{-1}$ to $10^{-6}$ and grown overnight. The colonies percentage of double transformants was calculated by dividing the number of CFU's on the dual antibiotic plate by the CFU's of the kanamycin plate. This procedure was repeated with library plasmid DNA in place of pJK_proJK1_LGK at a 100:1 ratio to determine the percentage of double transformants.

### 2.3.5 Growth-based Selections

Library cell stocks containing mixtures of pJK-series eGFP expression plasmids were thawed on ice and washed with M9 minimal media [23]. Cultures were inoculated to an $OD_{600}$ of 0.03 in M9 minimal media supplemented with 4 g/L glucose and carbenicillin (50 µg/mL). Cultures were grown at 37°C and 250 rpm to an $OD_{600}$ of 0.6. Cell growth was monitored every 45 minutes by $OD_{600}$ measured on a Genesys 20 spectrophotometer (Thermo Fisher Scientific, Waltham, MA). Cells were washed with M9 media. The cells were used to re-inoculate 2.5 mL of fresh media to an $OD_{600}$ of 0.03. Cultures were again grown to an $OD_{600}$ of 0.6 (8.6 total population-averaged generations). Following selection cells were stored in 1mL of M9 media and

7% (v/v) DMSO at -80°C until bacterial plasmid DNA was extracted using a Qiagen miniprep kit (Qiagen, Valencia, CA).

### 2.3.6 Primer Design

Two sets of primers were used to amplify stretches of DNA for sequencing. The inner set of primers was designed to be complementary to the regions of DNA at the 5' and 3' ends of the gene tile of interest. Forward and reverse primers were designed to have melting temperatures around 55°C. Sequences for the outer, universal set of primers were taken from the TruSeq Small RNA Sample Prep Kit. The outer primers attach the Illumina barcodes and adaptors for sequencing and are listed in Kowalsky et al.[22]

### 2.3.7 Gene tile amplification

Gene tiles are amplified by two-step PCR. The contiguous region containing mutations is amplified using tile-specific inner primers using Phusion High Fidelity Polymerase (NEB M0530). The three different methods used to amplify the target region are described in Table 1. 5 µl of the PCR products were run on a 2% agarose gel and visualized with SYBR-GOLD (Invitrogen) to ensure the presence of a single band of the expected size (~250 bp). Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) were used per the manufacturer's protocol to purify the PCR product. Samples were multiplexed using index sequences on the outer primers.

DNA concentrations were quantified using Quant-iT PicoGreen (Life Technologies, Carlsbad, CA) quantification and samples were mixed in equimolar quantities for sequencing. Library DNA was sequenced on an Illumina MiSeq with 150-bp PE reads.

**Table 1- Reaction Conditions for Illumina Sequencing Prep**

| | Method A | Method B | Method C |
|---|---|---|---|
| **Reaction Conditions (50uL)** | 10µL 5x Phusion Buffer<br>1µL 10mM dNTPs<br>2.5µL 5µM inner F primer<br>2.5µL 5µM inner R primer<br>2.5µL 10µM outer F primer<br>2.5µL 10µM outer R primer<br>0.5µL Phusion HF Polymerase<br>1ng template<br>Water to 50µL | 10µL 5x Phusion Buffer<br>1µL 10mM dNTPs<br>2.5µL 10µM inner F primer<br>2.5µL 10µM inner Rprimer<br>0.5µL Phusion HF Polymerase<br>1ng template<br>Water to 50µL | 10µL 5x Phusion Buffer<br>1µL 10mM dNTPs<br>2.5µL 10µM inner forward primer<br>2.5µL 10µM inner reverse primer<br>0.5µL Phusion HF Polymerase<br>1ng template<br>Water to 50µL |
| **PCR Conditions** | 98°C for 30s<br>25 cycles of:<br>98°C for 5s<br>53°C for 15s<br>72°C for 15s<br>72°C for 10 min | 98°C for 30s<br>16 cycles of:<br>98°C for 5s<br>53°C for 15s<br>72°C for 15s<br>72°C for 10 min | 98°C for 30s<br>9 cycles of:<br>98°C for 5s<br>53°C for 15s<br>72°C for 15s<br>72°C for 10 min |
| **Add** | N/A | 1.875µL 1:10 diluted ExoI | 0.5µL Phusion HF Polymerase<br>2.5µL 10µM outer forward primer<br>2.5µL 10µM outer reverse primer |
| **PCR Conditions** | N/A | 37°C for 30 min<br>95°C for 5 min | 98°C for 30s<br>14 cycles of:<br>98°C for 5s<br>53°C for 15s<br>72°C for 15s<br>72°C for 10 min |

**Table 1 (Cont'd)**

| | | | |
|---|---|---|---|
| **Add** | N/A | In new PCR tube:<br>10µL 5x Phusion Buffer<br>1µL 10mM dNTPs<br>2.5µL 10µM inner F primer<br>2.5µL 10µM inner R primer<br>0.5µL Phusion HF Polymerase<br>1µL product from prev. step<br>Water to 50µL | N/A |
| **PCR Conditions** | N/A | 98°C for 30s<br>16 cycles of:<br>98°C for 5s<br>53°C for 15s<br>72°C for 15s<br>72°C for 10 min | N/A |

### 2.3.8 Data Analysis

Enrich 0.2 software was used to compute enrichment ratios of individual mutants from the raw Illumina sequencing files [27]. Forward and reverse reads obtained for each section were used as input. Modifications were made to Enrich 0.2 in order to accommodate shifted and shortened protein alignment sequences. Enrichment ratios that were obtained were normalized as detailed below using custom scripts.

## 2.4 Theory

### 2.4.1 Normalization for Growth Rate Selections

When cells grow exponentially, the specific growth rate, $\mu_i$, of any individual mutant $i$ can be written as:

$$\mu_i = \ln\left(\frac{x_{fi}}{x_{oi}}\right)\frac{1}{t} \tag{1}$$

Where $x_{fi}$ is the final concentration of the mutant, $x_{oi}$ is the initial concentration, and $t$ is the time difference between the initial and final concentration of cells. In this formulation we are explicitly neglecting the effect of lag phases for growth. The equation for calculating the enrichment ratio, $\varepsilon_i$, of the same mutant is:

$$\varepsilon_i = \log_2\left(\frac{f_{fi}}{f_{oi}}\right) \tag{2}$$

Where $f_{fi}$ is the final frequency of the mutant in the library population and $f_{oi}$ is the initial frequency. These frequencies can be converted to cell concentrations by the equations below:

$$f_{oi} = \frac{x_{oi}}{\sum x_{oi}} \tag{3}$$

$$f_{fi} = \frac{x_{fi}}{\sum x_{fi}} \tag{4}$$

Where $\Sigma x_{oi}$ is the initial concentration of the culture and $\Sigma x_{fi}$ is the final concentration. The enrichment ratio can be rewritten as:

$$\varepsilon_i = \log_2\left(\frac{x_{fi}}{x_{oi}}\right) - \log_2\left(\frac{\sum x_{fi}}{\sum x_{oi}}\right) \tag{5}$$

Combining this equation with (1) leads to:

$$\mu_i \log_2 e = \frac{1}{t}\left(\varepsilon_i + \log_2\left(\frac{\sum x_{fi}}{\sum x_{oi}}\right)\right) \tag{6}$$

We can define the change in culture density between the initial and final conditions in terms of the number of average doubling periods ($g_p$) according to:

$$\log_2\left(\frac{\sum x_{fi}}{\sum x_{oi}}\right) = \#of\ Doublings = g_p \tag{7}$$

Similarly, we can remove time from (5) by redefining it as:

$$t = \frac{g_p \ln 2}{\mu_p} \tag{8}$$

where $\mu_p$ is equal to the bulk average growth rate of the population between the initial and final conditions.

Combining (7) and (8) into (6) leads to a description of the growth rate of mutant *i* as a function of its enrichment ratio:

$$\mu_i = \overline{\mu}_p\left(\frac{\varepsilon_i}{g_p} + 1\right) \tag{9}$$

It is often helpful to express the fitness of mutant *i,* $\zeta_i$, normalized to the growth rate of the starting construct (wild-type; $\mu_{wt}$)

$$\zeta_i = \log_2\left(\frac{\mu_i}{\overline{\mu}_{wt}}\right) \tag{10}$$

$$\zeta_i = \log_2\left(\frac{\frac{\varepsilon_i}{g_p}+1}{\frac{\varepsilon_{wt}}{g_p}+1}\right) \tag{11}$$

Since the starting construct is usually included in the population, fitness of each variant *i* can be normalized across different selection experiments given only the number of doubling periods as well as the enrichment ratios for the mutant and wild-type construct.

We can also rewrite the enrichment ratio as a function of growth rate:

$$\varepsilon_i = g_p\left(\frac{\mu_i}{\bar{\mu}_p}-1\right) \tag{12}$$

The enrichment ratio will increase linearly with the number of doubling periods so long as a mutant is able to exceed the population-averaged growth rate.

### 2.4.2 Theoretical effects of double transformation on enrichment ratios for growth-based selections.

Consider a microorganism transformed with a plasmid variant *i*. When grown exponentially, the time-dependent concentration of the culture ($x_{fi}$) can be written:

$$x_{fi} = x_{oi}e^{\mu_i t} \tag{13}$$

where $x_{oi}$ is the initial concentration, $\mu_i$ is the specific growth rate, and *t* is time.

Consider now a microorganism harboring two different plasmids: plasmid variant *i* and an unrelated plasmid variant *j*. For all variants $1 \leq j \leq$n in the population, the growth of microbes transformed with plasmid variant i and all other variants in the population ($x_{f\Phi i}$) can be represented by:

$$x_{f\phi i} = x_{o\phi i}\sum_{j=1}^{j=n} f_j e^{\frac{\mu_i+\mu_j}{2}t} \tag{14}$$

Where $f_j$ represents the frequency of plasmid $j$ in the sequenced population. The form of this growth equation is based on an assumption that a microbe double transformed with plasmids $i$ and $j$ will grow at an average of their individual growth rates (see below for further discussion).

Taking the doubly transformed population into account, the time-dependent concentration of cells harboring plasmid $i$ can be written as:

$$x_{fia} = (1-\phi)x_{fi} + \phi x_{f\phi i} \tag{15}$$

where $\phi$ is the fraction of the population that is doubly transformed.

The actual, measured enrichment ratio can be represented by:

$$\varepsilon_{ia} = \log_2\left(\frac{x_{fia}}{x_{oi}}\frac{\sum x_{oi}}{\sum x_{fi}}\right) \tag{16}$$

and the true enrichment ratio, defined as the enrichment ratio obtained in the absence of double transformants, can be represented by:

$$\varepsilon_{it} = \log_2\left(\frac{x_{fi}}{x_{oi}}\frac{\sum x_{oi}}{\sum x_{fi}}\right) \tag{17}$$

The measured enrichment ratio can also written as:

$$\varepsilon_{ia} = \log_2\left(\frac{x_{fia}}{x_{fi}}\frac{x_{fi}}{x_{oi}}\frac{\sum x_{oi}}{\sum x_{fi}}\right) = \log_2\left(\frac{x_{fia}}{x_{fi}}\right) + \varepsilon_{it} \tag{18}$$

Where $\log_2\left(\frac{x_{fia}}{x_{fi}}\right)$ represents a correction factor to the true enrichment ratio because of the doubly transformed populations. We can further represent $\frac{x_{fia}}{x_{fi}}$ as:

$$\frac{x_{fia}}{x_{fi}} = \frac{(1-\phi)x_{fi}+\phi x_{f\phi i}}{x_{fi}} = \left(1-\phi\right)+\phi\frac{x_{f\phi i}}{x_{fi}} \tag{19}$$

Substituting (13) and (14) into (19):

$$\frac{x_{fia}}{x_{fi}} = \left(1-\phi\right)+\phi\frac{x_{o\phi i}}{x_{oi}}\frac{\sum_{j=1}^{j=n} f_j e^{\frac{\mu_i+\mu_j}{2}t}}{e^{\mu_i t}} \tag{20}$$

Since

$$\frac{x_{o\phi i}}{x_{oi}} = \frac{\phi}{1-\phi} \tag{21}$$

We can rewrite (20) as:

$$\frac{x_{fia}}{x_{fi}} = \left(1-\phi\right)+\frac{\phi^2}{1-\phi}\sum_{j=1}^{j=n} f_j e^{\frac{t}{2}\left(\mu_j-\mu_i\right)} \tag{22}$$

As before, time can be represented in terms of population doubling periods, $g_p$, and the average growth rate of the population ($\bar{\mu}_p$):

$$t = \frac{g_p \ln(2)}{\bar{\mu}_p} \tag{23}$$

Combining terms, we derive the following form for the correction factor:

$$\log_2\left(\frac{x_{fia}}{x_{fi}}\right) = \log_2\left[\left(1-\phi\right)+\frac{\phi^2}{1-\phi}\sum_{j=1}^{j=n} f_j e^{\frac{g_p \ln(2)}{2\bar{\mu}_p}\left(\mu_j-\mu_i\right)}\right] \tag{24}$$

This correction factor is a function of the double transformation rate, the number of doubling periods, the average growth rate of the population, and the distribution of the starting growth rates in the beginning population. Because we do not know the distribution of growth rates

in a given population *a priori*, this correction factor cannot reliably be used. However, it can help illuminate conditions where the correction factor is expected to small.

The main assumption in deriving this correction factor is that doubly transformed cells grow at an average rate of cells transformed with the individual plasmids. This assumption is likely to be correct for plasmids that are segregated evenly upon cell division, as well as where activity of an individual protein is linearly proportional to the growth rate of cells harboring its plasmid. More complicated growth models can be assessed using the framework laid out here.

### 2.4.3 Normalization for Fluorescence-Activated Cell Sorting

For comparisons of variants across different populations, we desire a method to reconstruct mean fluorescence for each mutant, $\overline{F}_i$ from its enrichment ratio $\varepsilon_i$. In fluorescence-activated cell sorting (FACS), populations are screened by collecting cells with fluorescence above a certain gating threshold. A clonal population of cells will exhibit a mean fluorescence with a certain variance according to cell size, surface density of displayed proteins, or other factors. Thus, only a fraction of cells for each variant will exceed the fluorescence threshold needed for collection. Since fluorescence measurements of clonal population of cells are log-normally distributed in flow cytometry, $\overline{F}_i$ can be determined using regular statistical calculations:

$$\overline{F}_i' = \ln\left(F_g\right) - \sigma' \sqrt{2}\, erf^{-1}\left(1 - 2\frac{x_{fi}}{x_{oi}}\right) \tag{25}$$

Here, $\overline{F}_i'$ is the mean of the natural log of the fluorescence for variant *i,* $\sigma'$ is the natural log of the standard deviation of the data, $F_g$ is the fluorescence gating threshold for the experiment, and the ratio $\frac{x_{fi}}{x_{oi}}$ is the fraction of variant *i* that is collected above the gating threshold. $\overline{F}_i'$ can be determined from $F_{wt}'$ by:

$$\overline{F}_i = e^{\left(\overline{F}_i' + \frac{\sigma'^2}{2}\right)} \tag{26}$$

It remains to find $\frac{x_{fi}}{x_{oi}}$ in terms of experimentally measurable values. The flow cytometer used to analyze the culture records the percentage of the total population sampled that is collected, $\phi$. This value can be written as:

$$\phi_i = \frac{\sum x_{fi}}{\sum x_{oi}} \tag{27}$$

From sequencing data, the enrichment ratio of each mutant, ε, is also known and can be written as:

$$2^{\varepsilon_i} = \frac{x_{fi}}{x_{oi}} \frac{\sum x_{oi}}{\sum x_{fi}} \tag{28}$$

Combining equations (26-28), we end up with:

$$\frac{x_{fi}}{x_{oi}} = \phi 2^{\varepsilon_i} \tag{29}$$

Finally, combining this relation into equations (25-26) leads to:

$$\overline{F}_i = \exp\left[\frac{\sigma'^2}{2} + \ln\left(F_g\right) - \sigma'\sqrt{2}\, erf^{-1}\left(1 - \phi 2^{\varepsilon_i + 1}\right)\right] \tag{30}$$

As with the growth-based selection, it is often helpful to express the fitness of mutant *i* normalized to the fluorescence of the starting construct (wild-type; $\overline{F}_{wt}$)

$$\zeta_i = \log_2\left(\frac{\overline{F}_i}{\overline{F}_{wt}}\right) \tag{31}$$

$$\zeta_i = \log_2\left[e\sqrt{2}\sigma'\left(erf^{-1}\left(1 - \phi 2^{\varepsilon_{wt}+1}\right) - erf^{-1}\left(1 - \phi 2^{\varepsilon_i+1}\right)\right)\right] \tag{32}$$

To normalize fluorescence measurements, $\phi$ is set by the experiment, and the enrichment ratios $\varepsilon_i$ and $\varepsilon_{wt}$ are obtained from analysis of the raw sequencing files. In this derivation we assume

that the log-transformed standard deviation is the same between the individual variant and the wild-type sequence. We have not rigorously tested this assumption. Note that the form of the fitness metric used in this work has the standard deviation as a scalar which is unlikely to vary much using the same cell type and flow cytometer; thus, ratios of fluorescence measurements can be related between populations using only the enrichment ratios and the gating threshold.

### 2.5 Results and Discussion

We have developed a standardized method to map the sequence-function relationships of entire gene sequences encoding full-length proteins. This process is applicable to a wide variety of proteins, including binding proteins, fluorescent proteins, and enzymes. With some modifications, the method can also be extended to membrane proteins and transcription factors. The protein class determines the selection method: binding proteins can be screened or sorted using phage or yeast display techniques, whereas growth-based selections are preferable for enzymes [9, 12, 17, 28, 29]. Regardless of the protein category, the initial sequence should encode some level of functional activity as a basis to distinguish active and inactive proteins.

Figure 2 outlines the basic steps covering target selection, gene tiling, library preparation, selection, deep sequencing library preparation, and data analysis and normalization. We have written custom scripts and modified published scripts to facilitate data generation and analysis. Additionally, we have formulated optimal selection criteria and derived equations governing the normalization of results across different selection conditions. Practical considerations for each step are listed in Supplementary Note S1 (Appendix A). In the following sections we consider each step in the overall process in detail.

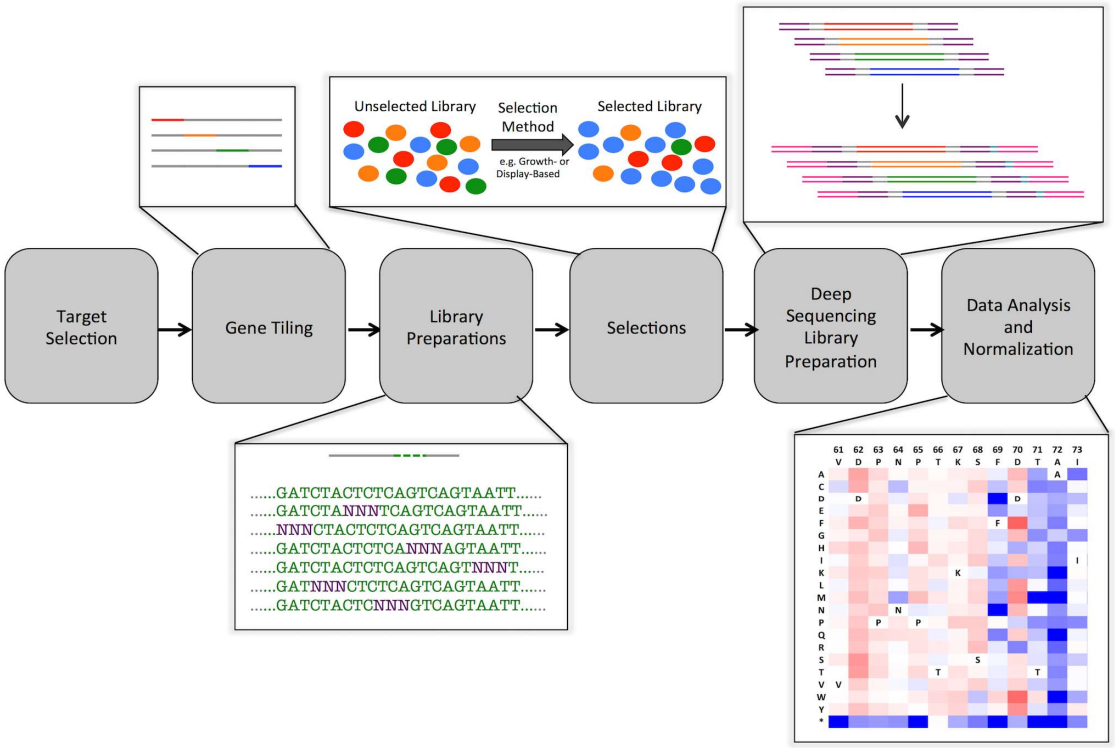



**Figure 2- Overview of high-resolution sequence-function mapping process.**

*Target Selection.* Proteins of interest are selected for interrogation of sequence function relationships. A plasmid containing the gene-encoding sequence is generated. *Gene tiling.* Starting from this gene sequence, semi-overlapping tiles are generated to cover the entire gene. These tiles are either 150, 250 or 300 bp in length in order to be sequenced in paired-end mode on Illumina deep sequencing platforms. *Library Preparation*. The single-pot PFunkel method is used to generate a comprehensive single-site saturation mutagenesis library. *Selections.* Growth-based selections and FACS screens are used to resolve library populations; these selections should not completely converge on a few members of the population. It is important that the initial protein shows activity toward the selection method. *Deep Sequencing and Library Preparation*. After selection, cells are lysed and plasmid DNA is purified. The specific mutated tile region of the gene of interest is then amplified using overhang PCR, at which time Illumina sequencing primers and adaptors with selection-specific indexes are attached. *Data analysis and normalization.* Barcoded DNA is sequenced on a standard Illumina platform, analyzed, and normalized using custom scripts. The end result of this analysis is a comprehensive portrait of the effects of sequence on function for thousands of single point mutants in the gene of interest. These portraits can be used for various purposes such as improving protein binding affinity and specificity or improving enzymatic catalytic efficiency.

### 2.5.1 Gene Tiling

A protein of 250 residues is encoded by a gene of 750 bp, which is longer than high-quality read lengths of existing sequencing platforms. Previous approaches to map sequence to function for full-length proteins involved sequencing the entire gene as smaller amplified segments (Figure

3a) [9]. Because there should be only one mutation per gene, reads from amplified regions other than the one containing the mutation yield no information and are wasted. Figure 3b shows the percentage of total sequence reads that provide information as a function of gene length. As gene length increases, the percentage of usable sequencing data decreases and, consequently, more reads are needed to ensure proper coverage. For example, using this previous method results in usable information for only 33% of the sequencing reads in a gene of length 450 bp.

We have improved the efficiency of scanning long genes by dividing the gene into multiple "tiles," each of which is effectively treated as a distinct gene. Each tile is independently mutagenized, subjected to selection, and sequenced before our analysis pipeline normalizes and merges the count data to generate the sequence-function map of the full gene. Tile regions are designed to be slightly shorter than a sequencing read, and within each parallel mutagenesis reaction, mutations are restricted to the corresponding tile. For example, tiles designed for 150-bp read lengths would consist of a central 120-bp mutated region flanked by 15-bp constant regions for PCR primer annealing. Multiple, partially overlapping libraries are prepared for each gene to ensure full coverage of the protein. This approach eliminates excess wild-type sequencing because only the region containing the mutation is sequenced (Figure 3c). However, the tradeoff is that assessing the function of a full-length sequence requires multiple independent selections. Since population dynamics may vary among selections the enrichment ratios must be normalized to allow comparisons across tiles (see Theory section and below).

### 2.5.2 Library Mutagenesis Preparation

Our objective is to map the function of every single nonsynonymous (NS) mutation of a protein-encoding sequence. In an ideal system, 1) There would be exactly one NS mutation per protein-encoding sequence; 2) The library would contain complete uniform coverage of all

possible single NS mutations; 3) The library prep method would be as reliable, fast and inexpensive as possible; and 4) Each cell would harbor a single protein-encoding sequence.

Numerous methods have been described for the creation of mutant libraries [30-33]. Certain protocols, like QuikChange or Kunkel mutagenesis, introduce mutations at specified locations with specific primers. Because each residue targeted for mutation requires a separate primer and a separate reaction, creation of a single-site saturation mutagenesis (SSM) library for a 250- residue protein requires 250 unique primers and 250 separate reactions, limiting scalability. A newly developed method named Pfunkel incorporates the benefits of Kunkel mutagenesis while minimizing library preparation time by combining the individual SSM reactions into a single-pot [25].

To evaluate the performance of Pfunkel, we created a SSM library of the first forty residues on a codon-optimized gene encoding levoglucosan kinase (LGK) from *L. starkeyi* (GenBank: EU751287.1) [34]. A SSM library incorporating NNN codons should theoretically contain 2520 (63 codons at 40 positions) unique NS mutations. The mutagenesis primer set was manually designed using the Agilent QuikChange primer design calculator, and a Pfunkel reaction was performed essentially as described in Firnberg *et al.* [24, 25]. The resulting library was sequenced using 150-bp paired-end (PE) reads on an Illumina MiSeq. The quality of the mutagenesis procedure was evaluated based on the percent coverage of mutations at the DNA and amino-acid levels, the percentage of starting (wild-type) DNA sequences, and the percentage of sequences with more than one mutation in the coding sequence.

Coverage analysis of the SSM library showed 99% of the 2520 possible codon mutations were incorporated into the SSM library. Additionally, we observed 100% coverage of single base mutations and coverage of two and three base substitutions higher than previously reported (Table

2). The number of transformed colonies in the Pfunkel procedure did not impose a bottleneck on library complexity since the number of transformed colonies exceeded the library size by seven-fold, corresponding to a theoretical 99.9% library coverage [35]. Based on this analysis, we independently conclude that Pfunkel can produce comprehensive SSM libraries. The single-pot reaction can produce high-coverage SSM libraries in two days with minimal hands-on time.

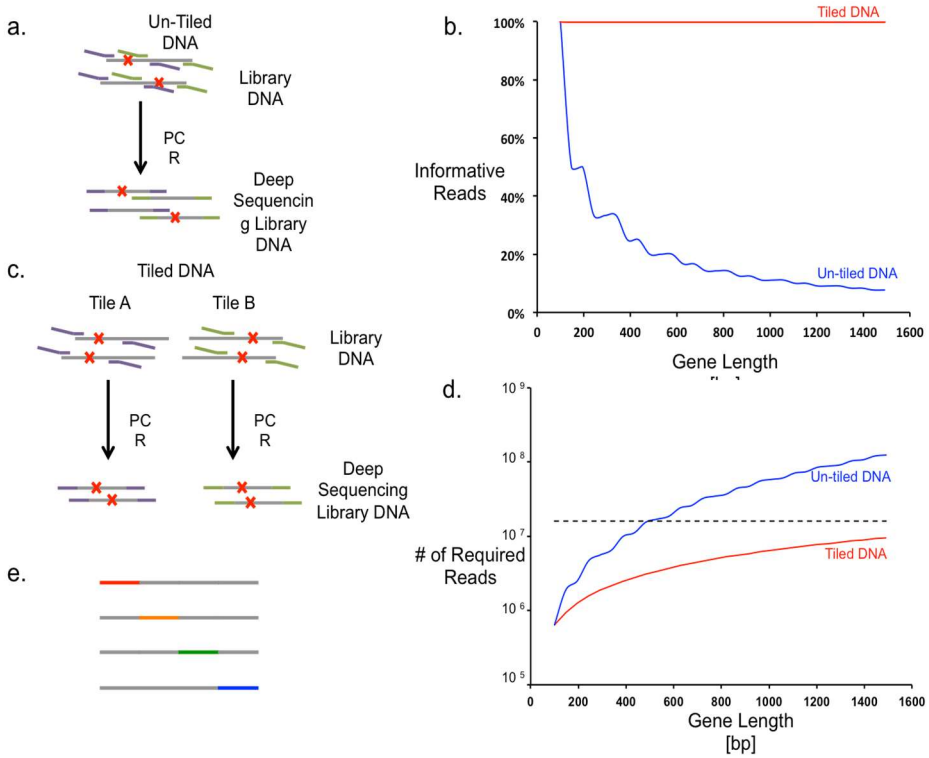



**Figure 3- Gene tiling increased the efficiency of deep sequencing for sequence-function mapping.**

a. Deep sequencing without using gene tiles. Gene sequences are represented by grey lines, mutations by red x's, and sequencing primers by purple and green lines. Several previous methods to amplify target DNA (left) amplify both mutated and non-mutated regions. The latter result in wasted reads and increase the sequencing capacity necessary to resolve the entire library. b. Percent of usable reads as a function of gene length with (red line) and without (blue line) gene tiling. c. Gene tiling has the ability to reduce the number of DNA sequencing reads necessary by targeting the region with a mutation in PCR amplification for sequencing purposes. To implement gene tiling, separate libraries are prepared and sorted for each tile. d. Number of sequence reads required for 300-fold average coverage of nonsynonymous mutations with (red line) or without (blue line) gene tiling. The horizontal dashed line represents the average number of DNA sequences from a single MiSeq lane. e. In gene tiling, short contiguous stretches of DNA (tiles) are targeted for mutations. Gene tiles are indicated by the colored dashed lines and cover the entire gene sequence among the different libraries.

While Pfunkel is a simple and reliable method to create high-coverage SSM libraries, the costs associated with primer synthesis are not trivial. For a protein of length *L*, the cost of the primer set is \$3.90**L* (\$0.10 per base and 39 bases per primer) (2014, Integrated DNA Technologies, Corralville, IA). Accordingly, we looked for ways to improve the Pfunkel method by reducing method cost. 1.) Shorter primer lengths would decrease cost. Our initial primer set was designed using a QuikChange calculator that suggested longer primer lengths than the custom primer design script provided by in the Pfunkel paper. 2.) Recovering plasmid DNA in liquid culture would reduce both cost and time. In the current procedure following transformation, cells are plated on expensive BioAssay plates.

We hypothesized that shorter primers would produce SSM libraries with equally high coverage but a decreased percentage of reads containing exactly one mutation. To test this we produced a second primer set using the custom primer design script from Firnberg et al. (referred to as scripted Pfunkel primers) [25]. This primer set averaged 27 bp in length while the QuikChange primers were, on average, 39 bp. We evaluated the two primer sets using three variables that contribute to inefficient sequencing: 1.) percentage of wild-type reads; 2.) fractional library coverage; and 3.) the number of double mutants. In comparison to the QuikChange primer set, libraries prepared with the scripted Pfunkel primers had a much higher rate of wild-type sequences (62.6%), lower library coverage (99.5%), but a lower rate of double mutants (3.3%) (Table 2). Although for the QuikChange primer set there is a higher rate of double mutants, all are accounted for in the sequencing and so do not influence later data analysis. Thus, the cost of synthesizing longer QuikChange primers is more than balanced by the benefit of a high-quality SSM library, which requires fewer DNA sequencing reads for full library coverage.

In the original Pfunkel method, following transformation cells were plated on large BioAssay plates and grown at 37°C overnight. Recovering the library in solution without plating could save cost and time. To determine whether library quality suffers without plating, two parallel Pfunkel reactions were performed with the QuikChange or Pfunkel Scripted primer sets. Following transformation, half of the cells were plated on a selective plate while the other half was grown to an $OD_{600}$ of 0.1 in a liquid culture. Cells were then harvested, and plasmid was recovered and sequenced. Cells recovered in liquid culture showed 73-93% coverage of all possible codon substitutions, much lower than the 98.3-99.3% observed for the libraries that were plated. The liquid culture data also showed a bias against NS mutations. For example, using the QuikChange primer set and plating the cells resulted in 70.6% of the reads containing exactly one NS mutation, whereas growing cells in culture resulted in only 40.0% of reads containing one NS mutation (Table 2). Based on these experiments, we conclude that plating cells following transformation is necessary to produce high-quality SSM libraries.

Theoretically, mutational frequencies caused by saturation mutagenesis with NNN codons should be equal across bases. However, consistent with the results presented by Firnberg *et al.*, we find that guanosine (G) bases are enriched relative to theoretical predictions (Table 2) [25]. Since we see very little difference in the incorporation of single bases at the DNA level between the two sets of mutagenic primers, the artificial enrichment of G bases is likely the result of improper machine mixing of the NNN mutations in primer synthesis, as previously suggested [25]. While hand mixing of the nucleotides during primer synthesis may reduce the bias, it would substantially increase primer cost. Alternatively, the enrichment of G bases could be introduced by a bias in primer annealing as suggested by Jain and Varadarajan [36]. Neverless, since in our protocols the average library member is counted at least 100 times, the observed level of bias is tolerable.

The plasmid DNA encoding the SSM library must be transformed into the host organisms used for selections. If multiple plasmids are transformed into a single cell, gain-of-function variants could potentially compensate for weaker variants. To account for this, we have derived a correction factor for the measured enrichment ratio as a function of percentage of double transformants (See Theory). For libraries with less than 10% doubly transformed cells this correction can be neglected because its absolute magnitude correction is less than 0.35 (Figure 4**)**, which is comparable to the experimental error in determining enrichment ratios from sequencing data for loss-of-function variants. However, at higher percentages of doubly transformed cells this effect may be significant (Figure 4**)** and controls must be run to minimize artifacts [37].

In our typical workflow we use *E. coli* for growth-based selections and *S. cerevisiae* for yeast display of binding proteins [38]. The most common yeast display plasmid contains a CEN6/ARSH4 ori maintaining a low plasmid copy number, such that co-transformed plasmids are segregated well before FACS [39]. For many *E. coli*-based systems however, plasmids of medium to high copy numbers do not efficiently segregate and the percentage of double transformants needs to be quantified. Goldsmith et al. suggested a strategy which we follow here [37]. The starting plasmid pJK_proJK1_kanR_LGK was modified by changing the antibiotic resistance from kanamycin to ampicillin, forming the plasmid pJK_proJK1_LGK. These two plasmids were mixed at a mass ratio of 1:100, respectively, and 40 ng of this mix was transformed into 40 μL of electrocompetent Tuner cells. The reaction was plated on ampicillin, kanamycin and

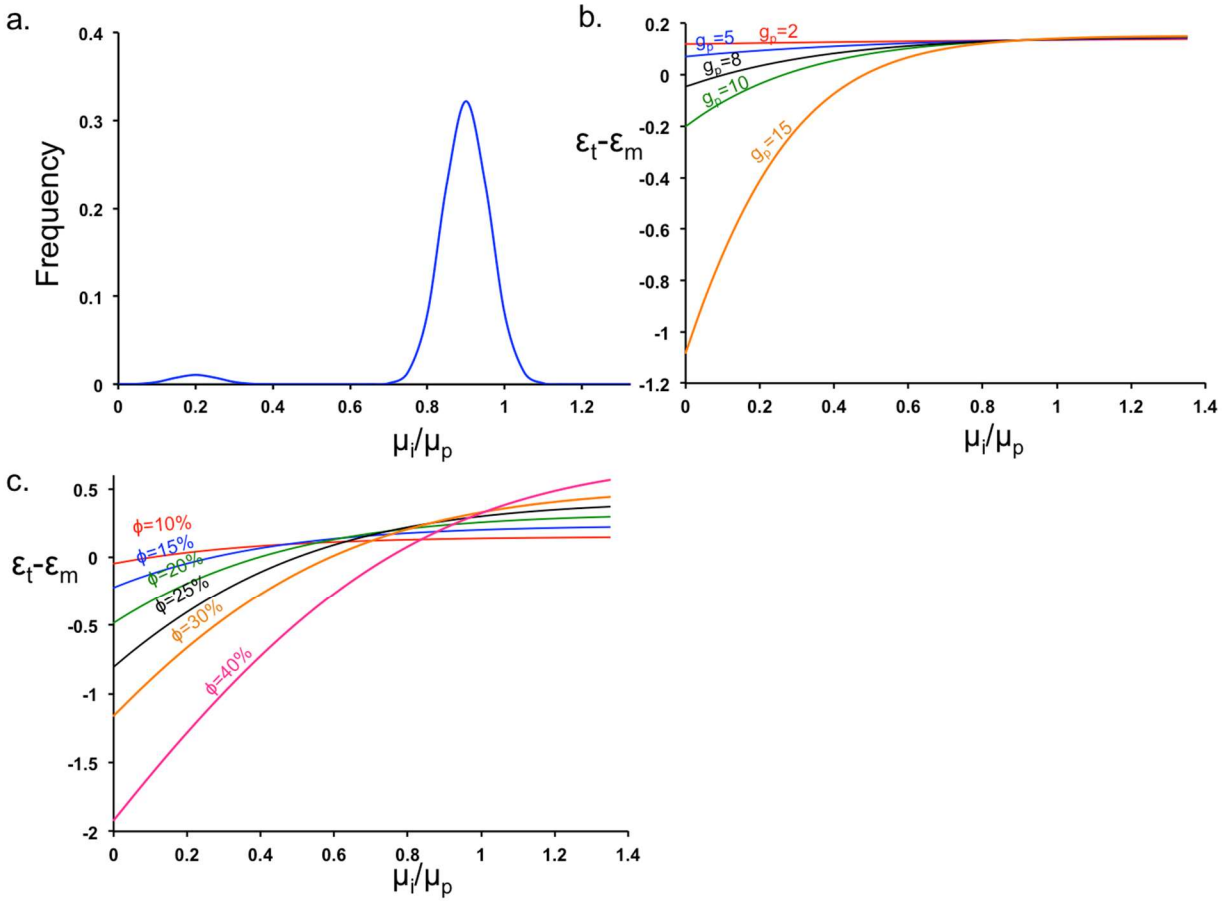



**Figure 4- Enrichment correction factor from double transformation artifacts.**

The enrichment correction factor is characterized by the true enrichment ratio ($\varepsilon_t$) minus the measured enrichment ratio ($\varepsilon_m$) for different, variant growth rates relative to the population growth rate. a. The distribution of the individual growth rates in the population was assumed to be a bimodal-guassian with means of $\mu_i/\mu_p$ = 0.2 and 0.9 and a standard deviation of 0.06, broadly consistent with individual variant growth rates observed for a library. b. Here the assumed double transformation rate is assumed to be 10% and the correction factor is plotted for different numbers of population doubling periods. (2 red, 5 blue, 8 black,10 green). c. Here the assumed population doubling periods is 8 and the correction factor is plotted for different double transformation rates. If the double transformation rate is less than 10% and the population doubling periods is about 8 the correction factor is negligible and does not need to be considered.

ampicillin/kanamycin selective plates and grown overnight. The colonies were counted and the percentage of double transformants was calculated by taking the ratio of the number of dual antibiotic resistant colonies over the number of solely kanamycin resistant colonies (Table 3). Under these conditions, the rate of double transformants is on the order of 2%, well below the 10%

threshold. Additionally, in this specific case, we found that the number of solely ampicillin transformants is more than sufficient to support the degeneracy of a library size of 2,560. We recommend re-running this experiment for every library, as transformation conditions often vary.

### 2.5.3 Selections

Methods for selection are chosen based on protein function. Selections should be designed such that the widest range of activity levels can be resolved. Protein binding activity is usually screened/sorted by phage, bacterial, or yeast display platforms [40-43]. The latter two methods resolve the population by FACS; in this section we derive equations governing the grouping of different variants by FACS and suggest optimal experimental parameters. We also show equations that govern the appropriate choice of experimental parameters for growth-based selections.

To ensure proper coverage, selections are designed such that on average there is 200-500 fold coverage of each variant in the unselected population. Sequencing to this depth requires on the order of 500,000 quality 150-bp reads. The enrichment of an individual variant is described as the $\log_2$ ratio of its frequency in the selected population to the unselected population. For a library containing 2,500 members and sampled at 200-fold coverage, there is a lower enrichment limit of -7.5 for mutants counted once after selection and an upper bound of 11.3 for a variant that completely overtakes the population. Because intrinsic error (Poisson noise) is lowered when the counting threshold is set much higher than 1, and because allowing a single variant to overtake the population provides no data about the remaining positions, the practical dynamic range for the selection range spans enrichment values of -4 to 4. Selections should be designed to best span this range of enrichment values. The dynamic range will vary minimally with

**Table 2 - Mutagenesis statistics for different experimental conditions. Comparisons to theoretical predictions and previous literature data are shown as reference [25].**

| | *Theoretical* | *Firnberg et al. results* [25] | QuikChange Primers | | Scripted Pfunkel Primers | |
|---|---|---|---|---|---|---|
| | | **Plated** | **Plated** | **Culture** | **Plated** | **Culture** |
| **Sequences (reads)** | | *787,488* | 414,410 | 319,179 | 510,126 | 436,135 |
| NNN primer base composition | | | | | | |
| **T** | *25.0%* | *17.1%* | 15.2% | 16.7% | 22.8% | 19.2% |
| **A** | *25.0%* | *18.3%* | 14.0% | 15.7% | 20.2% | 19.7% |
| **C** | *25.0%* | *18.3%* | 17.5% | 18.5% | 16.8% | 19.8% |
| **G** | *25.0%* | *46.3%* | 53.3% | 49.0% | 40.2% | 41.4% |
| Percent of possible codon substitiutions observed | | | | | | |
| **1-base substitution** | | *99.6%* | 100.0% | 100.0% | 100.0% | 100.0% |
| **2-base substitutions** | | *97.7%* | 99.6% | 96.4% | 98.6% | 83.7% |
| **3-base substitutions** | | *95.3%* | 98.7% | 88.3% | 97.3% | 54.4% |
| **All substitutions** | | *97.0%* | 99.3% | 93.5% | 98.3% | 73.5% |
| Percent of reads with | | | | | | |
| **No nonsynonymous mutations** | *1.6%* | *26.2%* | 22.4% | 55.0% | 62.6% | 87.0% |
| **One nonsynonymous mutation** | *98.4%* | *56.7%* | 70.6% | 40.0% | 34.1% | 12.0% |
| **Multiple nonsynonymous mutations** | *0.0%* | *17.1%* | 7.0% | 5.0% | 3.3% | 1.0% |
| Coverage of possible single nonsynonymous mutations | | | 99.9% | 98.0% | 99.5% | 89.1% |

**Table 3- Experimental results used for determination of double transformation percentage.**

| *Mass of Plasmid (ng)* | *Plasmid Antibiotic Resistance* | *Number of CFU's on Plate* | | | *% Double Transformants* |
|---|---|---|---|---|---|
| | | **Amp Plate** | **Kan Plate** | **Amp+Kan Plate** | |
| **0** | - | <10 | <10 | <10 | N/A |
| **10** | Amp | 690,000 | <10 | <10 | N/A |
| **5** | Kan | <10 | 250,000 | <10 | N/A |
| **40** | 100:1 Amp:Kan | 2,200,000 | 16000 | 310 | 1.9% |

increasing library coverage: every 2-fold increase in coverage results in decreasing the lower bound of the enrichment ratio by 1 unit. Figure 5 shows a mutant with an enrichment ratio of -4 and makes up 0.0024% of the selected population, while a variant with an enrichment ratio of 4

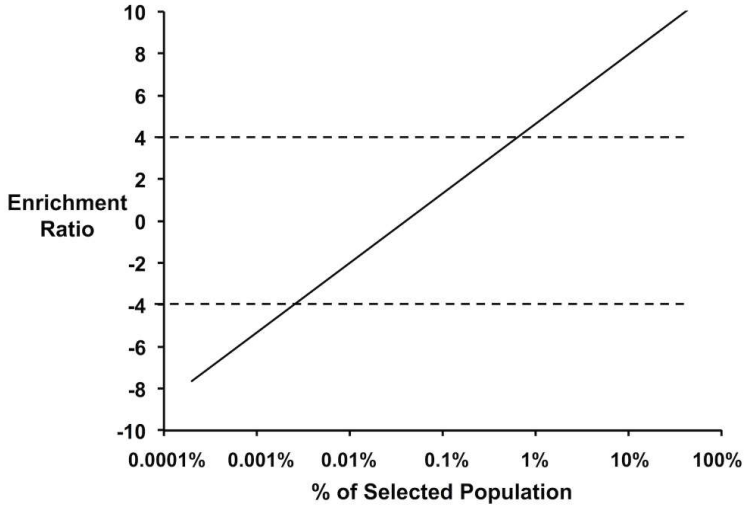



**Figure 5- Enrichment ratio of a clone as a function of its abundance in the selected population.**

The dynamic range of the method lies between enrichment ratios of -4 to 4 (indicated by horizontal dashed lines) such that (i.) single clones do not dominate the selected population; and (ii.) loss-of-function clones are not completely removed from the population.

and makes up 0.8% of the selected population. Among 500,000 sequence reads from the selected library, the former variant is observed 12 times and the latter 3,400 times.

#### 2.5.3.1 Enzymes: Growth selections

The enrichment value ($\varepsilon_i$) of an individual variant *i* depends on the average growth rate of the library population ( $\overline{\mu}_p$ ), the number of doubling times the culture is allowed to grow ($g_p$), and the growth rate of the individual variant ($\mu_i$):

$$\varepsilon_i = g_p\left(\frac{\mu_i}{\overline{\mu}_p} - 1\right) \tag{12}$$

Growth selections should be designed such that the number of generations the culture is allowed to grow fits a reasonable time frame (under 2 days) and there is high resolution of fitness for the entire library. Figure 6 shows the enrichment ratios for a range of specific growth rates relative to the population-averaged growth rate for different numbers of doubling times. According to these results, the dynamic range of protein activities is maximized between five and ten doubling times. This range allows resolution of all variants with growth rates above 0.2 of the population-averaged growth rate. Furthermore, limiting the number of doublings minimizes the effect of spontaneous mutations in the background strain [44].

#### 2.5.3.2 Protein Binders/Transcriptional Regulators/Membrane Proteins: FACS Screens

FACS is used in many different screening scenarios including protein binding, transcriptional activation, gene silencing, and localization studies [9, 10, 45-47]. In each of these screens the presence of cellular fluorescence corresponds to some underlying protein activity. In yeast

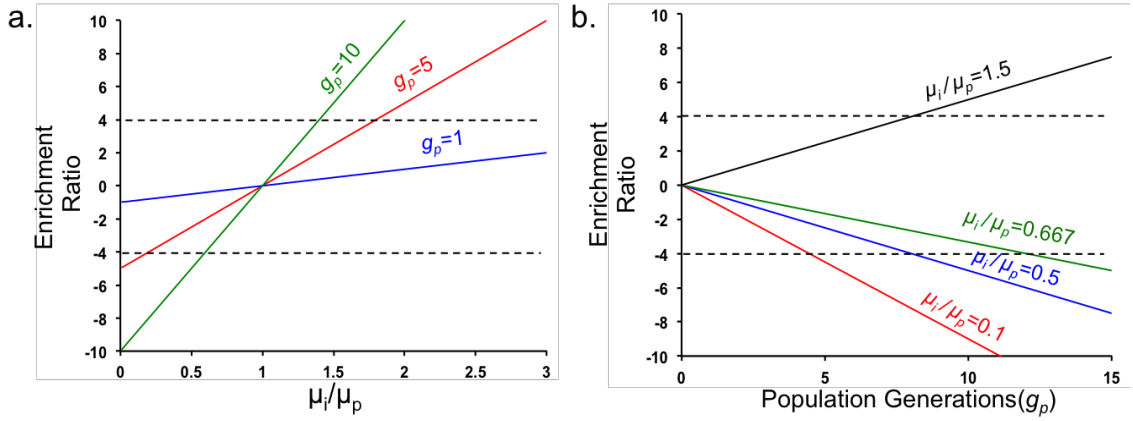



**Figure 6- Growth Selection Parameters.**

The parameters of growth-based selections should be chosen such that the range of enrichment ratios for the population lies between -4 and 4. a. Enrichment ratios as a function of the individual growth rate compared to the population growth rate. Following one generation of population growth (blue) the enrichment ratios remain around zero. Increasing the number of population generations the experiment is allowed to grow (1 generation, blue, 5 generations, red and 10 generations, green) increases the experimental resolution in discriminating mutant growth phenotypes. b. Enrichment ratios as a function of average population generation (gp) for various μi/μp. For μi/μp values less than one (0.1, red; 0.5, blue; and 0.667, green) the enrichment ratios decrease with increasing population generations. Variants with values μi/μp values above one (1.5, black) show enhanced enrichment with increasing population generations.

display, the binding affinity of a given protein-protein or protein-small molecule interaction is assessed by binding of a biotinylated protein or small molecule (present at a concentration near the dissociation constant for the interaction) to a surface-displayed protein, followed by labeling with fluorescently-conjugated streptavidin [38]. In this case, higher fluorescence indicates increased binding affinity for the biotinylated protein.

The distribution of fluorescent intensity for individual cells is log normally distributed (Figure 2.6a) with a mean fluorescence $\overline{F}_i'$ and a clone-independent standard deviation $\sigma'$. To sort populations, square (normal to one axis) or diagonal gates (normalizing for surface expression) are usually drawn (Figure 2.6b); these gates sort a specific fraction of the population, $\phi$, that exceeds a gating fluorescence, $F_g$. Sorting cells using one-color (square gate) is most common. However,

two-color sorting (diagonal gating) is often used to correct for intrinsic noise caused by distributions in cell size, among other factors [48]. Similarly, two-color sorting can be used in protein display techniques to normalize for cell-to-cell variation in surface expression [49]. Two-color sorting results in a log normal distribution for the transformed fluorescence but with a significantly reduced standard deviation. As such, these sorts can be described by the FACS equations derived in the Theory section. Gating the top fraction of the fluorescent distribution enriches the sorted population in variants with enhanced activity. The enrichment ratio of a single clone can be described by rearranging equation (18):

$$\varepsilon_i = \log_2 \left[ \frac{\left( 1 - erf\left( \frac{\frac{\sigma'}{2} + \ln \frac{F_g}{\bar{F}_i}}{\sigma' \sqrt{2}} \right) \right)}{\phi} \right] - 1 \qquad (21)$$

where $\varepsilon_i$ is the enrichment ratio of an individual clone, $\sigma'$ is the standard deviation of the single-clone log-normal fluorescence distribution, $\bar{F}_i$ is the mean fluorescence of an individual mutant, $F_g$ is the gating fluorescence and $\phi$ is the gating percentage (Figure 2.6a**)**. $\sigma'$ can be calculated independently for a clonal population of the starting variant using the fluorescence distribution.

To determine optimal sorting parameters, we have plotted enrichment ratios as a function of the ratio of individual fluorescence to the gating threshold for fluorescence $\left(\frac{\bar{F}_i}{\bar{F}_{wt}}\right)$at different gating percentages (Figure 2.6c). It should be noted that the gating threshold is dependent on the fraction of cells that will be collected. A less-stringent gate, with $\phi$ equal to 10%, provides a distribution of enrichment values for many of the clonal populations but will not resolve differences in binding above $\frac{\bar{F}_i}{F_g}$>1.3. A stringent gate at $\phi$=1% enriches strong binders to a ratio of about 6, providing little information about poor binders. We find the optimal $\phi$ to be around 5%, where the enrichment

ratios of both poor binders and strong binders (relative to the original binding interaction) fall within the dynamic range of 4 to -4 (Figure 2.6c). FACS selections should be designed such that the ratio of the fluorescence for the starting construct relative to the anticipated gating threshold is less than 0.5. The actual gating threshold $F_g$, however, is governed by $\phi$, the percentage of the cells that will be collected. Another parameter that can be modified is the log-transformed standard deviation of the fluorescent distribution. For example, this standard deviation can be decreased by drawing a diagonal gate so that the populations are sorted by two fluorescent parameters, which compensates for certain sources of noise.

Figure 2.6d shows the enrichment ratio of clones as a function of the ratio of individual fluorescence to the gating fluorescence at a single gating fluorescence for different standard deviations. Populations with a smaller standard deviation show a smaller range for collection than those with a larger standard deviation. It is recommended that for applications where elucidation of gain-of-function and loss-of-function variants is desired, a square gate should be used. However, a diagonal gate should be used for enriching the population to uncover mostly improved variants.

Finally, in the specific case of yeast surface display of protein binders we label the displayed proteins at levels approximately half of the dissociation constant for the starting protein-ligand interaction. Optimal labeling concentrations can be calculated using parameters set by

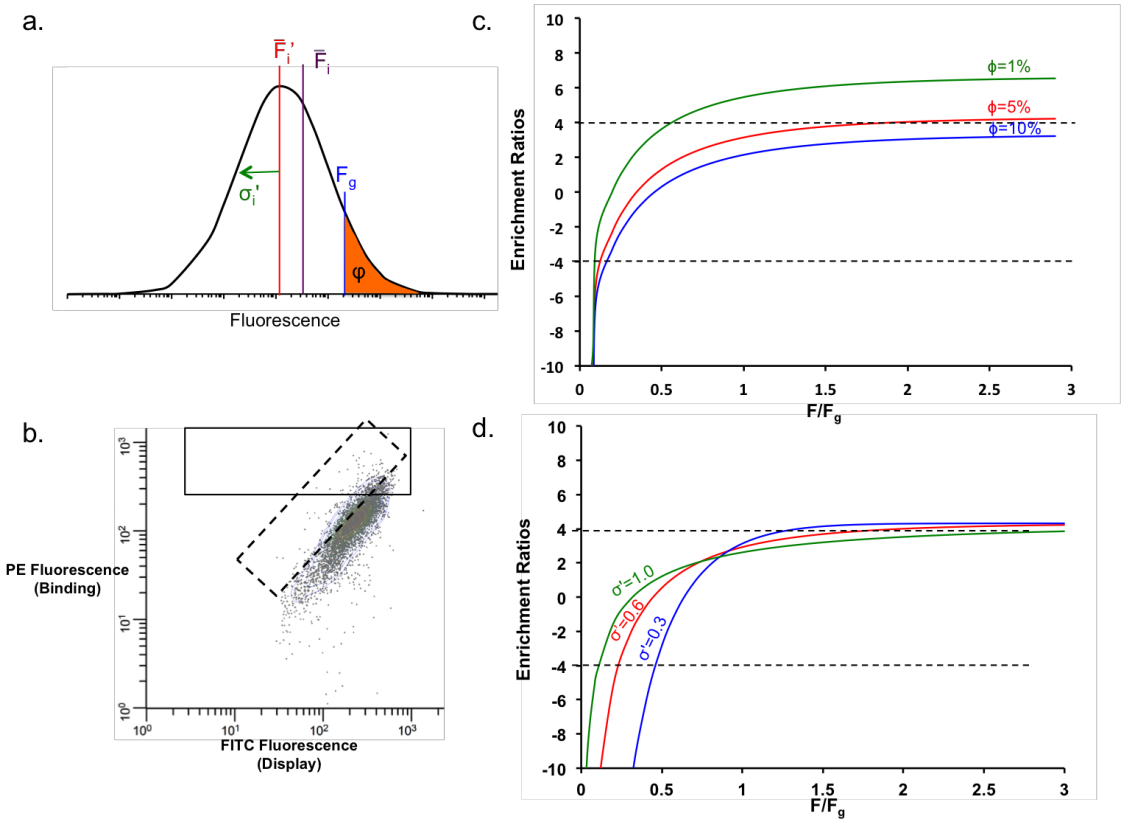



**Figure 7- FACS Selection Parameters.**

a. Individual fluorescence from a clonal population of cells is log-normally distributed with a log-transformed standard deviation σ′, log-transformed mean fluorescence $\overline{F}_i'$ and mean fluorescence $\overline{F}_i$. Cells are collected based on the gating fluorescence, $F_g$, which controls the fraction of cells collected, φ. b. Sample FACS readout for yeast-surface display. The x-axis represents the fluorescence of the displayed population, whereas the y-axis represents the fluorescence of the binding activity of interest. Both square (solid line) and diagonal (dashed line) gates can be drawn around the population to be sorted. Diagonal gates will decrease the standard deviation of the transformed fluorescence distribution, narrowing the range of protein activities that can be resolved. c. The enrichment values as a function of the ratio of the individual fluorescence to the gating fluorescence for different gating percentages (1%, green; 5%, red; 10%, blue) for σ′=0.6. In more stringent sorts, resolution is lost in the enrichment ratios for poor binders. The dynamic range for the fraction of cells collected is between 5 and 10%. d. The enrichment values as a function of the ratio of the individual fluorescence to the gating fluorescence for different standard deviations (1.0, green; 0.6, red; 0.3, blue) for a gating percentage of 5%. Setting smaller standard deviations by two-color sorting using diagonal gates narrows the dynamic range of enrichment values for the sorted populations.

Boder and Wittrup [50]. Higher activity variants are often isolated using multiple sorts from yeast display or other display-based methods. However, our normalization equations only allow

quantitative comparisons between populations occurring during a single sort. Theoretically using one sort is sufficient to resolve most of the population while minimizing time and down stream processing for FACS. As necessary, further sorting can be done to finely discriminate among the enhanced binding variants (Figure 2.6c**)**. In the specific case of yeast surface display, the labeling concentration for the second sort can be set at a much lower level than the first sort. Analysis of the population frequencies after the second sort compared to the first sort can be done using the same normalization equations as above.

### 2.5.4 Deep Sequencing Library Preparation

Deep sequencing was used to obtain count data of each variant in the population using an Illumina MiSeq in 150-bp paired end mode. Plasmid DNA was extracted using a Qiagen miniprep kit (for *E. coli*) or a modified smash and grab protocol (for *S. cerevisiae*)[48]. Following plasmid extraction, a modular two-step PCR method was used to amplify the gene tile and to add the Illumina sequencing, adaptor, and barcode sequences (Figure 2.7). The two-step PCR procedure involves two sets of primers. The first, inner, set amplifies out the gene tile using the gene sequence up- and downstream of the tile and attaches a segment of the sequencing primer. Inner primers are specific to each tile and can be designed using a custom script. The outer primers attach the Illumina adaptors and a barcode on the 3' end of the gene. These primers are a universal set and can be used across different experiments.

Three different PCR methods (Methods A, B, and C, Table 1) were used to attach both sets of primers to an unselected library of LGK variants, and frequencies of each variant were quantified by deep sequencing. If there were no differences in PCR bias among methods, the error in calculating the normalized amount (frequency) of each variant in the population would approach the Poisson limit. Comparisons of variant frequency between Method A and Method B show error

between methods approaching this theoretical minimum (Figure 9a). By contrast, Method C shows much larger differences with respect to Method A (Figure 9b**)** indicating a bias in the PCR method. Figure 10 shows the protein mutation distribution compared to the

theoretical coverage for this unselected library following preparation for sequencing by each of the different methods. Amino acid mutations are enriched in proline (CCN), alanine (GCN), histidine (CAT, CAC) and arginine (CGN) most likely because of the overrepresentation of G bases in the NNN codons in the primer set as discussed above. While proline is grossly overrepresented in the library as a consequence (20% reads vs. 5% from theoretical expectation),

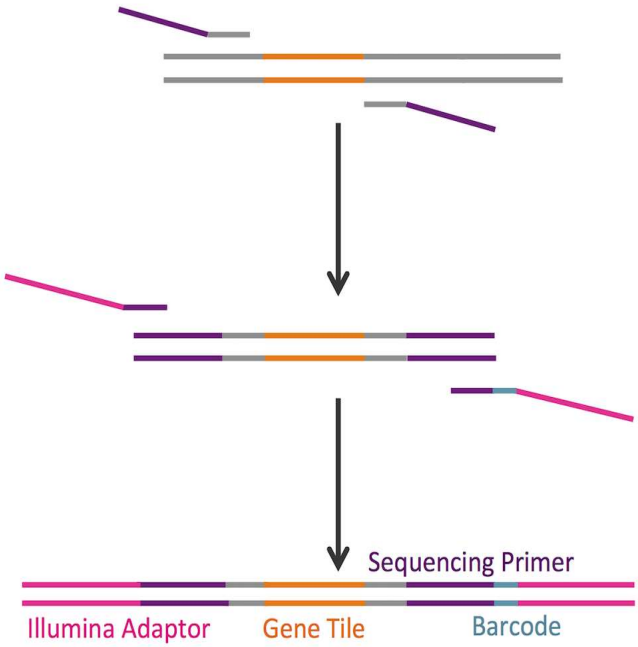



**Figure 8- 2-step PCR method for deep sequencing preparation of libraries.**

PCR reactions are shown for two separate gene tiles containing single mutations (orange and green). Primers are designed to be complementary to flanking regions (grey) of each tile, with encoded single mutations. The first set of primers includes the flanking regions and Illumina sequencing primers (purple). In the next step, outer primers add the Illumina adaptor (pink) and multiplexing index (teal) sequences to the gene. The PCR reaction is performed in a single tube using a 1:2 molar ratio of inner to outer primers and bead purified to remove primer dimer products. The purified library is ready for sequencing without further modifications. While the first set of primers is specific to a single gene tile, the outer primer set is universal.

this bias is tolerable because of oversampling of population members in sequencing. From these results we recommend Method A as it requires the least amount of hands-on and setup time.

### 2.5.5 Normalization and Data Analysis

The frequency of individual variants in selected and unselected populations is extracted from raw sequencing files using the Enrich software suite [27]. Briefly, the forward and reverse reads are aligned, errors between reads are resolved, and the combined sequence is aligned to the starting DNA sequence. Each mutation is recorded and counted, these counts are normalized to frequencies, and the enrichment ratios are found by comparison of the frequency of a given mutant in the selected to the unselected population. To facilitate comparisons of variants across different selection conditions, we have derived normalization equations that transform these enrichment ratios ($\varepsilon_i$) to an objective fitness metric. If a variant is not present in the unselected library then we are unable to determine the fitness metric for that variant.

For growth-based selections, this fitness metric is defined as:

$$\zeta_i = \log_2 \left( \frac{\frac{\varepsilon_i}{g_p} + 1}{\frac{\varepsilon_{wt}}{g} + 1} \right) \tag{11}$$

This metric requires two additional pieces of information. First, the enrichment ratio of the starting or reference sequence must be known ($\varepsilon_{wt}$). Fortunately, this reference variant is generally present in the library, regenerated at each position by the appropriate NNN primer. Second, the number of doubling periods ($g_p$) for the culture must be calculated from the initial and final optical cell density.

To determine whether this relation could reproduce the fitness of individual variants in different populations, we grew populations of *E. coli* harboring plasmids expressing different

levels of eGFP expression [23]. Differential expression results in growth differences among individual strains of nearly 2-fold (n=11; range $0.46 \leq \mu_i \leq 0.76$ $h^{-1}$). Initially, we mixed these variants into a single population and determined individual fitness values after 8.6 average population doublings. Then, we mixed subsets of these variants into two different populations and again determined individual fitness values. Ideally, these fitness values would be exactly the same across the different populations. A best-fit regression line of the fitness values for individuals compared across the different populations gives a slope of 1.04 ($R^2$=0.96), very close to the ideal case of 1 (Figure 11). Based on these results, we conclude that the derived relation is an effective way to normalize the fitness of individual mutants across different populations, thus allowing quantitative comparisons across different selected populations.

In the case of FACS, fitness can be measured across populations according to the following relation:

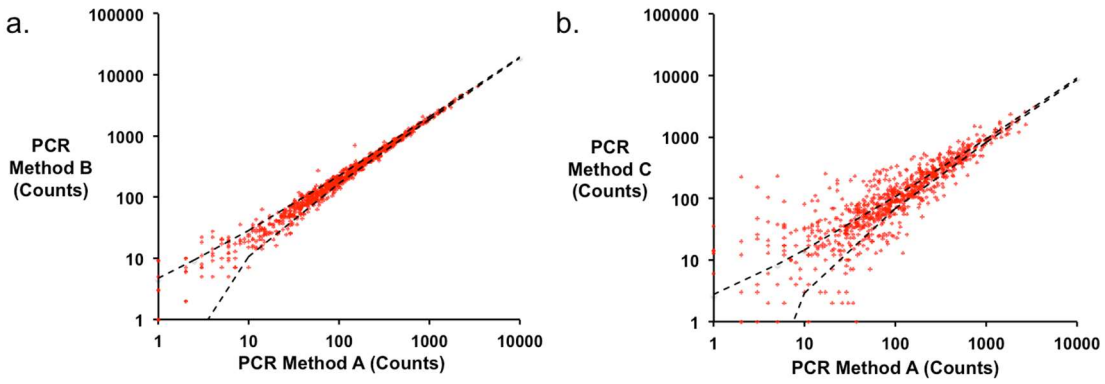



**Figure 9- Errors introduced by different PCR methods.**

Identical mutant libraries were prepared for sequencing using three different PCR methods. The number of counts for each library member was compared across the different methods. Each point represents a specific mutant sequence. For each panel, dashed black lines represent the 95% confidence interval for the Poisson noise between different methods. a. Method B v. Method A. Above 10 counts, the data fall almost completely within minimal error predicted by Poisson noise. b. By contrast, Method C shows significant variance in counts relative to Method A.

$$\zeta_i = \log_2\left[e\sqrt{2}\sigma'\left(erf^{-1}\left(1-\phi 2^{\varepsilon_{wt}+1}\right)-erf^{-1}\left(1-\phi 2^{\varepsilon_i+1}\right)\right)\right] \quad (20)$$

Thus, the fitness of an individual variant $\zeta_i$ can be derived from its enrichment ratio ($\varepsilon_i$) given the enrichment ratio of the wild type ($\varepsilon_{wt}$), the percentage of the entire population collected under the sorting gates ($\phi$), and a $\log_2$-transformed standard deviation of fluorescence for a variant ($\sigma'$) . Since this standard deviation is a scalar in the function, rank ordering of fitness across different populations can be done without directly measuring this quantity. By contrast, the gating percentage of the library is easily measured. Thus, experimentally measured parameters, combined with this relation, allow unambiguous comparisons of variants across different populations.

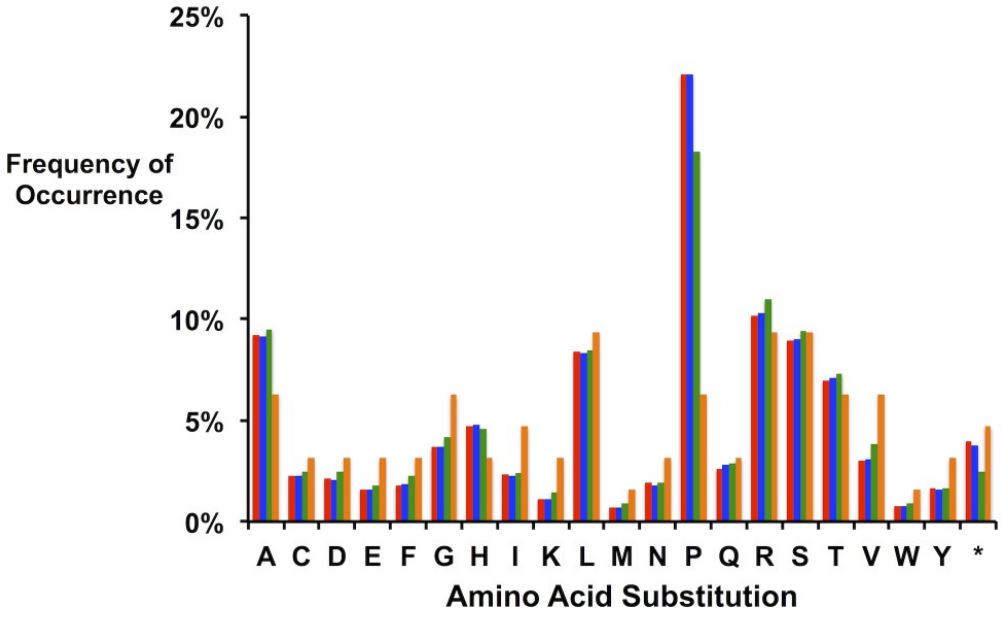



**Figure 10- Library amino acid distribution and PCR bias determination.**

The distribution of incorporated amino acids was used to determine any bias introduced by PCR methods to prepare the library for deep sequencing. The frequency of each mutation overall in the 40 residue region was compared to the theoretical frequency (orange) for each residue type. Method A (red) and Method B (blue) show little difference between the distribution of amino acid substitutions, while Method C (green) shows slight differences at some residue types. The artificial enrichment in proline, alanine and histidine occurs because degenerate primers used for mutagenesis contained an overabundance of guanine bases. The different PCR methods do not show a specific bias toward any single residue.

## 2.6 Conclusions

In this paper we have presented a standardized method for producing the sequence function determinants for entire protein sequences. Furthermore, we have derived equations that allow users to identify optimal selection conditions for their target of interest and to directly compare variants across different populations. Using this method, users can create functional landscapes for full-length genes quickly and efficiently. These landscapes can be applied to protein engineering, for antibody-epitope mapping, and for many different end uses. Additionally, these landscapes can be integrated with computational design methods, either by highlighting existing shortcomings of computational prediction software or as experimental data to guide computational trajectories in search algorithms [51].

The best practices and a step-by-step protocol governing each step in the process are listed in Supplementary Note S2 (Appendix A). These guidelines add to the body of literature for recent sequence-function mapping protocols [13-15, 52, 53]. Notably, many of the individual steps presented here are fully compatible with, and can enhance, these other published protocols. For example, the general fitness equations derived for growth-based selections can be used to optimize experimental set-up for the EMPIRIC approach [15]. Additionally, the general gene tiling and primer design strategy can be applied for assessing full-length sequences with EMPIRIC.

Because of the gene tiling approach, there is no practical upper limit on a gene sequence to be tested. In principle, this approach can be applied to targets much larger than single gene products like complete metabolic pathways. One downside of current approaches is that short read lengths inherent in existing sequencing platforms limit libraries to single mutants or coupled mutants that are proximal in a contiguous stretch of the gene. Resolving this limitation requires new sequencing methods able to resolve long reads with very low error rates. In the near future, perhaps

sequencing-function mapping of multiple simultaneous mutations can be used as a way to fine tune cooperation effects between different beneficial mutations or neutral mutations identified from a single-site saturation mutagenesis library.

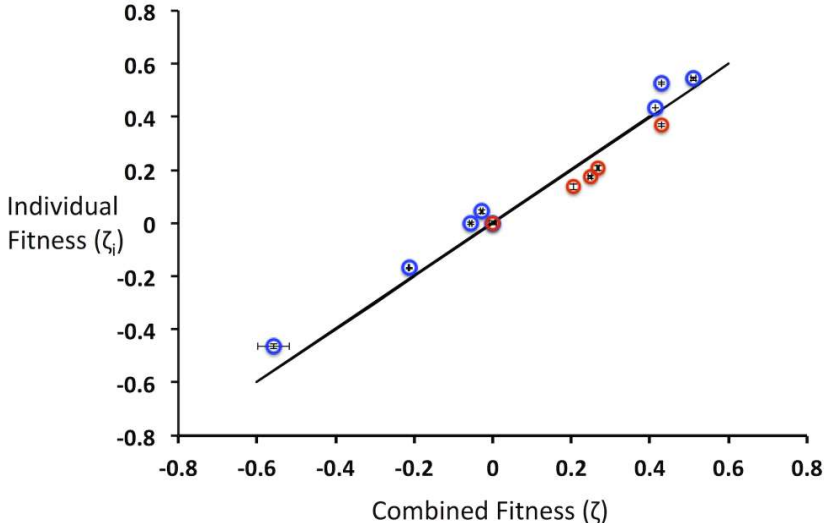



**Figure 11- Experimental validation of growth rate normalization relation.**

*E. coli* harboring 11 different plasmids driving differential eGFP levels were grown in a single population or in two separate populations. The fitness from the separate populations (represented by blue and red open circles) and combined populations were evaluated and compared. Error bars represent one standard deviation from two independent experiments. The solid black line represents the theoretically ideal relationship between individual and combined fitness.

# CHAPTER 3

## 3. Determination of binding affinity upon mutation for type I dockerin-cohesin complexes from *Clostridium thermocellum* and *Clostridium cellulolyticum* using deep sequencing


### Abstract

The comprehensive sequence determinants of cohesin to binding affinity towards type I dockerin from *Clostridium thermocellum* and *Clostridium cellulolyticum* was evaluated using deep mutational scanning coupled to yeast surface display. We measured the relative binding affinity to dockerin for 2,970 and 2,778 single point mutants of *C.thermocellum* and *C. cellulolyticum*, respectively, representing over 96% of all possible single point mutants. The interface ΔΔG for each variant was reconstructed from sequencing counts and compared with three independent experimental methods. The computational software packages FoldX and Rosetta were used to predict mutations that disrupt binding by more than 0.4 kcal/mol. The area under the curve of receiver operator curves was 0.82 for FoldX and 0.77 for Rosetta, showing reasonable agreements between predictions and experimental results. Destabilizing mutations to core and rim positions were predicted with higher accuracy than support positions. This dataset may be useful for protein engineering of orthogonal type I dockerin/cohesin interactions for designer cellulosomes. This benchmark dataset may also be useful for developing new computational prediction tools for the prediction of the mutational effect on binding affinities for protein-protein interactions.


### 3.1 Introduction

Protein-protein interactions are fundamental to biological life, including driving adaptation in the immune system and intercellular communications that regulate organismal development decisions. Given this outsized importance, a grand challenge in protein science is the quantitative prediction of affinity and specificity in protein-protein interactions. Numerous computational

prediction methods have been developed [78-83], but in many cases give poor predictions [84]. One major contributor to improved prediction methods is the presence of robust, strong experimental benchmark datasets. Several such datasets have been compiled over the past fifteen years [82, 85, 86]. These benchmarks are manually curated and contain mutants that have been validated through *in vitro* binding measurements. However, the majority of the mutations in these benchmark sets are of mutations to alanine. As such, the effect of all possible amino acid substitutions on binding affinity at a given position is under-represented.

One major reason that the above benchmark sets are sparsely populated is that they rely predominantly on *in vitro* binding measurements, which can be laborious. Orders of magnitude more data can be generated by experiments measuring the relative binding affinities of entire populations using techniques such as phage display or yeast surface display [35, 87]. In recent years, the ability to deep sequence an entire population has enabled such experiments to be performed routinely [4, 12, 13, 88, 89]. The principle behind these deep mutational scanning experiments is simple: a protein library of mutational variants is passed through a selection or screen for binding affinity. The entire population is sequenced after the selection and compared with a reference population. The change of frequency of a given variant in the population can be calculated directly from deep sequencing counts, and this "enrichment ratio" gives a measure of fitness of that variant under the given selection condition. In previous deep mutational scanning experiments, it was not apparent that the enrichment ratios could be directly related to binding affinities [4, 84]. Recently we have developed selection criteria and normalization relations in a yeast surface display experimental pipeline to unambiguously reconstruct the mean fluorescence of a clonal variant directly from sequencing counts [48]. This approach was recently validated on several different protein-protein interactions [88]. These advances allow for generation of deep mutational scanning benchmark sets

where the relative change in binding affinity can be evaluated for each point mutant in a given protein sequence.

Some anaerobic bacteria degrade cellulosic biomass using extremely large enzyme complexes called cellulosomes [90-92]. The core of the cellulosome is formed from a protein scaffold that non-covalently links several enzymes at specific spatial locations. The scaffold is thought to be functionally important for several reasons: 1.) carbohydrate binding modules arrayed alongside the scaffold disrupt the intra-chain cellulose hydrogen binding, increasing the concentration of enzymatically labile β1-4 glycosidic bonds; 2.) the scaffold links the cellulose to the organism enhancing local concentration effects; and 3.) synergy between spatially constrained enzymes increase overall cellulase activity. The organization of the enzymes onto the cellulosome scaffold is composed of a modular protein-protein interaction recognition sequences. The first characterized interacting partners were type I dockerin/cohesin complexes, with a dockerin domain of approximately 60 residues genetically fused to the enzymes and cohesin domains of 140 residues are dispersed throughout the scaffold. This modular framework results in many different potential enzymatic combinations, the placement of which has not necessarily been optimized for cellulase activity. Indeed, other groups have reconstituted designer cellulosomes comprising orthogonal dockerin/cohesin pairs isolated from diverse microorganisms, and have shown enhanced activity *in vitro* and *in vivo* compared to controls [92-96]. Orthogonal type I dockerin/cohesin interactions have also found utility in disparate metabolic engineering strategies involving enzyme compartmentalization [95, 96].

Type I dockerin/cohesin complexes are a useful model system to evaluate sequence determinants to specificity and affinity for several reasons. Structures of the complexes have been solved from *Clostridium thermocellum* (Ct) and *Clostridium cellulolyticum* (Cc) [97, 98]. The

interactions for both involve a classical protein-protein interface involving buried hydrophobic surface area with a ring of polar and charged residues [99]. The main interaction surface area is comprised of rigid secondary structural elements with a flat beta sheet surface on cohesin and two alpha helices on the dockerin side. One unusual feature of the interaction is that dockerin is able to bind the cohesin in two distinct binding modes [97, 100]. These dual binding modes are thought to be biologically relevant [101]. While several different studies have evaluated sequence determinants to affinity and specificity for Type I dockerin-cohesins [101-107], most have evaluated a handful of mutations to residues thought to be most important for binding affinity. Knowledge of the binding activity for every single amino acid mutation for the complex would improve knowledge of binding affinity and improve computational predictions.

In this contribution, we use deep mutational scanning to evaluate the effect of binding affinity to dockerin for nearly all-possible single point mutants on the type I cohesin domain for both *Clostridium thermocellum* and *Clostridium cellulolyticum* species, giving a comprehensive picture of one side of the affinity landscape for these protein-protein complexes. We evaluate the ability of common computational prediction software packages Rosetta and FoldX to discriminate mutations that disrupt affinity from neutral or gain of affinity mutations. This sample set is, to our knowledge, the largest deep mutational scanning benchmark set with explicit calculations of affinity. As such, these datasets will find utility in developing the next generation of computational prediction software. This manuscript also supplies a blueprint for designing completely orthogonal dockerin and cohesin proteins for use in designer cellulosomes.

## 3.2 Materials and Methods

### 3.2.1 Reagents

All reagents were purchased from Sigma Aldrich (St. Louis, MO) except where noted.

### 3.2.2 Constructs

#### 3.2.2.1 Strains

*E. coli* strains used in this study: **XL1-Blue** (Agilent, Santa Clara, CA) recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F' proAB lacI1$^{q}$ZΔM15 Tn10 (Tet$^{r}$)]; **BL21*** (New England BioLabs, Ipswich, MA) fhuA2 [lon] ompT gal [dcm] ΔhsdS. Yeast strain used in this study: **EBY100** (American Type Culture Collection, Manassas, VA) MATa AGA1::GAL1-AGA1::URA3 ura3-52 trp1 leu2-delta200 his3-delta200 pep4::HIS3 prb11.6R can1 GAL.

#### 3.2.2.2 Plasmids

The plasmid pETCON_ctCohesin was created by inserting a codon optimized gene encoding *Clostridium thermocellum (Ct)* cohesin [98] (GenScript, Piscataway, NJ) using NdeI/XhoI restriction sites. The plasmid pETCON_ccCohesin was created by inserting a gBlock encoding *Clostridium cellulolyticum (Cc)* cohesin [100] (IDT, Coralville, IA) using NdeI/XhoI restriction sites. The plasmid pMAL-C5G_ctDockerin was created by inserting a gene encoding *Ct* Dockerin [98] (GenScript, Piscataway, NJ) with His(x6) as the C-terminal tag using NdeI/BamHI restriction sites. The plasmid pMAL-C5G_ccDockerin was created by inserting a gBlock encoding *Cc* Dockerin [100] (IDT, Coralville, IA) with His(x6) as the C-terminal tag using NdeI/BamHI restriction sites. All genes were codon optimized for *S. cerevisiae* or *E. coli*. All plasmids are deposited in Addgene (www.addgene.org).

### 3.2.3 Protein Expression

*Ct*- and *Cc*-dockerin protein domains were expressed as C-terminal fusions to maltose binding protein (MBP). pMAL–C5G_ctDockerin and pMAL–C5G_ccDockerin were transformed into *E. coli* BL21* (DE3) and protein produced by Studier auto-induction [108] by incubation at 37°C for 6 hours and then at 18°C overnight. Protein was purified according to Bienick et al. [49] except

using a different loading buffer for $Ni^{2+}$-NTA chromatography (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 10 mM $CaCl_2$ and 15 mM imidazole). Proteins were desalted using gravity flow PD-10 desalting columns (GE Healthcare, Little Chalfont, Buckinghamshire, UK) into 3.5 mL Dulbecco's phosphate buffered saline (+$CaCl_2$, +$MgCl_2$) (DPBS) (Life Technologies, Grand Island, NY), pH 7.5. The protein purity was at least 95% as determined by SDS-PAGE.

*Ct* and *Cc* MBP-dockerin were biotinylated with the EZ-link Sulfo-NHS_Biotin bintinylation kit (Thermo Scientific, Waltham, MA) using a protein to biotin molar ratio of 1:20. MBP-*Ct*Dockerin protein was frozen in aliquots and stored at -80°C and thawed on ice when used. MBP-*Cc*Dockerin protein was expressed, purified, and biotinylated fresh when used in experiments.

### 3.2.4 Yeast Clonal Titrations

Yeast clonal titrations were done according to Chao et al. [63] to determine the binding affinity of individual mutants. Point mutants were made by the method of Kunkel [56].

### 3.2.5 Library Preparation

*Ct*- and *Cc*-cohesin genes were segmented into 4 (150 bp paired end reads) or 2 (250 bp paired end reads) tiles, respectively, according to Kowalsky et. al. [48]. Single-site saturation mutagenesis primers containing an NNN degenerate codon were designed using the online QuikChange Primer Design module (http://www.genomics.agilent.com/primerDesignProgram.jsp). For each tile Pfunkel mutagenesis [51] was performed to create a site saturation mutagenesis library. After plasmid midiprep, EBY100 cells were transformed with plasmid library according to Benatuil et.al. [64]. The transformation efficiency ranged from 2.0-6.7x$10^5$ cfu/transformation. Following transformation, cells were harvested by centrifugation, resuspended in 250 mL of SDCAA media, and grown overnight at

30°C and 225 rpm. Cells were re-inoculated to an $OD_{600}$ =1.0 in 250 mL and allowed to grow overnight at 30°C and 225rpm. $1x10^7$ cells were stored at -80°C in 1 mL of yeast storage buffer (20 mM HEPES 150 mM NaCl pH 7.5, 20% (w/v) glycerol) until sorting.

### 3.2.6 Yeast Display Selections

$1x10^7$ cells were grown in 2 mL SDCAA for 6 hours at 30°C and re-inoculated at $OD_{600}$ =1.0 in SGCAA at 22°C for 18 hours. $3x10^7$ cells were labeled with a biotinylated binding protein for 30 minutes at room temperature in DPBSF (DPBS with 1 g/L fraction V BSA). Experiments were performed to confirm that 30 minutes is sufficient labeling time for equilibrium binding. The binding concentration was set to half of the experimentally determined dissociation constant on the yeast surface. Displayed proteins were then labeled with anti-cmyc-FITC (Miltenyi Biotec, San Diego, CA) and streptavidin-phycoerythrin (Thermo Fisher, Waltham, MA). Sorting was done on a Sony Biotechnologies SY3200 or a BD Influx Cell Sorter flow cytometer. Three populations were collected: cells passing through the cell sorter, cells which displayed the protein on the surface and the top 5% of the cells in the fluorescence channel associated with binding. 400,000-500,000 cells were collected of each population (~200x sampling of the theoretical diversity of the library, Table 4). Following the first day of sorting cells recovered populations were grown for 48 hours in SDCAA (pH4.5). $1x10^7$ cells were stored in 1mL of yeast storage buffer at -80°C for each sorted population.

### 3.2.7 Deep Sequencing

Yeast plasmid DNA was prepared for deep sequencing following the protocol in Kowalsky et. al. [48]. Library DNA was sequenced using the 150x2 and 250x2 Illumina MiSeq kits (Illumina, San Diego, CA) at the Michigan State University Sequencing Core.

**Table 4- *Ct*Cohesin and *Cc*Cohesin Sorting Statistics**

| | Tile Length (AA) | Minimum Transformants for 99.9% Coverage | Sort Labeling Conditions (nM) | Events Collected for Binding Population | Percent Sorted (Display) | Percent Sorted (Binding) |
|---|---|---|---|---|---|---|
| ***Ct* Cohesin Tile 1** | 40 | 17,920 | 0.5 | 500,000 | 57.0% | 5.5% |
| ***Ct* Cohesin Tile 2** | 40 | 17,920 | 0.5 | 500,000 | 58.9% | 5.4% |
| ***Ct* Cohesin Tile 3** | 41 | 18,368 | 0.5 | 500,000 | 55.0% | 5.5% |
| ***Ct* Cohesin Tile 4** | 41 | 18,368 | 0.5 | 500,000 | 43.3% | 7.7 % |
| ***Cc* Cohesin Tile 1** | 76 | 34,048 | 0.5 | 500,000 | 49.8% | 4.4% |
| ***Ct* Cohesin Tile 2** | 76 | 34,048 | 0.5 | 400,000 | 61.3% | 5.9% |

### 3.2.8 Data Analysis

A modified version of Enrich-0.2 as described in Kowalsky et al. [48, 53, 88] was used to compute enrichment ratios of individual mutants from the raw Illumina sequencing files. An

enrichment ratio of variant i ($\varepsilon_i$) is defined as the log 2 transform of the frequency of variant i in the selected population relative to a reference population.

To normalize the data across the multiple tiles we define the fitness metric for variant i $\left(\zeta_i\right)$ as the binary logarithm of the mean fluorescence of variant i ( $\overline{F}_i$) to the mean fluorescence of the wild-type sequence $\left(\overline{F}_{wt}\right)$[48]:

$$\zeta_i = \log_2\left(\frac{\overline{F}_i}{\overline{F}_{wt}}\right) \tag{1}$$

This results in the following equation in terms of experimental observables:

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

Where $\phi$ is the percentage of cells collected, $\sigma'$ is the log-normal standard deviation of a clonal population, and the subscript *wt* denotes the wild-type.

Under the sorting conditions used in the experiment, the fitness values can be converted to the change in binding energy upon mutation ($\Delta\Delta G_i$) using the following relation [88]**:**

$$\Delta\Delta G_i \left(\frac{kcal}{mol}\right) = RTln(-\frac{1}{2} + \frac{3}{2^{\zeta_i+1}}) \tag{3}$$

Where R is the gas constant and T is temperature set to 300 K.

A lower bound for the error on the energetic calculations can be estimated from Poisson noise of the sequencing counts and was calculated according to Kowalsky et al. [88]. Positional Shannon entropy was calculated according to Kowalsky et al. [88]. Custom Python scripts used to calculate the fitness metric, Shannon entropy, and statistics are at Github [user: JKlesmith] (www.github.com). The full deep sequencing datasets are provided at figshare (www.figshare.com).

### 3.2.9 Computational Analysis and Evaluation

Computational predictions were compared with experimental observations by area under the curve of a receiver operator curve according to Sirin et al. [86]. Classification of positions in Cohesin sequence was done according to Levy [109].

#### 3.2.9.1 FoldX modeling.

Structures of CtCoh-CtDock (PDB ID 1OHZ) and CcDock-CcCoh (PDB ID 2VN5) were downloaded from PDB and cleaned using the RepairPDB application within FoldX [81]. ΔΔG of individual mutants was assessed using the PSSM application within FoldX using the default flags.

#### 3.2.9.2 Rosetta modeling

Rosetta 2015.38 was downloaded from rosettacommons.org and used for analysis. Structures of CtCoh-CtDock (PDB ID 1OHZ) and CcDock-CcCoh (PDB ID 2VN5) were downloaded from PDB and cleaned using the cleanpdb.py script in the Rosetta release. Structures were prepared for the ddgmonomer application[110] by performing a pre-minimization as follows:

```
./rosetta_bin_mac_2015.38.58158_bundle/main/source/bin/minimize_with_cst.linuxgccrelease -in:file:l lst -in:file:fullatom -ignore_unrecognized_res -fa_max_dis 9.0 -database ./rosetta_bin_mac_2015.38.58158_bundle/main/database/ -ddg::harmonic_ca_tether 0.5 -score:weights talaris2013 -ddg::constraint_weight 1.0 -ddg::out_pdb_prefix min_cst_0.5 -ddg::sc_min_only false > mincst.log
```

ΔΔG of individual mutants was assessed using the low-resolution ddgmonomer application using the following command line:

```
./rosetta_bin_mac_2015.38.58158_bundle/main/source/bin/ddg_monomer.linuxgccrelease -in:file:s /path/to/min_cst_0.5.[PDB]_0001.pdb -resfile ./path/to/mutations.res -ddg:weight_file soft_rep_design -ddg:minimization_scorefunction talaris2013 -database ./rosetta_bin_mac_2015.38.58158_bundle/main/database -fa_max_dis 9.0 -ddg::iterations 50 -ddg::dump_pdbs false -ignore_unrecognized_res -ddg::local_opt_only false -ddg::min_cst true -
```

constraints::cst_file /path/to/input.cst -ddg::suppress_checkpointing true -in::file::fullatom -ddg::mean false -ddg::min true -ddg::sc_min_only false -ddg::ramp_repulsive true -unmute core.optimization.LineMinimizer -ddg::output_silent true -override_rsd_type_limit

### 3.3 Results

To map the sequence determinants to binding for the type-I cohesin dockerin complexes from *Clostridium thermocellum (Ct)* and *Clostridium cellulolyticum (Cc)* we followed a deep mutational scanning approach developed by Kowalsky et al. [88]. We used yeast display[63] to express *Ct*Cohesin (*Ct*Coh) and *Cc*Cohesin (*Cc*Coh) on the surface of *S. cerevisiae* (Figure 12b).

Next, a comprehensive saturation mutagenesis library of all possible single non-synonymous mutations was made for the genes encoding cohesin [51]. These libraries were displayed on the surface of yeast [63] and incubated with the dockerin domain at a labeling concentration of half of the observed dissociation constant. The libraries were sorted by fluorescence activated cell sorting (FACS) into three different populations: a reference population collected without applying a sorting gate, a population of all cells surface displaying cohesin variants, and a population

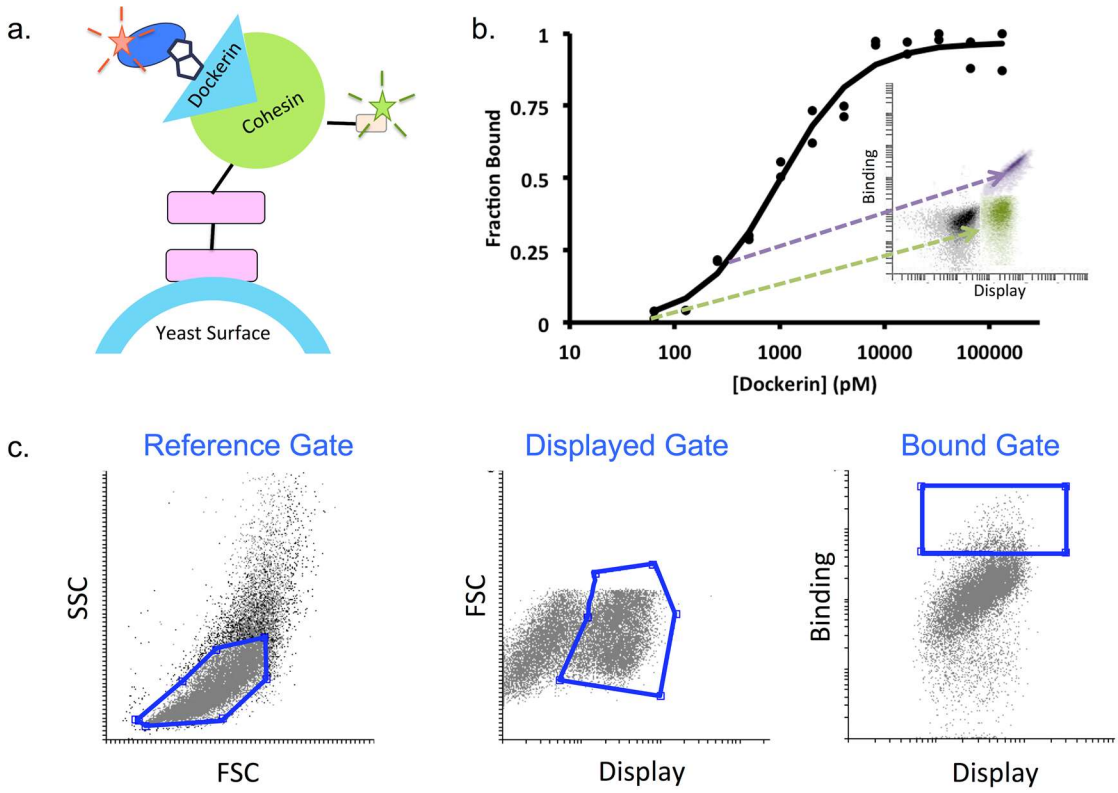



**Figure 12- Method Overview.**

a. Yeast-surface display is used to monitor binding activity. Both *Ct* and *Cc* cohesin domains are independently displayed on the surface of yeast. Dockerin domains are solubly expressed and chemically biotinylated. A C-terminal c-myc epitope tag is used to monitor protein display and a streptavidin-conjugated fluorophore is used to monitor cohesin-dockerin binding. b. Binding titration curve for the Ctcohesin-Ctdockerin interaction as determined by flow cytometry. Inset shows no display or binding activity (black), display only (green) and the binding activity of a clonal population at sorting conditions (purple). c. Populations are discriminated by three sorting gates. The reference gate drawn around forward and side scatter channels (left) captures all yeast cells passing through the flow cytometer. The displayed gate (middle) captures all yeast cells displaying the cohesin domain, while the bound gate captures approximately the top 5% of the binding population (right). Collected populations are grown overnight and the plasmid DNA containing Cohesin variants is extracted, deep sequenced, and analyzed.

containing approximately the top 5% of the binding population (Figure 12c). Sorting statistics are listed in Table 5. Plasmid DNA from the collected populations was then prepared and deep sequenced. We filtered the sequencing counts to include data on a given mutation only if it was counted at least 10 times in the reference population [48].

Using this method we were able to evaluate 3133 out of 3240 (96.7%) and 2824 out of 2900 (97.3%) possible single nonsynonymous mutations, including stop codons, in the protein encoding sequence for *Ct* and *Cc* cohesin domains, respectively. Statistics on the quality of the mutational libraries are given in Table 5**.** We first addressed whether the yeast display system could identify mutations that disrupted the fold of the cohesin proteins. To evaluate this, we compared the frequency of each variant in the population gated by the fluorescence channel associated with binding to a C-terminal c-myc epitope tag ("display" population) compared to the reference population. Consistent with a previous experiment [4], almost every single point mutant in both *Ct* and *Cc* population displayed with near the same frequency as the reference population (Figure 13). These include mutations that are predicted to appreciably destabilize the protein fold ($\Delta\Delta G_{unfolding}$ > 1.0 kcal/mol), including introduction of charged residues in the

protein interior and mutations of small to large interior residues predicted to cause steric clashes. Notably, the sorting gate used on the displaying population was able to discriminate mutants not displaying the epitope tag as premature stop codons that remove the c-myc epitope tag used to identify the displaying populations are significantly depleted (p-value $<10^{-16}$ using a one-tailed paired t-test).

We next compared the frequency of the binding population to the frequency of the reference population. The new reference population was combined from the nearly identical displaying and the initial reference populations in order to lower the intrinsic counting error. We used a Shannon (sequence) entropy metric [88] to plot sequence conservation on the structures of *Ct* and *Cc* cohesin. Consistent with expectations, we find that the conserved residues map to the known binding modes of their respective dockerins (Figure 13).

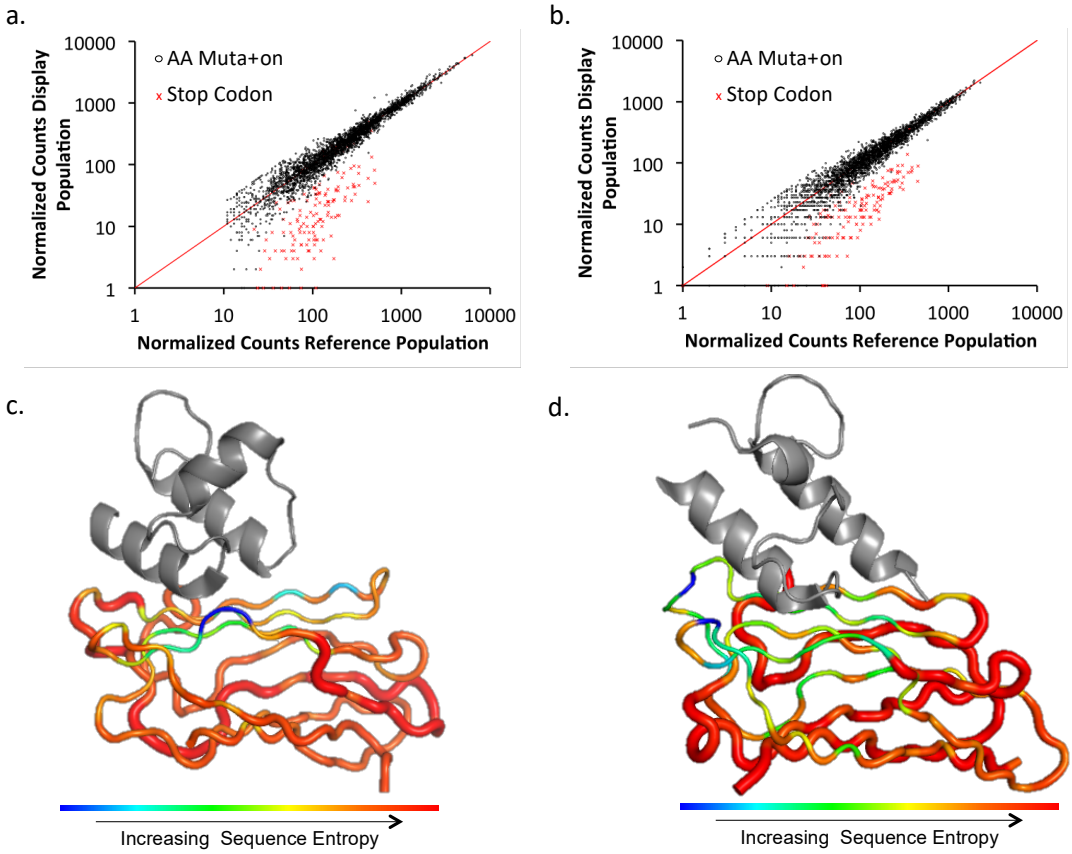



**Figure 13- Deep mutational scanning of CtCoh and CcCoh**

**a.**,**b**. Counts in the display population relative to counts in the reference population for CtCoh (**a.**) and CcCoh (b**.**). Black open circles represent a single point mutation to another amino acid, whereas red crosses indicate mutation to a premature stop codon. The counts were normalized to account for different depth of coverage by deep sequencing. **c.**,**d.** Cartoon showing the solved structures of CtCoh-CtDock (**c**; PDB ID 1OHZ) and CcCoh-CcDock (d; PDB ID 2VN5). The dockerin domain is shown in grey, while the cohesin domain is colored according to positional sequence entropy. The positions of low sequence entropy map to the binding interface.

Next, we converted the enrichment ratios in yeast display sorts to a fitness metric ($\zeta_i$) representing mean fluorescence of a variant relative to mean fluorescence of the wild-type sequence [48]. Variants with higher fluorescence in the binding channel will be selectively enriched in the population. Conversely, variants with lower fluorescence will

**Table 5- Library Statistics**

| | *Ct* Cohesin | | | | *Cc Cohesin* | |
|---|---|---|---|---|---|---|
| | **Tile 1** | **Tile 2** | **Tile 3** | **Tile 4** | **Tile 1** | **Tile 2** |
| **Reads passing through enrich for reference library** | 171600 | 451027 | 193222 | 741965 | 346198 | 736768 |
| **Percent of possible codon substitutions observed** | | | | | | |
| **1-base substitution** | 100.0% | 99.4% | 99.2% | 100.0% | 100.0% | 100.0% |
| **2-base substitutions** | 95.6% | 96.9% | 96.4% | 97.7% | 96.7% | 97.9% |
| **3-base substitutions** | 93.4% | 96.4% | 96.1% | 97.6% | 96.4% | 96.1% |
| **All substitutions** | 95.3% | 97.0% | 96.7% | 97.9% | 97.1% | 97.4% |
| **Percent of reads with** | | | | | | |
| **No nonsynonymous mutations** | 17.8% | 17.1% | 19.0% | 28.1% | 36.4% | 23.9% |
| **One nonsynonymous mutation** | 78.6% | 78.5% | 76.9% | 60.4% | 53.6% | 58.6% |
| **Multiple nonsynonymous mutations** | 3.6% | 4.4% | 4.2% | 11.4% | 9.9% | 17.5% |
| **Coverage of possible single nonsynonymous mutations** | 96.2% | 97.5% | 96.1% | 98.4% | 96.9% | 97.8% |

be depleted. The fluorescence of any given cell depends on the number of molecules of target protein displayed on the surface, the number of fluorophores per capture protein that binds to the target protein, and the fraction of targets bound by the capture molecule. Under conditions of equilibrium binding, which we have validated in the present case, the fraction of bound targets is set by the labeling concentration of the capture molecule and the binding dissociation constant of the interaction.

In the following step we converted the fitness metric to a change in binding energy (interface $\Delta\Delta G_i$) for each variant, *i*, using the following equation:

$$\Delta\,\Delta G_i\left(\frac{kcal}{mol}\right) = RTln(-\frac{1}{2}+\frac{3}{2^{\zeta_i+1}}) \quad (3)$$

Where R is the gas constant and T is temperature set to 300 K. This deep sequencing-derived interface ΔΔG is calculated using the above fitness metric representing the relative fluorescence of a variant at a single labeling concentration (Figure 14). This conversion results in a relatively narrow dynamic range arising from variants either at the maximum or minimum fluorescence value at the labeling conditions used (Figure 14). Given the sorting conditions used in the experiment, there is no discrimination among mutations with interface ΔΔG greater than 0.5 kcal/mol [88]. At the other extreme, this method can also detect stabilizing mutations, as we have designed the experiment such that no one variant takes over the entire selected binding population. This gives an upper bound of interface ΔΔG to -0.8 kcal/mol.

There are several assumptions needed to derive interface ΔΔG directly from sequencing counts, and any major deviations in these assumptions can substantially affect the validity of this reconstruction. These assumptions include that (1.) the maximum fluorescence is much greater than the minimum fluorescence in the binding channel on the cell sorter [88]; (2.) the labeling conditions occur under equilibrium binding conditions; (3.) there is a 1:1 binding stoichiometry between the surface displayed protein and the protein in solution; (4.) The titration curve can be represented with a Hill coefficient equal to 1; (5.) the mean and variance of target molecules displayed on the surface is the same for each variant; and (6.) the number of sequencing counts of a variant accurately reflects its frequency in the population. While assumptions 1-4 can be externally validated or are otherwise reasonable for many protein-protein interactions, the latter two assumptions are known to be simplifications. For example, in certain cases protein variants have been shown to have very different average surface display [111, 112]. Additionally, intrinsic counting error (also known as Poisson noise) decreases with increasing sequencing depth.

Accordingly, one would expect that the error in determining the frequency of a variant sequenced 10 times in the reference population will be greater than that of a variant represented 1000 times.

In order to test the above assumptions, we reasoned that assessing the interface ΔΔG for all synonymous mutations to the wild type sequence would show an average interface ΔΔG of zero, and that the distribution of calculated interface ΔΔG at these positions would give a baseline error in the method. We evaluated the interface ΔΔG for all synonymous mutations on *Ct*Coh (Figure 14). Of these 378 variants, the range of interface ΔΔG is 0.33 to -0.55 kcal/mole, with an average of -0.02 kcal/mol and a standard deviation of 0.10 kcal/mol. The interface ΔΔG is calculated based off sequencing counts, and lower counts will increase the intrinsic counting error. Accordingly, we hypothesized synonymous mutations with higher counts in the reference population would have a narrower distribution than lower counts.

Figure 14 shows the distribution for variants with less than or more than 100 counts in the reference population. As predicted, those with less than 100 counts had a standard deviation of 0.13 kcal/mol and more than 100 counts had a standard deviation of 0.07 kcal/mol. To test the assumption that the mean number of target molecules displayed on the surface is the same for each variant, we assessed interface ΔΔG for surface positions on *Ct* Cohesin far from the interface.

Figure 14 shows the distribution for variants with less than and more than 100 counts in the reference population. Similar to the synonymous mutants, those with less than 100 counts had a standard deviation of 0.16 kcal/mol and more than 100 counts had a standard deviation of 0.08 kcal/mol.

As a final test of the estimation of ΔΔG values from deep sequencing counts, we measured the binding affinities of nine individual variants using yeast clonal titrations (Figure 14**).** The best-

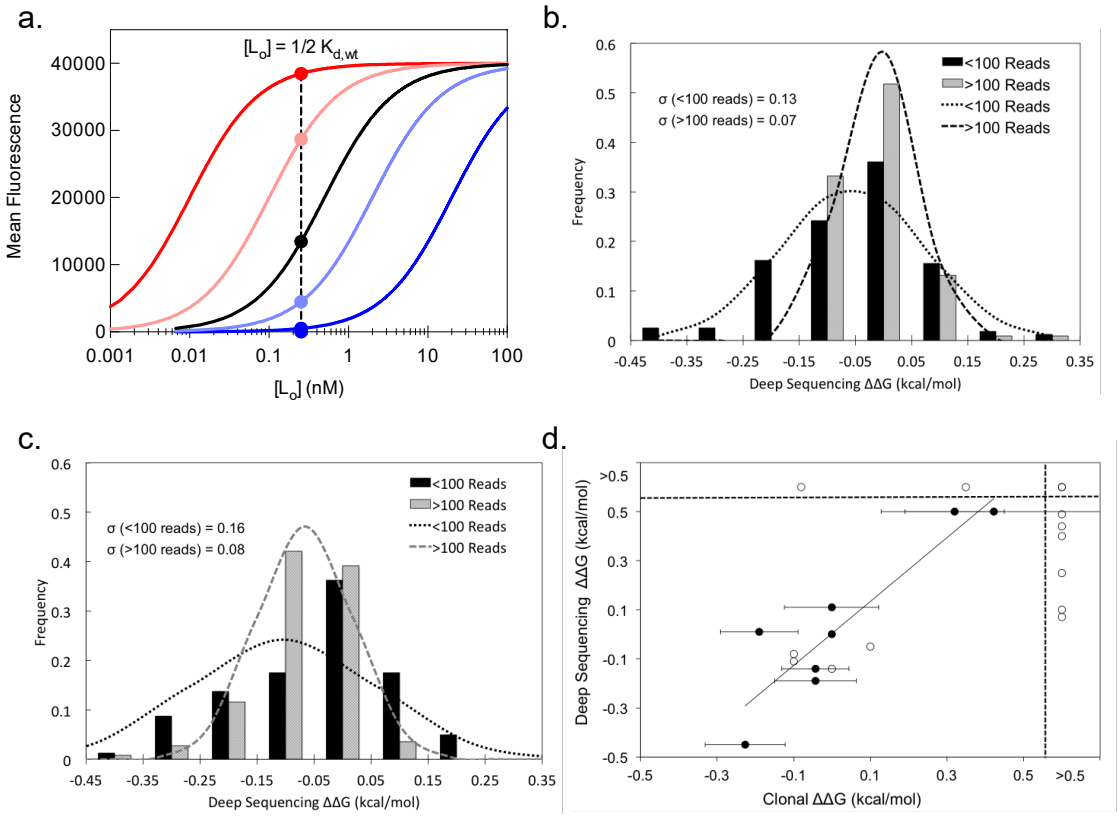



**Figure 14- Experimental estimates of error in interface DDG**

**a.** Theoretical titration curves for wild-type (black) and variant (red-blue) clonal titrations. The deep sequencing method reconstructs for each variant the mean fluorescence at the labeling concentration, represented as a closed circle on the dashed vertical line. Deep sequencing interface ΔΔG values are then reconstructed from these single-point titration curves. **b.** Histogram for deep sequencing-calculated interface ΔΔG for synonymous mutations to wild-type CtCoh. Mutations with less than 100 counts in the reference population are in black and mutations with more than 100 counts in the reference population are shown as diagonal stripes. Dashed lines represent best-fit Gaussian distributions for the different populations. Bin sizes are 0.1 kcal/mol and are centered on the axis markings. **c.** Histogram of calculated interface ΔΔG for surface residues far from the interface. Bin sizes are 0.1 kcal/mol and are centered on the axis markings. The black bars are for all selected surface mutations with less than 100 counts in the unselected population and the striped bars are for greater than 100 counts in the unselected population. **d.** Scatter plot of interface ΔΔG calculated by deep sequencing compared against clonal titrations. Literature values are shown as open circles, while yeast clonal titrations done in this work are represented with filled in circles with error bars shown as 1 s.d. of 3 independent measurements. Errors for literature values are not reported in the original references. The best fit line shown for yeast clonal titrations shows a $R^2$= 0.82. For all measurements, $R^2$= 0.52.

fit line of the relationship between ΔΔG calculated from deep sequencing and clonal titrations has

a correlation coefficient of 0.82 (Figure 14). Restricting the correlation to the identity line (y=x)

still maintains a correlation coefficient of 0.77. Next, we compared our relative dissociation constants with literature values that span a wide range of ΔΔG (Figure 14). As a final test of the estimation of □□G values from deep sequencing counts, we measured the binding affinities of nine individual variants using yeast clonal titrations. The best-fit line of the relationship between □□G calculated from deep sequencing and clonal titrations has a correlation coefficient of 0.82 (Figure 3.14). Restricting the correlation to the identity line (y=x) still maintains a correlation coefficient of 0.77. Next, we compared our relative dissociation constants with literature In this case, the correlation coefficient for the identity line was 0.52 [106, 107, 113]. We next asked which residues were energetically important using sequence entropy as a measure of conservation to the *Ct* Dockerin-Cohesin interface (Figure 15)[88]. Of all CtCoh residues within 8Å of CtDock, Asn37, Asp 39, Val41, Ala72, Tyr74, Val81, Leu83, Glu86, Gly122, Gly123, and Ala125 were found to

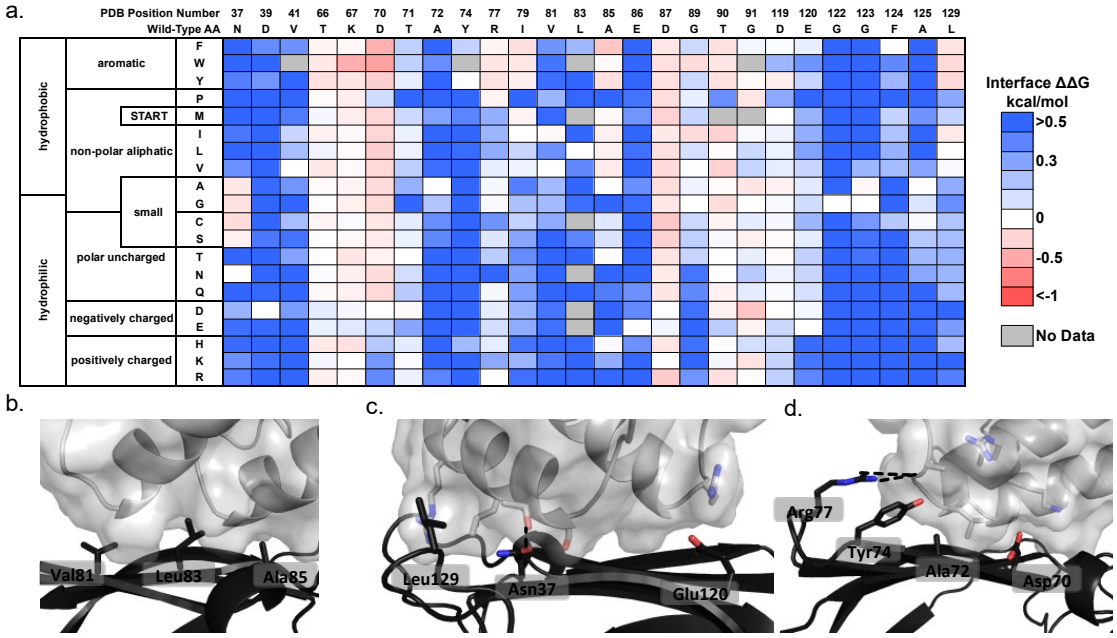



**Figure 15- Interface ΔΔG reconstruction for *Ct* Cohesin – Dockerin interaction.**
a. Interface ΔΔG map for *Ct* Cohesin residues within 8Å of the interface. ΔΔG values for individual point mutations are shown for deleterious (blue), neutral (white) and beneficial (red) substitutions. b.-d. Interface views of selected residues for the interaction between cohesin (dark grey cartoon) and dockerin (light grey cartoon and surface).

have the lowest sequence entropy, and therefore the highest conservation. These residues have been mapped on to the structure of *Ct* cohesin (**Figure 15).** Leu83 is located within a hydrophobic patch at the cohesin-dockerin interface, Asn37 and Asp39 are part of a hydrogen bond network previously found to be essential to the interaction, and Glu86 is part of a conserved salt bridge [107]. Among these conserved positions, only 8/228 mutations have an interface ΔΔG less than 0.1 kcal/mol: Asn37Ala, Asn37Gly, Asn37Cys, Asn37Ser, Val81Ile, Gly123Ala, Ala125Gly, and Ala125Cys (**Figure 15**). All of these are conservative, small-to-small mutations consistent with the positioning of these residues in the center of the interface.

Consideration of all residues on *Ct* cohesin within 8Å of the bound dockerin domain revealed five residues of particular interest: Asp70, Arg77, Ile79, Asp87 and Leu129. For Asp70 all mutations except for those to positively charged Arg and Lys show interface ΔΔG of less than 0.2 kcal/mol (Figure 15). Indeed, substitutions to hydrophobic and aromatic amino acids are tolerated. We hypothesize that Arg and Lys are not tolerated at position 70 because the proximity of Lys18 and Arg19 on *Ct*Dock creates an unfavorable electrostatic interaction. We speculate that larger residue substitutions can be tolerated because the Asp70 Cα-Cβ vector points away from the interface. Arg77 makes hydrogen bonding contacts with Arg23 on *Ct*Dock, yet tolerates most substitutions except for Asn (Figure 15). Ile79 appears well packed in the crystal structure, yet readily tolerates substitutions to hydrophobic and/or aromatic amino acids. Another interface residue that accepts aromatic mutations with improved binding (interface ΔΔG less than -0.15 kcal/mol) is Leu129, which is at the periphery of the interface. Finally, residue Asp87 accepts all amino acid substitutions with essentially no energetic penalty.

Reconstruction of interface ΔΔG was less straightforward for the *Cc*Coh-*Cc*Dock interface.

**Table 6- Interface ΔΔG comparison between clonal and deep sequencing data for *Ct*Coh-*Ct*Dock interface.**

| Mutation | Clonal ΔΔG [kcal/mol] | Deep Sequencing ΔΔG [kcal/mol] | Experiment |
|---|---|---|---|
| **E120R** | 0.4 ± 0.3 | >0.5 | |
| **I79F** | -0.2 ± 0.1 | -0.19 | |
| **D70F** | -0.1 ± 0.1 | -0.45 | |
| **R77E** | -0.1 ± 0.1 | 0.11 | |
| **R77N** | 0.2 ± 0.1 | > 0.5 | |
| **D70H** | -0.2 ± 0.1 | 0.01 | |
| **N37A** | -0.1 ± 0.1 | -0.14 | |
| **D39S** | 0.5 | 0.44 | Miras et. al.[40] |
| **Y74A** | -0.1 | > 0.5 | |
| **E86S** | 0.4 | > 0.5 | |
| **A94L** | 0.5 | 0.07 | |
| **G97S** | 0.1 | -0.05 | |
| **L83S** | 3.4 | 0.40 | Slutzki et al.[34] |
| **N37A** | 0 | -0.14 | |
| **N37D** | 2.5 | 0.25 | |
| **N37L** | >4 | > 0.5 | |
| **D39A** | >4 | 0.49 | |
| **D39N** | >4 | > 0.5 | |
| **E131A** | 1.8 | 0.10 | |
| **V41Y** | 0.7 | > 0.5 | |
| **D70S** | -0.1 | -0.11 | |
| **V81S** | 1.9 | > 0.5 | |
| **A85L** | -0.1 | -0.08 | |

The protein was split into two libraries with one comprising residues 1-76 and the other containing mutations in residues 77-154. The library comprising residues 1-76 yielded results qualitatively

similar to the CtCoh results (Figure 13, Figure 16**)**. However, the library comprising residues 77-154 resulted in very strong gain of function for nearly all mutations at positions Asn81, Gly82, and Thr83 that are peripheral to the interface (Figure 16). These mutations overtook the selected population. We speculate that Asn81-Thr83 contains a yeast N-linked glycosylation site (NXT/S) not present in the native interaction that sterically occludes access of *Cc*Dock to *Cc*Coh (Figure 16). Removal of glycosylation by mutation results in unimpeded access to the *Cc*Coh interface surface, resulting in large increases in binding affinity. Because of this complication we restricted our analysis to the first 76 residues of *Cc*Coh only, although we note that the unmodified glycosylation site does exist and may potentially affect the resulting data for the first 76 residues as well. In the first 76 residues, we found three highly conserved residues as determined by sequence entropy: Thr45, Asn47, and Tyr49 (Figure 16). These residues are found on an internal β-strand at the cohesin-dockerin interface. Only 2/57 mutations at these positions had an interface ΔΔG of less than 0.1 kcal/mol (Thr45Lue and Thr45Val). Two residues of additional interest are Asn74 and Ser76. While not located in direct contact with the interface, any mutation to Asn74 results in a destabilized interaction (ΔΔG >0.2 kcal/mol). This is because the side chain of Asn74 has many contacts with backbone elements of nearby residues. The other residue, Ser76, can accept neutral substitutions to Phe, Tyr, Ala, Gly, Asn, and Asp. Ser76Asp shows an increase of binding affinity possibly because of the formation of a salt bridge interaction with the dockerin domain.

We next evaluated the predictive ability of the computational software programs Rosetta and FoldX to discriminate among the affinity enhancing mutations. There was sufficient data for 686 mutations on CtCoh & CcCoh that met the interface classification of Levy [109]. According to classification of an interface by Levy, a protein interface can be classified into core, rim, or support

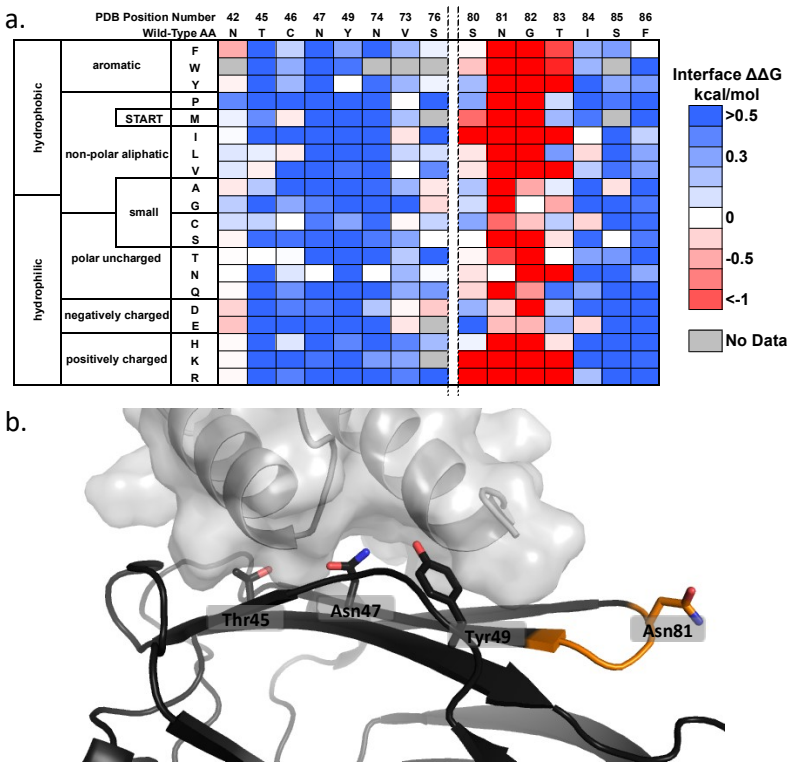



**Figure 16- Interface ΔΔG reconstruction for *Cc* Cohesin – Dockerin interaction.**

a. Interface ΔΔG map for selected interface residues of *Cc* Cohesin. ΔΔG values for individual point mutations are shown for deleterious (blue), neutral (white) and beneficial (red) substitutions. The dashed line demarcates the two separate libraries used for reconstruction of binding affinities. The right panel shows strong gains in binding affinity upon mutation of a predicted N-linked glycosylation site. b. Interface views of selected residues for the interaction between cohesin (dark grey cartoon) and dockerin (light grey cartoon and surface). The predicted site for N-linked glycosylation at Asn81 is shown in orange.

positions. Core positions are exposed in the monomer state but buried in the complexed state, support positions are those buried in both states, while rim positions maintain greater than 25% relative accessible surface area in both states. Of all interface positions, 242 (35.3%) were destabilizing by more than 0.4 kcal/mol. We first assessed whether these methods could discriminate the mutations experimentally determined to increase interface ΔΔG by greater than 0.4 kcal/mol. This was determined using receiver operator curves and evaluating the area under the curve (AUC). Mutations with a ΔΔG greater than 0.4 kcal/mol were treated as positives and all others were negative. FoldX gave an AUC for the entire dataset of 0.82 (Figure 17), whereas the

Rosetta modeling suite afforded an AUC of 0.77 (). The AUC for the complete *Ct*Coh dataset was 0.80 for both FoldX and Rosetta. We next asked whether there was a difference in computational prediction of the interfacial positions according to these Levy interface classifications of core, rim, and support positions. Mutations at core (AUC 0.85, 0.75) and rim (AUC 0.80, 0.81) were discriminated with a higher accuracy than support (AUC 0.76, 0.61) positions for FoldX and Rosetta, respectively.

Mutations at CtCoh residues Asn37, Asp70, Glu120, Asn127, and Leu129 and *Cc*Coh position Cys46 gave particular disagreements between computational predictions and experimental results. While Rosetta predicts that directional hydrogen bonding on *Ct*Asn37 contributes considerably to the binding affinity (computational prediction for Asn37Ala shows a $\Delta\Delta G$ >0.5 kcal/mol), *Ct*Asn37 can be replaced smaller residues Ala/Gly/Cys/Ser with no loss in binding affinity (Figure 15) as previously discovered [107]. *Ct*Asp70 is a rim position that tolerates most substitutions except for positively charged amino acids, whereas FoldX predicts strong conservation at that position. Rim position *Ct*Asn127 is predicted to tolerate most mutations by FoldX yet pays a small energetic penalty for all substitutions except for mutation to Asp/Glu presumably because of electrostatic interactions with *Ct*Dock Arg53. Substitution of *Ct*Leu129 with aromatics Phe/Trp/Tyr results in an increase in binding affinity, and Rosetta predicts a favorable interaction for Leu129Phe only.

Only 5/686 point mutations (0.7%) showed increased affinity greater than 0.4 kcal/mol: *Ct*Lys67Trp, *Ct*Asp70Phe/Trp, *Cc*Pro66Trp, and *Cc*Lys69Cys. Substitutions on *Ct* remove charged residues on the rim of the interface with aromatics that presumably increase van der Waals packing. FoldX was not able to identify any of these mutations. In fact, FoldX predicted that CtAsp70Phe/Trp would be appreciably destabilizing ($\Delta\Delta G$ >0.5). In contrast, while Rosetta

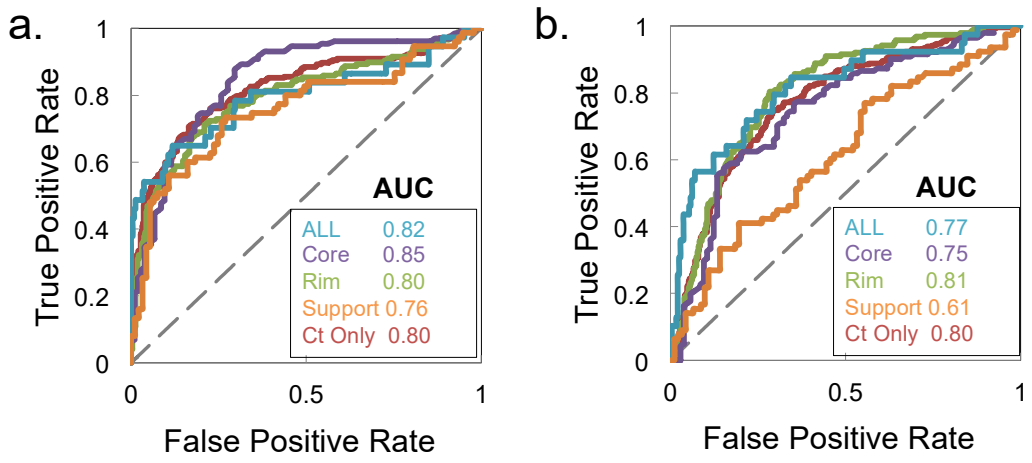



**Figure 17- Computational predictions of experimental interface ΔΔG.**

Receiver operator curves for computational prediction of all interface mutants (cyan), Ct-only mutants (red), core positions (color), rim positions (color), and support positions (color) (a.) FoldX and (b.) Rosetta. The area under the curve (AUC) for each subset is in the inset.

predicted all 3 beneficial mutations on CtCoh, the program also predicted gain of binding affinity for 76 other mutations at that interface.

### 3.4 Discussion

The comprehensive sequence determinants to binding affinity for type I dockerin-cohesin complexes from *Clostridium thermocellum* and *Clostridium cellulolyticum* were evaluated using yeast surface display coupled to deep mutational scanning. Interface ΔΔG could be estimated from sequencing counts. The computational software packages FoldX and Rosetta could predict mutations that disrupt binding by more than 0.4 kcal/mol with reasonable accuracy. Destabilizing mutations to core positions were predicted with higher accuracy than rim or support positions.

As with other systematic mappings of the sequence-function space of proteins, one must keep in mind certain limitations when interpreting results [114]. First, in the present work the effective interface ΔΔG estimated from the deep sequencing data was limited to a narrow dynamic range of approximately 1.3 kcal/mol (approximately a 10-fold range of dissociation constants). Other reports have shown that one can sort yeast populations under different or multiple labeling

concentrations or use multiple sorting gates in order to discriminate moderately destabilizing mutations from true hot spot residues [23, 115, 116], although such experiments significantly increase time and effort. Nevertheless, within this narrow range the deep sequencing estimated interface ΔΔG had a correlation coefficient of 0.77 with yeast clonal titration data and 0.52 for all mutations found in literature. From our analysis of the fitness metrics of synonymous mutation we observe an intrinsic error around 0.1 kcal/mol, with lower intrinsic error found with positions that have higher depth of coverage. In light of these issues, care must be taken to estimate an energetic contribution especially for mutations that are predicted to enhance binding affinity. Enrichment of a variant can result from an increase in fluorescence resulting from higher surface expression, not from its effect on binding affinity. Scanning the entire protein sequence, rather than just the interface residues, allows one to quantify the error associated with this assumption for the given protein of interest.

Second, as shown for the mutational library comprising residues 77-154 on *Cc*Coh, binding sites on the yeast-displayed protein can be partially blocked by N-linked glycosylation. Removing that steric hindrance by mutation can perturb the energetic analysis for the other mutants in the library. Prescreening and removing N-linked glycosylation sites prior to sorting can prevent this issue.

Finally, one critical assumption is that the energetics of binding determined from yeast surface display titrations is representative of binding affinities from the proteins in solution. While there has been some validation for this assumption, yeast surface display measurements yield worse correlations using computational prediction software compared with *in vitro* techniques like surface plasmon resonance [86].

In summary, we have used deep mutational scanning to determine the binding affinities for nearly every single point mutant in the *Ct*Coh and *Cc*Coh domains. The methodology used in this current contribution can be transferred to other protein systems. We anticipate a series of benchmark sets for many different protein-protein interactions in the near future.

# CHAPTER 4

## 4. Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing.


### Abstract

Knowledge of the fine location of neutralizing and non-neutralizing epitopes on human pathogens affords a better understanding of the structural basis of antibody efficacy, which will expedite rational design of vaccines, prophylactics, and therapeutics. However, full utilization of the wealth of information from single cell techniques and antibody repertoire sequencing awaits the development of a high-throughput, inexpensive method to map the conformational epitopes for antibody-antigen interactions. Here we show such an approach that combines comprehensive mutagenesis, cell surface display, and DNA deep sequencing. We develop analytical equations to identify epitope positions, and show the method effectiveness by mapping the fine epitope for different antibodies targeting TNF, Pertussis Toxin, and the cancer target TROP2. In all three cases, the experimentally determined conformational epitope was consistent with previous experimental datasets, confirming the reliability of the experimental pipeline. Once the comprehensive library is generated, fine conformational epitope maps can be prepared at a rate of four per day.


### 4.1 Introduction

Pinpointing the fine conformational epitope targeted by a given antibody affords a better understanding of the structural basis of its mechanism of protection, which provides an intellectual property basis and can lead to improved prophylactic or therapeutic interventions against human diseases[4, 24-31]. Recent technical advances allow an unprecedented look at the adaptive immune response to an immunogen[117-119]**.** For example, single cell isolation methods coupled to deep sequencing have revealed the identification of thousands of patient-specific paired antibody heavy

and light chain sequences elicited in response to infection or vaccination, and such information has begun to be used in antibody discovery. While functional or neutralization assays can be used to determine the efficacy of individual members in these repertoires, a full utilization of this wealth of information awaits the development of a high-throughput method of determining conformational epitopes targeted by these antibodies[120].

Existing methods either do not identify conformational epitopes[121, 122] or are labor-intensive and costly. Co-crystallization provides unambiguous epitope identification but can require considerable effort and generation of many antigen variants in order to identify one that is compatible with crystallization[25]. Mass spectrometry-based methods utilizing hydrogen/deuterium exchange identify epitopes to a ca. 5 amino acid resolution only under rigorous control experiments that limit throughput[32, 33]. Competing display-based methods use many sorts[123], identify only partial epitopes[40, 124], or are limited by restricting mutations to alanine[7].

Recently, yeast surface display[63] coupled to deep mutational scanning[13] was used to understand the sequence effects of binding for nearly every single point mutant for two computationally designed proteins targeting a conserved epitope on Influenza hemagglutinin[4]. This method was used to confirm the paratope for both small proteins, as validated by crystal structures. More recently, other approaches using yeast display and deep sequencing for the purposes of conformational epitope mapping have been demonstrated [123, 124]. However, current implementations require several sorting steps that severely hinder throughput. Because additional inefficiencies exist at several stages in the deep sequencing and analysis workflow, we asked whether we could simplify the yeast display-deep sequencing pipeline to increase the method throughput, reduce cost, and improve the ability to resolve complete conformational epitopes for full-length proteins.

## 4.2 Materials and Methods

### 4.2.1 Strains

*E. coli* strains used in this study: XL1-Blue (Agilent, Santa Clara, CA) recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F' proAB lacI1$^{q}$ZΔM15 Tn10 (Tet$^{r}$)]; BL21* (Life Technologies, Carlsbad, CA) fhuA2 [lon] ompT gal [dcm] ΔhsdS; BL21(DE3) (New England BioLabs, Ipswich, MA) fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5. Yeast strain used in this study: EBY100 (American Type Culture Collection, Manassas, VA) MATa AGA1::GAL1-AGA1::URA3 ura3-52 trp1 leu2-delta200 his3-delta200 pep4::HIS3 prb11.6R can1 GAL.

### 4.2.2 Plasmids

The plasmid pETCON_TNF was created by inserting a codon optimized gene encoding the Gly57-Leu233 extracellular portion of tumor necrosis factor (TNF) (GenScript, Piscataway, NJ) into the pETCON plasmid using flanking NdeI/XhoI restriction sites. The plasmid pETCON_PTx-S1-220 was created by amplifying PTxS1-220 from pAK400_PTx-S1-200K[125] and inserting into pETCON at NdeI/XhoI sites. To create pETCOn_TROP2Ex, DNA was isolated from HeLA cells with the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich, St. Louis, MO). This DNA was used as a template for amplication of the ectodomain of Trop2 encompassing residues 27-274, which was then inserted into pETCON at Nde/XhoI sites. Polypeptide sequences of the variable regions for the heavy and light chains of Infliximab were obtained[126] and used to generate codon optimized DNA sequences (GenScript Piscataway, NJ). A $(Gly_4Ser)_3$ linker was placed between the C-terminal residues of the heavy and N-terminal residue of the light chains. The plasmid pET29b_inflix_scfv was prepared by inserting the inflix_scFv gene into the pET-29b(+) vector (EMD BioSciences, Billerica, MA) using NdeI/XhoI restriction

sites. m7e6 heavy and light chains were subcloned into 293-6E expression vector pTT5 were custom ordered from Genscript (Piscataway, NJ). Plasmids for the yeast display constructs have been deposited in the AddGene plasmid repository (www.addgene.org).

### 4.2.3 Preparation of inflix_scFv

pET29_inflix_scFv was transformed into chemically competent BL21*(DE3). Cultures were grown to an $OD_{600}$ of 0.8 and where then induced with 1 mM IPTG and incubated with shaking at 18° C for ~18 hours. Inflix-scFv was isolated from inclusion bodies and refolded using existing protocols[127, 128]. After the refolding procedure the sample was centrifuged at 17,000 xg to remove precipitated protein. The concentration was determined using the Bradford method using BSA as a protein standard and was biotinylated at a molar ratio of 1:20 protein:biotin using the EZ link NHS-biotin kit following the manufacturers instructions (Life Technologies, Carlsbad, CA).

### 4.2.4 Preparation of Trop2 and PTxS1 Fabs

PTxS1 antibody hu1B7 was prepared according to previous reports[129]. Anti-TROP2 monoclonal antibody m7E6[130] was produced in 293-6E cells and purified by protein A column by Genscript (Piscataway, NJ). Fabs were produced using the Pierce Fab Preparation Kit (Life Technologies, Carlsbad, CA). Concentrations were determined using A280 with the recommended estimated extinction coefficient (1 mg/mL) of 1.4 and was biotinylated at a molar ratio of 1:20 protein:biotin using the EZ link NHS-biotin kit following the manufacturers instructions (Life Technologies, Carlsbad, CA).

### 4.2.5 Dissociation Constant Determination

Equilibrium dissociation constants ($K_D$) were determined using clonal population yeast display titrations according to Chao et al. [63]. Fab concentrations between 50 pM and 1 μM were tested.

### 4.2.6 Yeast Display Sorts

$1x10^7$ cells were grown in 2 mL SDCAA for 6 hours at 30°C and re-inoculated at $OD_{600}$ =1.0 in 2 mL SGCAA at 20°C for 18 hours. $3x10^7$ cells were labeled with biotinylated Fab or scFv for 30 minutes at room temperature in DPBSF (Dulbecco's Phosphate-Buffered Saline with 1 g/L BSA) at a concentration of half of the experimentally determined dissociation constant on the yeast surface. Cells were then secondarily labeled with anti-cmyc-FITC (Miltenyi Biotec, San Diego, CA) and streptavidin-phycoerythrin (SAPE) (Thermo Fisher, Waltham, MA). Sorting was done on an Influx Cell Sorter (Becton Dickinson, Franklin Lakes, NJ). FSC/SSC (gate 1), FSC/FITC (gate 2), and PE/FITC (gate 3) gates were set. Three populations were collected: an unselected population satisfying gate 1, a displayed population satisfying gates 1 and 2, and a bound population satisfying all three gates. The number of cells collected for each population was at least 100-fold higher than the theoretical library complexity. Sorting statistics for each population collected are listed in Table 7. Following the sort, recovered populations were grown for 48 hours in 10 mL SDCAA[63], and $1x10^7$ cells from this culture cells were stored in 1mL of yeast storage buffer at -80°C.

### 4.2.7 Deep Sequencing Preparation

Yeast plasmid DNA was prepared for deep sequencing following the protocol in Kowalsky et al.[131] 5 µl of the PCR products were run on a 2% agarose gel stained with SYBR-GOLD (Thermo

**Table 7- Sorting Statistics**

| | Tile Length (AA) | Minimum Transformants for 99.9% Coverage | Sort Labeling Conditions (nM) | Events Collected for Binding Population | Percent Sorted (Display) | Percent Sorted (Binding) |
|---|---|---|---|---|---|---|
| **TNF-Infliximab Tile 1** | 60 | 26,880 | 32 | 200,641 | 46.0% | 6.9% |
| **TNF-Infliximab Tile 2** | 60 | 26,880 | 32 | 200,437 | 43.0% | 5.1% |
| **TNF-Infliximab Tile 3** | 57 | 25,536 | 32 | 206,251 | 31.9% | 6.5% |
| **PTxS1-hu1B7 Tile 1** | 72 | 32,256 | 3 | 400,000 | 43.0% | 7.7 % |
| **PTxS1-hu1B7 Tile 2** | 74 | 33,152 | 3 | 400,000 | 45.0% | 6.1% |
| **cdPTxS1-hu1B7 Tile 3** | 73 | 32,704 | 3 | 400,000 | 46.2% | 6.4% |
| **Trop2-m7E6 Tile 1** | 82 | 36,736 | 22 | 400,000 | 19.8% | 6.6% |
| **Trop2-m7E6 Tile 2** | 83 | 37,184 | 22 | 400,000 | 24.7% | 6.7% |
| **Trop2-m7E6 Tile 3** | 83 | 37,184 | 22 | 400,000 | 16.9% | 6.6% |

Fisher, Waltham, MA) to ensure one band was obtained at the correct size (~250-350 bp). Agencourt AMPure XP PCR Purification (Beckman Coulter, Beverly, MA) was used per the manufacturer's protocol to purify the PCR product. Library DNA was sequenced on an Illumina MiSeq using the either the 300x2 or 250x2 Illumina MiSeq kits (Illumina, San Diego, CA) at the Michigan State University Sequencing Core.

### 4.2.8 Data Analysis

A modified version of Enrich-0.2 as described in Kowalsky et al. [131] was used to compute enrichment ratios of individual mutants from the raw Illumina sequencing files. To normalize the data across the multiple tiles we define the fitness metric for variant i $\left(\zeta_i\right)$ as the binary logarithm of the fluorescence of variant i to the fluorescence of the wild-type sequence $\left(\overline{F}_{wt}\right)$[131]:

$$\zeta_i = \log_2\left(\frac{\overline{F}_i}{\overline{F}_{wt}}\right) \tag{1}$$

This results in the following equation:

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

Where $\phi$ is the percentage of cells collected, $\varepsilon_i$ is the enrichment ratio for variant I, $\sigma'$ is the log normal standard deviation of a clonal population, and the subscript *wt* denotes the wild-type. Custom python scripts used to calculate the fitness metric and statistics are at Github [user: JKlesmith] (www.github.com). The full deep sequencing datasets are provided at figshare (www.figshare.com). Sorting parameters needed for each tile normalization are listed in Table 7.

Shannon (sequence) entropy values at a given position j in the protein sequence ($E_j$) were calculated by the following equation:

$$E_j = -\frac{\ln(20)}{\ln(X_j)}\sum_{i=A}^{Y}\frac{2^{\varepsilon_{ij}}}{\sum_{k=A}^{Y}2^{\varepsilon_{kj}}}\,ln\frac{2^{\varepsilon_{ij}}}{\sum_{k=A}^{Y}2^{\varepsilon_{kj}}} \tag{3}$$

Where $\varepsilon_{ij}$ is the enrichment ratio of substitution i at position j, and $X_j$ is the number of mutants with adequate sequencing counts in the unselected population at position j. The derivation for equation (3) is shown in the THEORY section. We excluded residues with an $X_j<12$ from analysis. Enrichment ratios for stop codons were not included in the Shannon entropy analysis.

### 4.2.9 Soluble expression of PTxS1

BL21(DE3) cells containing pAK400_ PTx-S1-220K[125] with an added C-terminal lysine residue were grown in TB at 25ºC to an $OD_{600}$ of 1.5, then induced with 1mM IPTG for 5 hours. Cell pellets were collected and the outer membrane lysed by osmotic shock. Lysate was purified by immobilized metal affinity chromatography, followed by size exclusion chromatography in PBS (S75, AKTA FPLC) as previously reported[125].Yields were between 0.2-0.6 mg/L. 3 μg of PTx (26.1 KDa) or PTx-S1-220K (25.9 KDa) were run on a 12% poly-acrylamide gel at 130V and stained with GelCode blue.

### 4.2.10 PTxS1 ELISA Binding Assay

A high-binding 96-well plate (Costar, Corning, NY) was coated with either 4nM PTx or 4nM PTx-S1-220k in PBS and incubated overnight at 4C. The plate was blocked with a solution of 5% nonfat dry milk in PBS with 0.05% Tween (PBSTM) for 1 hour at room temperature. hu1B7 was serially diluted across the plate in duplicate with a starting concentration of 5 μg/mL in PBSTM and incubated for 1 hour at room temp. Secondary antibody was GαhFc-HRP prepared at 1:2500 in PBSTM and incubated for 1 hour at room temp. Plate was developed with TMB, quenched with HCl, and absorbance read on a plate reader at 450 nm.

### 4.2.11 PTxS1 Western Blot

0.3 μg of PTx or PTx-S1-220K were run on a 12% poly-acrylamide gel at 130V and transferred to a PVDF membrane. The membrane was blocked for 1 hour at room temperature with PBSTM, then incubated for 1 hour at room temperature with 1 μg/mL of hu1B7A in PBSTM. The secondary antibody was GαhFc-HRP prepared at 1:10,000 in PBSTM and incubated for 1 hour at room temp, followed by washing and development using SuperSignal West Pico Chemiluminescent Substrate (Pierce, Gran Island, NY) and a 30s exposure time to x-ray film.

### 4.2.12 Cell Culture and Transfection

MCF10A human breast epithelial cell lines were obtained from Dr. Kathleen Gallo at Michigan State University. MCF10A cells were cultured in Dulbecco's modified Eagles's media (DMEM)/F12(1:1) (Life technologies, Grand Island, NY, USA) with 5% horse serum, 10 ug/ml insulin, 20 ng/ml EGF, 100 ng/mlcholeratoxin, 0.5 ug/ml hydrocortisone and penicillin/streptomycin (Invitrogen, Carlsbad, CA, USA). MDA-MB-231 human breast cancer cell lines were obtained from ATCC and cultured in DMEM with 10% fetal bovine serum (FBS), 2mM glutamine and penicillin/streptomycin (Invitrogen), as described[132, 133]. Cells were maintained at 37°C and 5% $CO_2$ environment.

SiGENOMESMARTpool ON-TARGETplusTROP2siRNA and scramblesiRNA were purchased from Thermo Scientific (Asheville, NC, USA) and Ambion (Grand Island, NY, USA). The SMARTpoolsiRNAs and the transfection reagent Lipofectamine were diluted with Opti-MEM (Invitrogen, Grand Island, NY, USA), respectively as described[134, 135]. The diluted SMARTpoolsiRNAs were mixed with RNAiMAX to form siRNA-RNAiMAX complexes. The cell culture medium was replaced with antibiotic-free medium containing the siRNA-RNAiMAX complexes at a final concentration of 10nM siRNA. Media were changed after 12hr and the cells were incubated in fresh media.

### 4.2.13 TROP2 Western Blot Analysis

The protein concentrations of the cell extracts were measured by the Bradford method. Protein samples of 15-30 μg were subjected to Western blot analysis as previously described[135, 136] using TROP2 antibody (Abcam, Cambridge, MA), beta-actin (Sigma, St. Louis, MO, USA), anti-mouse and anti-rabbit HRP-conjugated secondary antibodies (Thermo Scientific, Asheville, NC)

and donkey anti-goat IgG-HRP (Santa Cruz Biotech, Dallas, TX). The blots were visualized by SuperSignal West Femto maximum sensitivity substrate (Thermo Scientific, Asheville, NC).

### 4.2.14 Total mRNA Extraction and Quantitative real time PCR

Total mRNA from cells was extracted using the RNeasy Plus kit (QiaGen, Valencia, CA, USA) according to the manufacturer's instructions. The total mRNA was reverse transcribed into cDNA using the cDNA synthesis kit (BioRad, Hercules, CA, USA) as previously described[137, 138]. The following primer sets (Operon, Huntsville, AL, USA) were used for PCR:human actin (5'-tggacttcgagcaagagatg-3' and 5'- aggaaggaaggctggaagag-3') and human TROP2 (5'-gagattcccccgaagttctc-3' and 5'-aactccccccagttccttgat-3'). Quantitative real-time PCR was performed using iQSYBR Green Supermix and Real-Time PCR Detection System (BioRad, Hercules, CA). The cycle threshold values were determined by the MyIQ software (BioRad, Hercules, CA).

### 4.2.15 Transwell Migration and Invasion Assays

Chemotactic migration or invasion was quantified using a Boyden chamber transwell assay (8 μm pore size; Corning Costar, Cambridge, MA, USA), with either uncoated or matrigel-coated filters, respectively. Cells were deprived of serum overnight, trypsinized and introduced into the upper chamber. MitomycinC (Sigma Aldrich, St. Louis, MO) was added to the cultured media. The chemoattractant in the lower chamber was medium supplemented with 5% FBS. After 8 hours incubation at 37°C, the cells were fixed and stained. Migrated cells in five randomly chosen fields were counted. The experiments were performed in triplicate wells and each experiment was performed three times as indicated.

### 4.2.16 Wounding-Healing Assay

MDA-MB-231 cells were grown to confluence. The growth medium was replaced with fresh medium containing 5% FBS and supplemented with mitomycinC (1 mg/ml) (Sigma Aldrich,

St. Louis, MO) and the monolayer of cells was subsequently scratched using a 200 µl pipette tip. Wound width was monitored over time by microscopy.

### 4.2.17 Cytotoxicity and proliferation assays

Cytotoxicity and proliferation were assessed using LDH and AlamarBlue microplate assays, respectively. Cells were seeded in 96-well plates and treated with either PBS (control) or various concentrations of m7EG IgG or Fab for 48 hours. Cytotoxicity was detected using the LDH cytotoxicity detection kit (Roche, San Francisco, CA) following the manufacturer's protocol. Proliferation was determined by AlamarBlue assay (Pierce, Waltham, MA) following the manufacturer's protocol.

### 4.2.18 Confocal microscopy

MDA-MB-231 cells were seeded in glass-bottom 24-well plates (In Vitro Scientific, Sunnyvale, CA) and treated with either PBS (control), m7E6 IgG or Fab. After treatment, cells were washed 2x with ice-cold PBS, fixed with 3.7% formaldehyde for 15 min. at 37°C and washed 3x with PBS. Next, the samples were incubated in blocking buffer (1% BSA, 0.5% TritonX-100 in PBS) for 1 hour at 37°C. The cells were incubated with TROP2 intracellular domain-specific primary antibodies (EMD Millipore #ABC425, 1:500 dilution) in incubation buffer (0.5% BSA, 0.5% TritonX-100) overnight at 4°C. The samples were then washed 3x with PBS and incubated in respective secondary antibodies (AlexaFluor 488) diluted in PBS for 1 hour at 37°C in the dark. Cells were washed 2x with PBS and incubated in nuclear counter stain Hoechst 3342 (Invitrogen) for 10 min. at room temperature. After the final incubation, cells were washed twice with PBS and covered with anti-fade solution for imaging. Images were recorded with an Olympus FluoView 1000 Inverted IX81 microscope, using a 10X or 60X oil objective using identical exposure and PMT settings for each primary antibody-fluorophore pair across the different treatment conditions.

### 4.2.19 Statistical Analysis

All experiments were performed at least three times. Representative results are shown as mean ± standard deviation. Statistical analysis were performed using an unpaired, two tail student t-test. * indicates $p<0.05$, ** indicates $p<0.01$ and *** indicates $p<0.001$.

## 4.3 Theory

### 4.3.1 Derivation of sequence entropy metric and calculation of estimated errors

At any given position j in a protein sequence, Shannon entropy for that position ($E_j$) is defined as:

$$E_j = -\sum_{i=A}^{Y} P_{ij} ln P_{ij} \quad (4)$$

Here, $P_{ij}$ is the probability of finding amino acid i at position j after sorting given an equal representation of all 20 amino acids in the starting population. To determine $P_{ij}$, we first define $p_{ij}$ for a single substitution at position j using the frequency of mutant i in the initial ($f_{o,ij}$) and final ($f_{f,ij}$) populations:

$$p_{ij} = \frac{f_{f,ij}}{f_{o,ij}} \quad (5)$$

This can be written in terms of an enrichment ratio of a given substitution i at position j ($\varepsilon_{ij}$), such that:

$$p_{ij} = 2^{\varepsilon_{ij}} \quad (6)$$

$p_{ij}$ gives the probability of a variant with a mutation i at position j passing through the sort. Because the summation of these probabilities will not necessarily sum to unity, we can normalize the probabilities over a single residue such that:

$$P_{ij} = \frac{2^{\varepsilon_{ij}}}{\sum_{k=A}^{Y} 2^{\varepsilon_{kj}}} \quad (7)$$

Combining with the definition of Shannon Entropy in (1):

$$E_j = -\sum_{i=A}^{Y} \frac{2^{\varepsilon_{ij}}}{\sum_{k=A}^{Y} 2^{\varepsilon_{kj}}} ln \frac{2^{\varepsilon_{ij}}}{\sum_{k=A}^{Y} 2^{\varepsilon_{kj}}} \quad (8)$$

Because some positions do not have adequate sequencing counts for all 20 amino acids, the sequence entropy metric must be normalized by the maximum possible Shannon entropy:

$$E_j = -\frac{\ln(20)}{\ln(X_j)} \sum_{i=A}^{Y} \frac{2^{\varepsilon_{ij}}}{\sum_{k=A}^{Y} 2^{\varepsilon_{kj}}} ln \frac{2^{\varepsilon_{ij}}}{\sum_{k=A}^{Y} 2^{\varepsilon_{kj}}} \quad (9)$$

Where $X_j$ is the number of mutants with adequate sequencing counts in the unselected population at position j. Equation (6) is the final form of the sequence entropy metric used in the manuscript. In practice, we excluded residues with an $X_j$<12. This removed 13/647 (2.0%) positions tested in the present work.

To estimate the expected error in $E_j$, we can define the variance in $E_j$ as:

$$\sigma_{E_j}{}^2 = \sum_{i=A}^{Y} (lnP_{ij} + 1)^2 \, \sigma_{P_{ij}}{}^2 \quad (10)$$

Similarly, the variance on $P_{ij}$ can be defined as:

$$\sigma_{P_{ij}}{}^2 = \left(\frac{1}{\sum_{i=A}^{Y} 2^{\varepsilon_{ij}}}\right)^2 \left[\left[\sigma_{2^{\varepsilon_{ij}}}{}^2 + \left(\frac{2^{\varepsilon_{ij}}}{\sum_{i=A}^{Y} 2^{\varepsilon_{ij}}}\right) \frac{2^{\varepsilon_{ij}}}{\sum_{i=A}^{Y} 2^{\varepsilon_{ij}}}\right\}^2 \left(\left(\sigma_{\sum_{i=A}^{Y} 2^{\varepsilon_{ij}}}\right) \sigma_{\sum_{i=A}^{Y} 2^{\varepsilon_{ij}}}\right)^2\right] \quad (11)$$

Because the minimal error associated with counting sequences approximates Poisson noise[4, 131], we can write the variance for the two unknowns as:

$$\sigma_{2^{\varepsilon_{ij}}}{}^2 = (log2)^2 \left(\log_2 e\right)^2 \left(\frac{1}{x_{f,ij}} + \frac{1}{x_{o,ij}}\right) (2^{\varepsilon_{ij}})^2 \quad (12)$$

Here, $x_{f,ij}$ and $x_{o,ij}$ are the raw sequencing counts of mutation i at position j in the final and initial population, respectively. We can write a similar derivation for the variance of the denominator in the probability term:

$$(\sigma_{\sum_{i=A}^{W} 2^{\varepsilon_{ij}}})^2 = (log2)^2 (\log_2 e)^2 \sum_{i=A}^{Y} (2^{\varepsilon_{ij}})^2 \left(\frac{1}{x_{f,ij}} + \frac{1}{x_{o,ij}}\right) \quad (13)$$

Accurate calculation of the variance on the sequence entropy for a given position requires the raw sequencing counts for each position in the unselected and selected populations. Using these numbers, equations (12) and (13) can be solved, which can then be plugged into (10) and (11) to solve for the unknown. The variance on sequence entropy should be maximized when the raw sequencing counts are just above the inclusion threshold in the unselected populations. Even in this case, the standard deviation of sequence entropy metric is 0.02. Accordingly, we are justified in not including sequence entropy errors in the determination of the conformational epitope.

### 4.3.2 Calculation of relative dissociation constants from sequencing counts

For a clonal population of yeast cells displaying variant i and labeled with antibody at a labeling concentration of $[L_o]$, we can write the mean fluorescence ( $\overline{F_i}$) as:

$$\overline{F_i} = \frac{(F_{max} - F_{min})}{\frac{K_d}{[L_0]} + 1} + F_{min} \tag{14}$$

Here $F_{max}$ and $F_{min}$ are, respectively, the maximum and minimum average fluorescence for clone i in the fluorescence channel used for antibody binding. Using the labeling conditions $[L_o] = ½ K_{d,wt}$, equation (14) becomes:

$$\overline{F_i} = \frac{(F_{max} - F_{min})}{2\frac{K_d}{K_{d,wt}} + 1} + F_{min} \tag{15}$$

Similarly, the mean fluorescence for the wild-type variant can be expressed as:

$$\overline{F_{wt}} = \frac{(F_{max} - F_{min})}{3} + F_{min} \tag{16}$$

The ratio of mean fluorescence for a variant to wild-type can be written in terms of the fitness metric ($\zeta_i$) derived from the sequencing data[131]:

$$2^{\zeta_i} = \frac{\overline{F}_i}{\overline{F}_{wt}} \tag{17}$$

Substitution of equation (16) into equation (17) leads to:

$$\overline{F_{wt}} = \frac{\frac{(F_{max}-F_{min})}{2\frac{K_d}{K_{d,wt}}+1}+F_{min}}{\frac{(F_{max}-F_{min})}{3}+F_{min}} \quad (18)$$

If we assume that $F_{max} >> F_{min}$, equation (18) simplifies to:

$$\frac{K_d}{K_{d,wt}} = \frac{3}{2^{\zeta_i+1}} - \frac{1}{2} \quad (19)$$

This assumption is valid for the ratio of $F_{max}/F_{min}$ typically seen in yeast display. The ratio depends on multiple factors, including the sensitivity of the fluorescent detection on the cell sorter, the quantum yield of the fluorescent dye used, the biotin labeling per antibody, and the surface expression of a given antigen. In our hands, we observe a range from 50-2000 for this ratio. Here the range of $K_d$ values that we can see is relative to the interval of fitness metrics observed from the sequencing counts. Practically speaking, the range on the lower end of fitness metrics should be the average fitness metric for the stop codons (-1.1 to -0.75 depending on the percent collected), which gives a relative dissociation constant of approximately 2.7. This highlights the sensitivity of the method for differentiating small energetic changes in binding activity ($\Delta\Delta G_{binding}$ ~ 0.1-0.4 kcal/mol). The drawback, however, is we are unable to discriminate smaller perturbations with larger energetic changes typically associated with interface "hot spots" ( $\Delta\Delta G_{binding}$ ~ 1-2 kcal/mol).

Another important consideration is the error associated with digital counting of variants from deep sequencing data as this will also introduce error on both the fitness metric and corresponding relative dissociation constants for variants with low numbers of counts in the unselected

population. We have previously shown that digital counting of variants from deep sequencing error using our methods result in minimal error associated with Poisson noise[131].

The variance on the fitness metric can be defined as:

$$\sigma_{\zeta_i}^2 = \sigma_{\varepsilon_i}^2 \left(\frac{\partial \zeta_i}{\partial \varepsilon_i}\right)^2 + \sigma_{\varepsilon_{wt}}^2 \left(\frac{\partial \zeta_i}{\partial \varepsilon_{wt}}\right)^2 \tag{20}$$

Where $\varepsilon_i$ is the enrichment ratio for variant i. The variance for $\varepsilon_i$ ($\sigma_{\varepsilon_i}^2$) can be estimated from Poisson noise as[139]:

$$\sigma_{\varepsilon_i}^2 = \left(log_2 e\right)^2 \left(\frac{1}{x_{fi}} + \frac{1}{x_{o_i}}\right) \tag{21}$$

Where $x_{oi}$ and $x_{fi}$ are the number of counts in the unselected and selected populations respectively. For variants with many counts the error approaches zero, highlighting the importance of sequencing depth of coverage in these experiments. The fitness metric is defined as[131]:

$$\zeta_i = (log_2 e)\sqrt{2}\sigma'\left(erf^{-1}(1 - \phi 2^{\varepsilon_{wt}+1}) - erf^{-1}(1 - \phi 2^{\varepsilon_i+1})\right) \tag{22}$$

Where $\phi$ is the percentage of cells collected from the respective gate(s), $\sigma$' is the log normal standard deviation of a clonal population, and the subscript wt denotes the wild-type.

Combining these we can estimate the variance of the fitness metric as:

$$\sigma_{\zeta_i}^2 = \pi\phi^2\sigma'^2 \left\{\sigma_{\varepsilon_i}^2 \left[(2^{\varepsilon_i+0.5}) e^{erf^{-1}\left(1-\phi 2^{\varepsilon_i+1}\right)^2}\right]^2 + \sigma_{\varepsilon_{wt}}^2 \left[(-2^{\varepsilon_{wt}+0.5}) e^{erf^{-1}\left(1-\phi 2^{\varepsilon_{wt}+1}\right)^2}\right]^2\right\} \tag{23}$$

The error here is largest with variants with low fitness metrics and few counts in the unselected population.

The error for the relative dissociation constant is defined as:

$$\sigma^2_{K_{d,i}/K_{d,wt}} = \sigma^2_{\zeta_i}\left(\frac{\partial^{K_{d,i}/K_{d,wt}}}{\partial\zeta_i}\right)^2 \quad (24)$$

Which can be written as:

$$\sigma^2_{K_{d,i}/K_{d,wt}} = \sigma^2_{\zeta_i}\left(\frac{-3ln(2)}{2^{\zeta_i+1}}\right)^2 \quad (25)$$

**4.4**

**4.5 Results**

The streamlined method is shown schematically in Figure 18**.** In the first step, single site saturation mutagenesis (SSM) libraries for 250-300 nt contiguous sections of the gene of interest are prepared by PFunkel mutagenesis[51] and transformed into yeast. These yeast libraries are labeled with a biotinylated Fab or scFv and sorted once by FACS. The labeling concentration and FACS gates are set such that the capture probability of any given variant is a monotonically increasing function of its binding affinity. Three distinct populations are collected: an unselected population of cells that pass through a cell size gate (unselected population), a displayed population of cells passing through the previous gate as well as a gate confirming display of the C-terminal c-myc epitope tag (displayed population), and a binding population of cells satisfying these two previous gates as well as a gate on the fluorescence channel associated with antibody binding (bound population). The DNA from each population is prepared and sequenced on an Illumina platform. Then, the frequencies of each variant in the population are compared and merged into a single fitness metric[131] that allows direct, quantitative comparisons across different mutational libraries. Together, this approach allows for the rapid and comprehensive reconstruction of the sequence-binding determinants of full-length proteins for a given antibody.

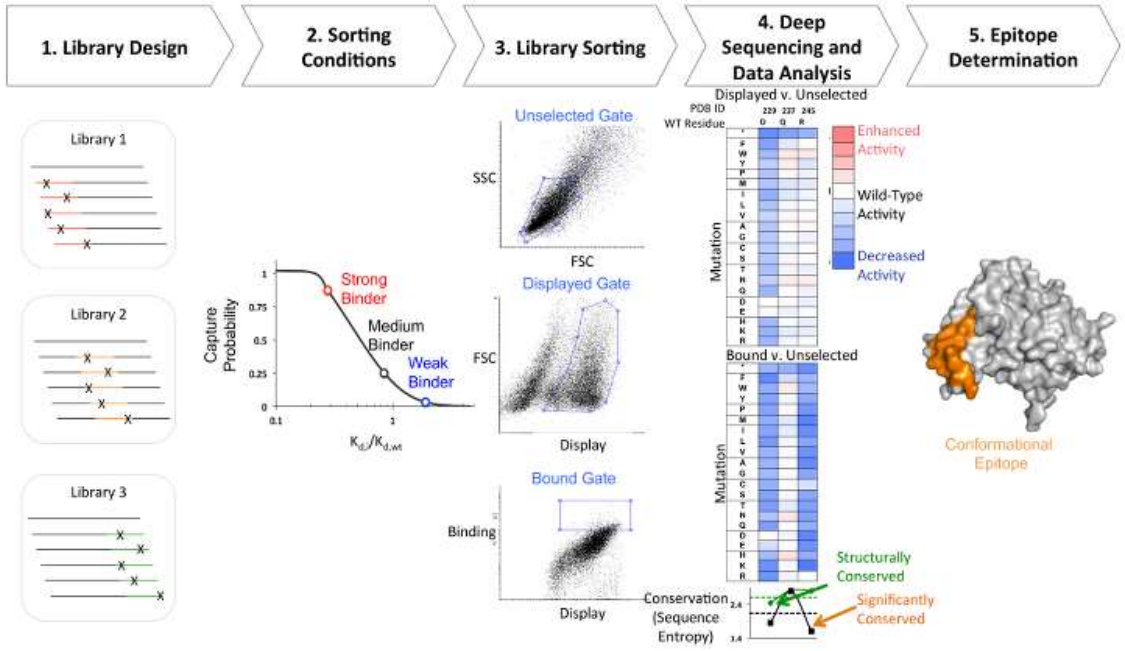



**Figure 18- Schematic of streamlined conformational epitope mapping process.**

*(1.)* SSM libraries are made for 250-300 nt contiguous sections along the gene of interest. Each library contains mutations in a different section of the gene. *(2.)* Sorting conditions are determined such that there is a higher probability of capturing stronger binding cells. *(3.)*Yeast libraries are labeled with biotinylated Fab and sorted by FACS. Three different gates are drawn: a gate on light scattering parameters SSC/FSC (top; unselected population), a gate set on FSC and fluorescence channel corresponding to display of antigen (middle; displayed population) and a binding gate that collects the top 5-10% of cells by fluorescence corresponding to channel for bound antibody (bottom; bound population). *(4.)* DNA from each population is extracted and prepared for deep sequencing on an Illumina platform. The frequency of each variant in the bound and displayed populations is compared against the unselected population and used to calculate a fitness metric. For each residue, sequence entropy (bottom) for bound (black) and displayed (green) sorts is used to determine the degree of conservation. *(5.)*Sequence entropy is used to identify conserved and non-conserved residues that are used to determine the conformational epitope (orange).

As a first test of our approach we chose to evaluate the binding of yeast-displayed Tumor Necrosis Factor (TNF; TNF□) on the monoclonal antibody Infliximab (marketed by Johnson & Johnson as Remicade). A structure for this complex is known[140], allowing assessment of the ability of the sequence-function method to demarcate discontinuous conformational binding epitopes. TNF is a homo-trimeric, multi-disulfide linked, marginally sTable protein and thus represents a

stringent test of the ability of yeast to surface display complicated proteins. We ordered a codon-optimized gene encompassing the Gly57-Leu233 extracellular portion of TNF and subcloned it into the yeast display plasmid pETCON[29]. Next, we created three single-site saturation mutational (SSM) libraries of TNF and induced yeast surface expression of library variants[63]. For each library we performed a single FACS sort using cells labeled with the biotinylated scFv of Infliximab (inflix_scFv) at 32 nM, which is half of the observed dissociation constant for the interaction. Approximately 200,000 cells from each library were collected for the unselected, displayed, and bound populations. Deep sequencing was used to determine the enrichment ratios of the bound and displayed populations compared to the unselected population. These enrichment ratios were then transformed to a fitness metric that allows direct comparisons across the different mutational libraries, allowing the sequence determinants of binding to be evaluated for nearly every possible single point mutant in the protein sequence (Figure 19). Overall, we observed 95.1% coverage of all possible single non-synonymous mutants in the extracellular TNF sequence (n=2985/3140) .

To identify the conformational epitope, we reasoned that residues essential to the protein-protein interaction would be conserved in the bound population. Conversely, residues that do not participate at the protein-protein interface would be mostly non-conserved. To discriminate among these positions we introduced a positional Shannon (sequence) entropy metric that is calculated using the enrichment ratios of every variant at a given position (see THEORY section). Because sequence conservation will depend strongly on the stringency of the sorting conditions, we next asked for cutoffs to discriminate between a conserved and non-conserved position. We considered a position to be conserved if the sequence entropy in the bound population compared with the unselected population was less than or equal to the midpoint of the sequence entropy range. Even using this stringent cutoff, 56/177 TNF residues were identified as conserved. Many of these

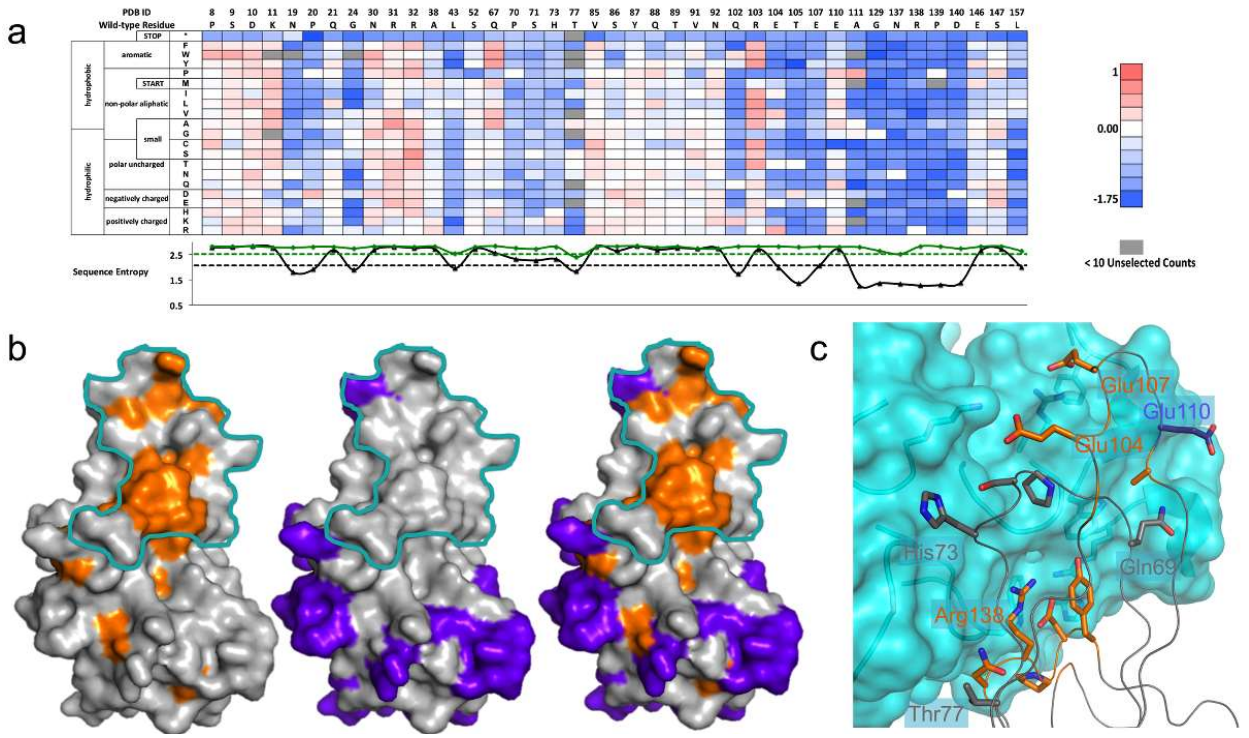



**Figure 19- TNF-Infliximab conformational epitope determination.**

a. A subset (41/177 residues) of the fitness-metric heat map for bound population of the TNF-inflix_scFv interaction. Sequence entropy for the display (green) and bound population (black) is plotted below with their respective cut-offs (dashed lines). b. Subtractive sequence entropy analysis for TNF-Infliximab interaction. Conserved residues (orange) are found mainly within the binding footprint of the TNF-Infliximab interaction (cyan). Non-conserved residues (purple) can also be mapped onto structure and fall outside of the footprint (middle). These non-conserved residues can be used to find regions where false positive conserved residues appear. For clarity, only one TNF monomer is shown. c. Close-up view of the structural interface between TNF (ribbon) and Infliximab (cyan surface). TNF residues are colored according to sequence conservation as in panel b.

residues are buried in the core of TNF, and presumably disrupt the fold of the protein. Positions located at the epitope can be partially discriminated from those that disrupt protein stability by calculating the sequence entropy of the displayed sort and using a cutoff of the midpoint of the sequence entropy range. This analysis removed 22 of the 56 residues from consideration as epitope positions. The removed residues were almost all buried, with a mean fraction solvent accessible surface area of 0.03 (range 0.00-0.22). The 34 remaining residues were a combination of surface

positions (using a fraction accessible surface area cutoff of 0.10, n=18) and buried position (Figure 19).

The conserved surface positions clustered in three regions on TNF (all subsequent residues use PDB numbering): the AB loop (Asn19, Pro20, Gly24), the EF loop (Gln102, Glu104, Thr105, Glu107, Ala111), and the GH loop (Asn137-Tyr141). There are also a handful of noncontiguous, partially surface-exposed positions scattered throughout (Figure 19). To further discriminate epitope from non-epitope positions, we reasoned that the epitope would be depleted in non-conserved positions. Identifying non-conserved positions as the upper quintile of the sequence entropy range removed 48/177 positions from consideration. Of these, only Glu110 was within 4Å of Infliximab in the bound complex. However, the C□-C□ vector of Glu110 is pointed away from the interface and its side chain does not make significant interactions to Infliximab (Figure 19) suggesting that its mutation to other amino acids would not disrupt the affinity of the TNF-Infliximab complex.

Considering both the conserved and non-conserved positions highlights the EF loop and GH loop as essential to the interaction. The single most conserved section documented by sequence entropy analysis is on the EF loop between Asn137-Tyr141, and these residues map neatly to the center of the experimentally determined binding region (Figure 19). Additionally, conservation of several residues on the EF loop is consistent with the Infliximab-TNF structure, including Glu107 that makes a salt bridge interaction across the interface. Furthermore, the importance of these two loops to the energetics of the interaction have been confirmed by mutagenesis[140].

Examination of non-conserved residues located at the interface identifies limitations in using a single metric for epitope determination. For example, Pro70, Ser71, and His73 on the CD loop and Thr77 in strand D are interface residues that are potentially energetically important but are

above the sequence entropy cutoff (Figure 19). For the CD loop residues the positions are conserved but the sequence entropy is just slightly above the cutoff. Proline at position 70 and serine 71 are the most favored amino acid, whereas a substitution of His73Lys is slightly favored. Thr77 is removed from consideration of the epitope as it is relatively conserved in the displayed population. Additional epitope residues that were not identified include the above-mentioned Glu110 and Gln67. However, neither is expected to be energetically significant as determined by alanine scanning mutagenesis and its position in the bound complex. Indeed, mutation of Gln67 to aromatic residue increases binding affinity for the Infliximab interaction (Figure 19). Based on this comprehensive mutagenesis dataset enabled by deep sequencing, we conclude that this improved yeast display-deep sequencing pipeline is effective in identifying fine conformational epitopes for antibody-antigen interactions.

We next asked whether the automated identification of the epitope using sequence entropy can be used to map binding footprints of other antibody-antigen interactions. To accomplish this, we evaluated the binding of yeast-displayed Pertussis Toxin subunit 1 (PTxS1) against a single humanized neutralizing antibody. Whooping Cough, a respiratory disease caused by the bacteria *Bordetella pertussis*, remains a major cause of infant mortality in both the developing and industrialized countries despite widespread vaccination[141]. Recently, Nguyen et al. demonstrated the ability of a binary antibody cocktail to halt whooping cough disease progression in a baboon model[129]. Although one of the cocktail antibodies, hu1B7, is able to bind to the S1 subunit on a Western Blot indicating a linear component of the epitope, previous studies using 15-mer peptides covering the entire S1 sequence were unable to identify a peptide showing binding activity against murine 1B7 [142]. Further information about the epitope on S1 targeted by hu1B7, one of the cocktail antibodies, will help elucidate its neutralizing mechanism.

Following a similar procedure to that of TNF, we used PFunkel mutagenesis to create a SSM library of nearly all possible single point mutants in the Asp1-Gly220 fragment of PTxS1 (PTx-S1-220) and performed a single FACS sort collecting 400,000 cells in each of the three populations. Soluble PTx-S1-220 can be expressed in *E. coli* and retains affinity for hu1B7 (Figure 20). Positional Shannon entropy was used to determine the most conserved residues at the interface using the same cutoffs identified in the TNF test case conservation of buried residues Phe84, Gly86 and His149 near the identified epitope indicate that the hu1B7 binding affinity depends somewhat on the conformation in the PTxS1 folded state.

As before, epitope residues were discriminated from residues that result in disruption of the protein fold by analysis of sequence conservation in the displaying population. Altogether, this procedure identified sixteen residues at the proposed antibody-antigen conformational epitope:

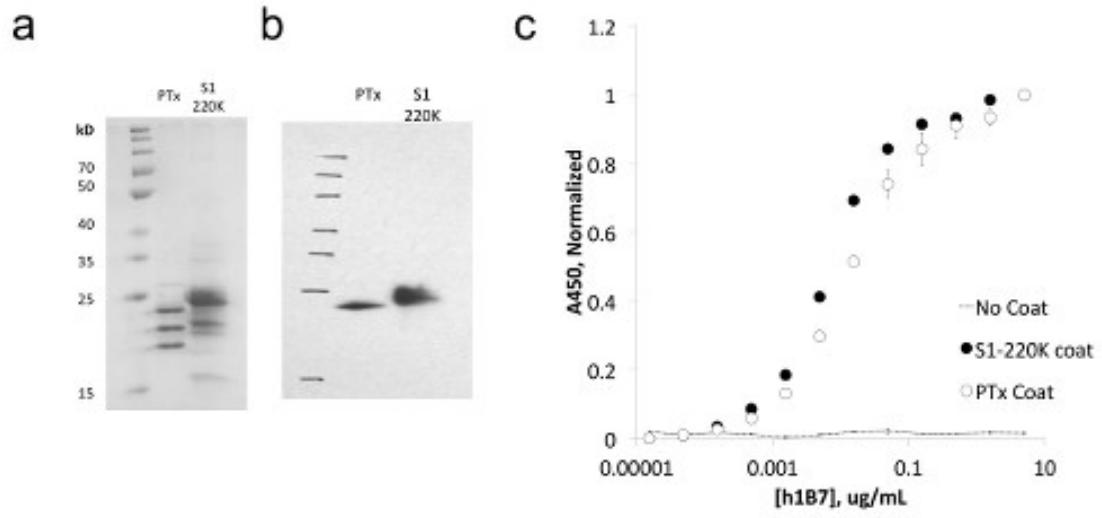



**Figure 20- A soluble version of the pertussis toxin S1 subunit can be expressed in *E. coli* and retains affinity for hu1B7.**

Truncated S1 in a pAK400 expression vector was produced in BL21(DE3), harvested by osmotic shock, and purified by immobilized metal affinity chromatography and size exclusion. a. SDS-PAGE of truncated S1 (S1-220K, 25.9 kDa) and full length PTx (26.1 kDa), b. Western blot of S1-220K and PTx, probed by hu1B7 and GαhFc-HRP, and c. ELISA of hu1B7 on a 4nM coat of PTx or S1-220K, detected by GαhFc-HRP.

Glu75, Gly78-His83, Ile85-Tyr87, Ala93, Tyr148, Asn150-Ile152, and Asn163. Mapping these residues onto the structure of the pertussis toxin (PTx)[143] Figure 21 shows that fourteen of sixteen residues are located in a spatially contiguous location. This proposed epitope is consistent with a previous alanine scanning dataset developed by Sutherland and Maynard[125]. The two conserved residues outside of this region, Ala93 and Asn163, are most likely of structural importance for the conformational epitope as they are buried in the protein core. The epitope surface is typical of antibody-antigen interactions with charged and aromatic residues along with hydrophobic patches (Figure 21).

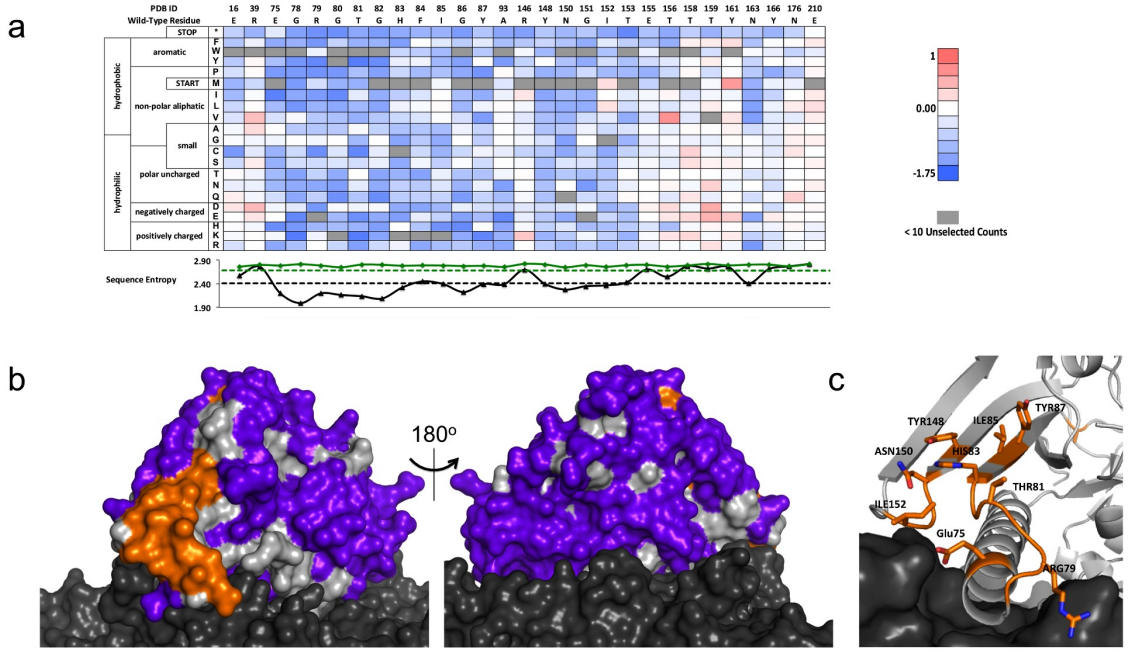



**Figure 21- PTxS1 Conformational Epitope Determination.**

a. A subset (29/220 residues) of the fitness-metric heat map for the PTxS1-hu1B7 interaction. Sequence entropy for the unselected/display population (green) and unselected/bound population (black) is plotted below with their respective cut-offs (dashed lines). b. Subtractive sequence entropy analysis for PTxS1-hu1B7interaction. The light grey surface represents the S1 subunit and the dark grey represents other subunits of PTx. Conserved residues (orange) are found on the S1 subunit proximal to the S5 and S6 subunits. Non-conserved residues (purple) are found over most of the solvent accessible surface area. c. Close up view of the conserved residues at the epitope interface. PTxS1 is represented with cartoon and sticks format, while the other subunits are represented as the dark grey surface.

In principle, relative dissociation constants ($K_{d,i}/K_{d,WT}$) for antibody-antigen interactions can be calculated directly from the digital counting[7] (see Theory section). However, our method relies on counting individual sequences after a single cell sort, resulting in a limited dynamic range. To determine whether dissociation constants calculated from deep sequencing results are quantitative, we compared our results with an alanine scanning dataset for *in vitro* binding of PTx-S1-220 and murine 1B7[125]. Our results are consistent with scanning data for 16 of the 17 mutations (Table 8). The one discrepancy, Arg39Ala, is not in spatial proximity of the highlighted epitope, potentially indicating different long-range interactions between PTx-S1-220 and the murine and humanized 1B7 antibodies used in the separate experiments. Consistent with the limited dynamic range of the deep sequencing method, the relative dissociation constants for hot spot residues Arg79, His83, Tyr148, and Asn150 are significantly underestimated in the deep sequencing datasets compared with *in vitro* measurements (Table 8). This limitation caused by digital counting a handful of sequences restricts the experimentally determined range of relative dissociation constants to 0.4-2.5 (Figure 22).

For further validation we tested four additional mutations identified from our deep mutational scanning datasets. PTxS1-220 variants T81K, T81H, I152M, and I152P were produced in *E. coli* and purified. A polyclonal anti-PTx antibody preparation was titrated against ELISA wells coated with 5 nM PTx S1 variants. Similar binding to all variants suggests that no variant has severe folding defects. Relative binding dissociation constants were calculated from observed $EC_{50}$ by titration of the hu1B7 antibody on an ELISA plate coated with 5 nM of the truncated PTx S1 or each variant (Table 8). *In vitro* binding for variants T81H, I152M, and I152P were quantitatively predicted by deep sequencing data. In contrast and consistent with the limited dynamic range highlighted above, the relative binding for binding knock-out variant T81K is

**Table 8- Comparison of fitness-metric based dissociation constant calculations with published experimentally determined dissociation constants.**
*Reported numbers are experimental relative EC50 values.

| | *in vitro* Binding | Deep Sequencing Data | |
|---|---|---|---|
| | $K_{d,i}/K_{d,wt}$ | $K_{d,i}/K_{d,wt}$ | Fitness Metric |
| WT | 1.00 ± 0.10 | 1.00 | 0 |
| R146A | 1.07 ± 0.10 | 1.08 ± 0.02 | -0.074 |
| E155A | 0.79 ± 0.09 | 0.97 ± 0.02 | 0.034 |
| T156A | 0.79 ± 0.09 | 1.04 ± 0.04 | -0.037 |
| T159A | 1.00 ± 0.16 | 1.06 ± 0.03 | -0.053 |
| Y161A | 0.93 ± 0.43 | 0.88 ± 0.04 | 0.116 |
| N176A | 0.79 ± 0.09 | 0.81 ± 0.03 | 0.199 |
| E210A | 1.14 ± 0.16 | 1.05 ± 0.02 | -0.052 |
| E16A | 1.14 ± 0.11 | 1.30 ± 0.08 | -0.266 |
| T81A | 1.21 ± 0.65 | 1.19 ± 0.03 | -0.168 |
| T158A | 0.93 ± 0.16 | 0.98 ± 0.03 | 0.020 |
| Y166A | 1.00 ± 0.16 | 1.10 ± 0.02 | -0.094 |
| R39A | 2.14 ± 0.32 | 0.84 ± 0.06 | 0.165 |
| T153A | 1.43 ± 0.30 | 1.48 ± 0.03 | -0.403 |
| R79A | 17.9 ± 3.1 | 2.62 ± 0.25 | -1.058 |
| H83A | 7.1 ± 1.5 | 2.14 ± 0.14 | -0.826 |
| Y148A | 20.7 ± 4.5 | 1.69 ± 0.16 | -0.546 |
| N150A | 5.7 ± 0.8 | 1.89 ± 0.08 | -0.670 |
| T81K | no binding* | 4.25 ± 0.88 | -1.664 |
| T81H | 1.51 ± 0.43* | 2.55 ± 0.24 | -1.025 |
| I152M | 0.68 ± 0.23* | 0.67 ± 0.11 | 0.359 |
| I152P | 1.36 ± 0.46* | 2.12 ± 0.06 | -0.806 |

significantly underestimated in the deep sequencing datasets. We conclude that while relative dissociation constants calculated directly from the deep sequencing data are consistent with *in vitro*

measurements, care must be exercised when calculating a quantitative energetic contribution.

Next, we asked whether the method could be used to map the conformational epitope of an antibody targeting tumor-associated calcium signal transducer 2 (TACST2, a.k.a. TROP2), a 323 amino acid, 36 kDa transmembrane glycoprotein. TROP2 is overexpressed in numerous human epithelial cancers[144] and identified as an oncogene in colon cancer, with metastatic and invasive abilities[137, 145]. Studies have linked Trop2 to increased tumor growth, since ectopic expression of Trop2 in cancer cell lines causes them to become highly tumourigenic when implanted in mice, whereas silencing Trop2 inhibits cell proliferation *in vitro*[146]. Furthermore, silencing Trop2 in breast cancer cell line MDA-MB-231 decreases migration as observed by transwell migration and wound-healing assays (Figure 23**).**

The extracellular portion of TROP2 (TROP2Ex) contains three domains: an N-terminal domain (ND), a middle TY domain, and a C-terminal domain (CD) (Figure 24). Like its close paralogue EpCAM (epithelial cell adhesion molecule), TROP2 is a nuclear signal transducer

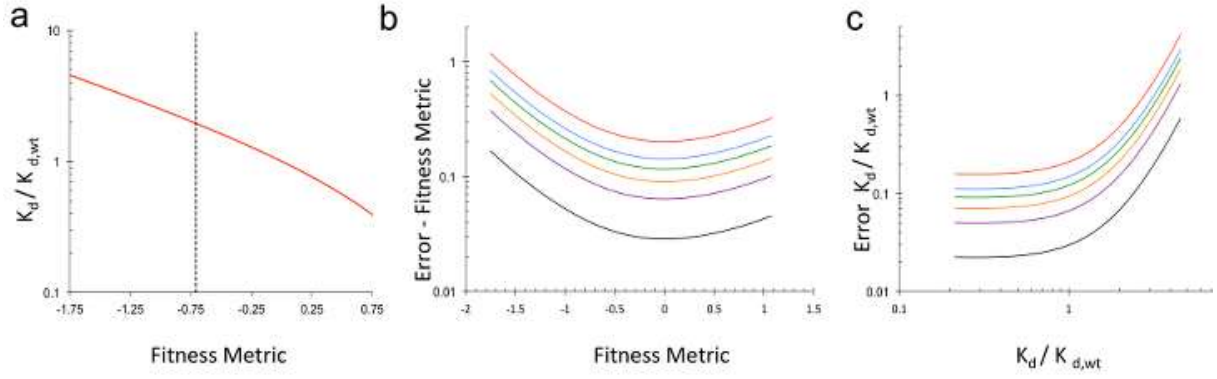



**Figure 22- Fitness metric and relative dissociation constant error.**

a. Relative dissociation constant as a function of fitness metrics. The vertical dashed line represents the average fitness metric for stop codon positions. b. Standard error as a function of fitness metric for different numbers of unselected counts (red, 10; blue, 20; green, 30; orange, 50; purple, 100; black, 500) Fitness metrics associated with lower number of counts have higher error. c. Relative dissociation constant error as a function of relative dissociation constant for different numbers of unselected counts. As the relative dissociation constant increases the amount of error increases.

activated by regulated intramembrane proteolysis (RIP)[147, 148]. TROP2Ex is first shed by proteolysis, followed by intramembrane cleavage to release intracellular TROP2 (TROP2IC). Because recombinant TROP2Ex forms a homodimer in solution[149], it has been speculated that destabilization of the extracellular region by local environmental changes or an agonist leads to RIP. However, neither the proteolytic cleavage sites nor potential agonist(s) have been unambiguously identified.

A recently developed mAb m7E6 (mouse7E6) targeting TROP2Ex inhibited tumor growth in the A431 xenograft model, and m7E6-drug conjugates induced long term regression in the BxPC3 xenograft model[130]. To investigate the structural basis of m7E6 efficacy, we prepared SSM libraries covering 95.8% of possible single non-synonymous mutations for the yeast-displayed TROP2Ex (residues Thr28-Thr274) and performed a single FACS sort against the biotinylated Fab of m7E6 at a labeling concentration of 22 nM. After the experimental workflow, the same sequence conservation analysis as above was used to determine residues contributing to the epitope. The subtractive sequence entropy measure completely removed most portions of TROP2Ex from consideration of the epitope, resulting in unambiguous determination of the binding footprint. In contrast to most previously described mAbs that bind at or near the N-terminal cysteine-rich domain, residues Asp171, Arg178, and the Ridge on CD (RCD) loop Gly241-Pro250 were identified as contributing to the m7e6- TROP2 interaction (Figure 24). This epitope is in agreement with m7e6 binding affinity results from domain swapping experiments[130]. Using a homology model of TROP2 guided by the structure of the paralogue epithelial cell adhesion molecule EpCAM[150], these residues map to a membrane-distal region opposite to the face on the CD domain that putatively makes specific inter-dimer contacts with the TY loop (Figure 24**).**

There are several reasons why targeting this epitope could be effective. For example, m7e6 could partially block the agonist(s) binding site on TROP2, preventing activation. Alternatively, m7e6 could prevent destabilization of the extracellular region by sterically blocking proteolytic cleavage or by preventing dimer dissociation after proteolytic cleavage. Since 4.1 nm separates the centers of the proposed epitope between the dimer subunits (Figure 24), an IgG could occupy both subunits of the dimer. In this scenario, we speculate that the antibody could prevent destabilization of the extracellular region or prevent an agonist site by the steric bulk of the IgG. As a test of this hypothesis, we reasoned that, owing to its monovalency, m7e6 Fab would not be as effective as the corresponding IgG in preventing dimer destabilization. We performed Boyden's chamber assays to investigate the influence of m7E6 IgG and Fab on migration rates in MDA-MB-231 cells. Whereas m7e6 IgG treatment inhibited migration ($p = 0.004$) (Figure 24), m7e6 Fab was unable to inhibit migration ($p=0.187$) at the highest tested concentration (40μg/ml) (Figure 24). We confirmed that both m7e6 Fab and IgG were able to label breast cancer cell line MDA-MB-231 (Figure 25). Additionally, we tested the influence of m7E6 IgG on inducing proliferation (metabolic activity) and cytotoxicity in MDA-MB-231 cells. We found that the treatment did not result in a statistically significant decrease in proliferation rates (One-way ANOVA $p=0.384$) or increase in cytotoxicity levels (One-way ANOVA $p=0.141$), indicating that the mechanism by which m7EG IgG resulted in reduced migration rates was IgG were able to label breast cancer cell line MDA-MB-231 (Figure 25). Independent of reduced proliferation or cell death (Figure 25).We further investigated the localization of the TROP2 intracellular domain in response to the IgG treatment using confocal microscopy. We found that nuclear expression levels of TROP2Ic were retained in both the IgG and Fab treated cells (Figure **26)** This scenario suggests that the effect of m7E6 binding to TROP2 on reducing migration rates may be mediated by blocking the agonist

binding site or by influencing the downstream signaling cascade. Whatever the exact mechanism behind m7e6 efficacy, the conformational epitope uncovered by the present method was used to predict that m7e6 Fab could not inhibit migration.

**4.6 Discussion**

The sequence-function mapping pipeline using a yeast-displayed antigen can be used to elucidate conformational, discontinuous epitopes of complex proteins. As demonstrated with TROP2, a solved structure of the antigen is not essential for identification of the conformational epitope. The methodological improvements developed in this paper allow us to complete the pipeline using a single cell sorter in 14 days for 24 different antibody antigen complexes at an approximate material and supply cost of $330 per antibody-antigen epitope. The cost and speed of this method offer significant advantages compared with competing display-based protocols. Notably, our method requires only a few micrograms of the starting antibody and so can be used directly downstream of immortalized B cell or hybridoma screening. Additionally, the ability to comprehensively map sequence determinants to binding may help elucidate potential escape mutants and be used to predict whether the antibody will maintain affinity for antigen homologs from model organisms. The sequence-function maps may also be integrated into computational prediction software to improve the predictions of specific antibody-antigen structural contacts at the atomic level or improve computational predictions of individual mutations on protein-protein interactions[151].

There are minor limitations in the current protocol. For example, antigens requiring multiple subunits to fold may be difficult to express on the yeast surface. Additionally, conformational epitopes requiring heterogenous peptide-glycosyl surfaces will not be able to be mapped. Nevertheless, our results show that antibody-binding surfaces for complicated homodimer and

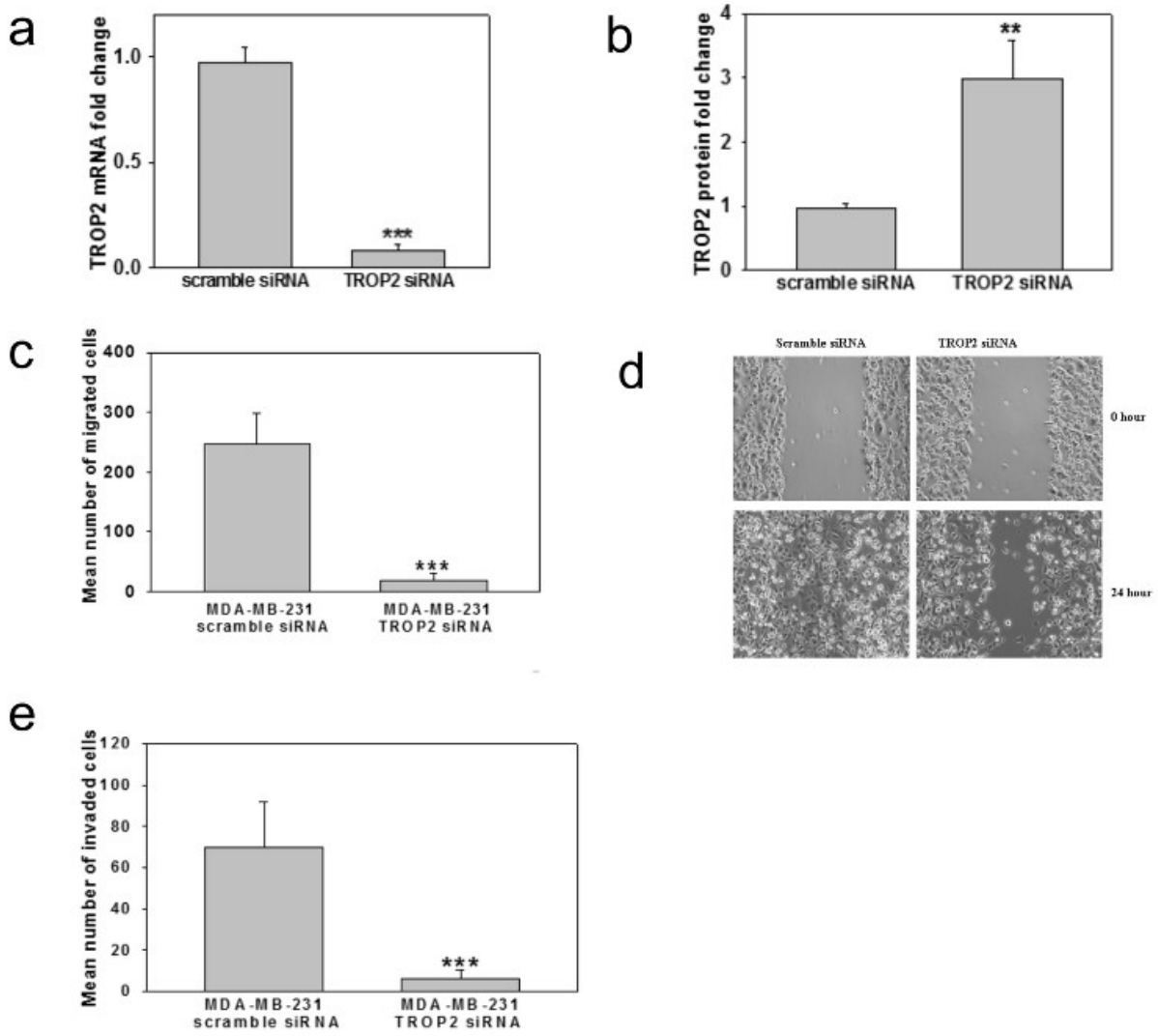



**Figure 23- TROP2 expression in MDA-MB-231 cells enhances migration and invasion.**

a,b. MDA-MB-231 cells were transfected with scramble siRNA or siRNA targeting TROP2 for 24hr. The (a.) mRNA and (b.) protein levels of TROP2 were measured by real-time PCR and Western blotting. n = 3. **: $p < 0.01$ and ***: $p<0.001$vs. control. c. Suppression of MDA-MB-231 migration by silencing TROP2. MDA-MB-231 cells were transfected with scramble siRNA or siRNA targeting TROP2, and then subjected to a transwell migration assay. The cells were allowed to migrate toward serum for 8 h. Triplicate wells were used for each condition in three independent experiments. ***: $p<0.001$ vs. control. d. Silencing TROP2 reduced the migration of highly invasive breast cancer cells. MDA-MB-231 cells were transfected with scramble siRNA or TROP2 siRNA and subjected to a wound-healing assay. Representative photographs at the indicated time points from three independent experiments, with each performed in triplicate wells. Magnification: 10X. e. Suppression of MDA-MB-231 invasion by silencing TROP2. MDA-MB-231 cells were transfected with scramble siRNA or siRNA targeting TROP2, and then seeded in a matrigel-coated Boyden chamber and subjected to a transwell invasion assay. The cells were allowed to migrate toward serum for 20 h. Triplicate wells were used for each condition in three independent experiments. ***: $p<0.001$ vs. control.

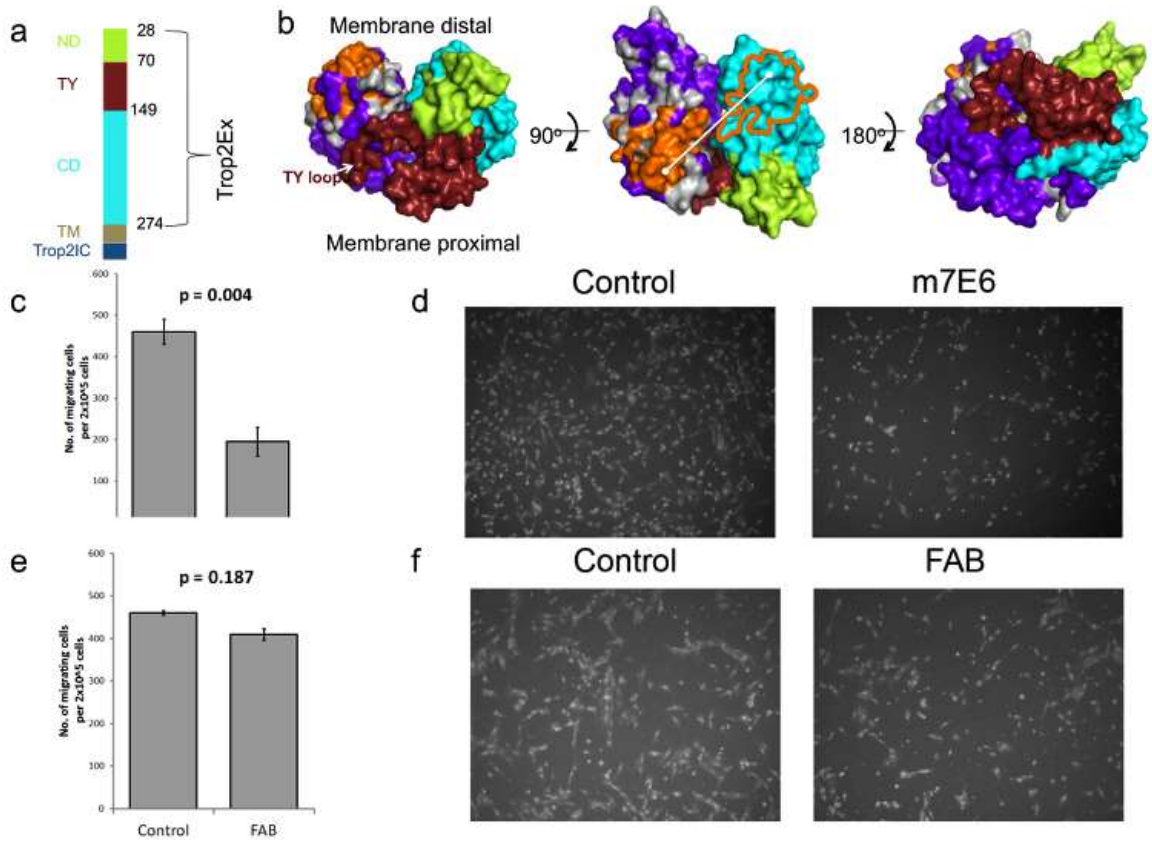



**Figure 24- TROP2 Conformational Epitope Determination.**

a. The domains of membrane protein TROP2. The extracellular portion of TROP2 (TROP2Ex) contains an N-terminal domain (ND, green), TY domain (brick red), and C-terminal domain (CD, cyan). TM indicates the membrane spanning portion, and Trop2Ic is the intracellular domain. b. Homology model of TROP2Ex homodimer shown in surface view. One subunit is colored by sequence entropy based on m7e6 binding (orange – conserved, purple – non-conserved, gray – intermediate). The other subunit is colored by domain using the same coloring scheme as in panel a. The fine epitope is located on the membrane distal face of the CD. The center of the epitope on one subunit is separated by 4.1 nm from the epitope on the adjacent subunit. c. Bar graphs showing the number of migrating cells in MDA-MB-231 cells treated with PBS (control) or m7EG IgG d. Representative bright field images of Boyden's chamber inserts showing cells that migrated across the 8μm membrane after 24 hours of treatment. e. Bar graphs showing number migrating cells with control or m7E6 Fab treatment. f. Representative bright field images showing migrating cells across the membrane. P-values indicate significance of difference in mean (n=3) determined using Student's two-tailed t-tests.

proteins can be mapped. While our approach can be used as is to interrogate antibody panels of 10-50 members, further improvements in speed and cost must be addressed for integration of the method with other high-throughput or single cell technologies. Methodological advances should

focus on replacing the bottleneck FACS step with a more high-throughput sorting technique and removing the need to prepare multiple libraries for each antigen.

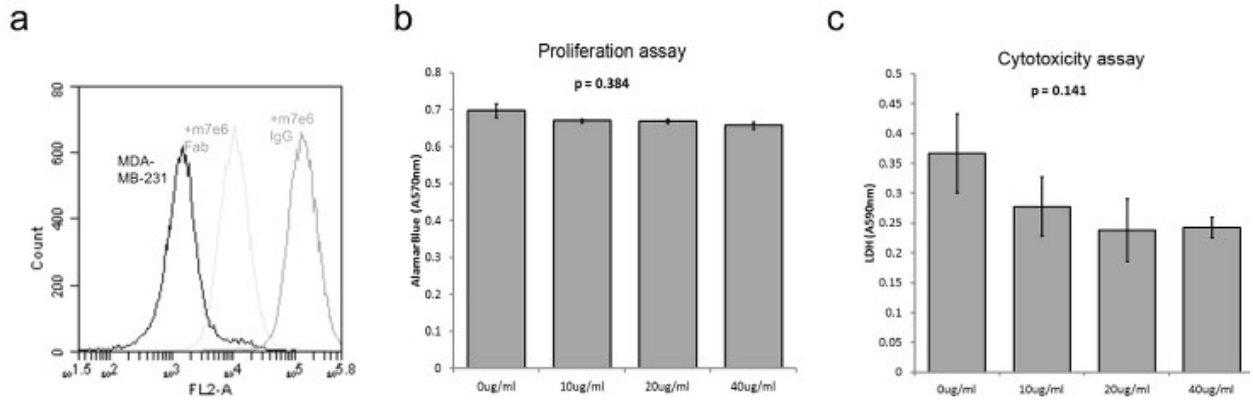



**Figure 25- TROP2 Characterization.**

a. $10^5$ MDA-MB-231 cells were labeled with 50 nM biotinylated m7e6 IgG (orange), 50 nM biotinylated m7e6 Fab (cyan), or nude (blue) in buffer PBSF for 30 min at room temperature. After washing, cells were secondarily labeled with streptavidin-phycoerythrin and processed by flow cytometry. b. Bar graphs indicating the average (n=3) AlamarBlue assay absorbance (570nm) in cells treated with PBS (control) or increasing concentration of m7E6 IgG (10 to 40 μg/ml). c. Bar graphs indicating the average (n=3) LDH absorbance (590nm) in control and m7EG IgG treated cells. P-values indicate the significance levels of the influence of increasing concentrations of the IgG treatment on proliferation or cytotoxicity levels determined using one-way ANOVA.

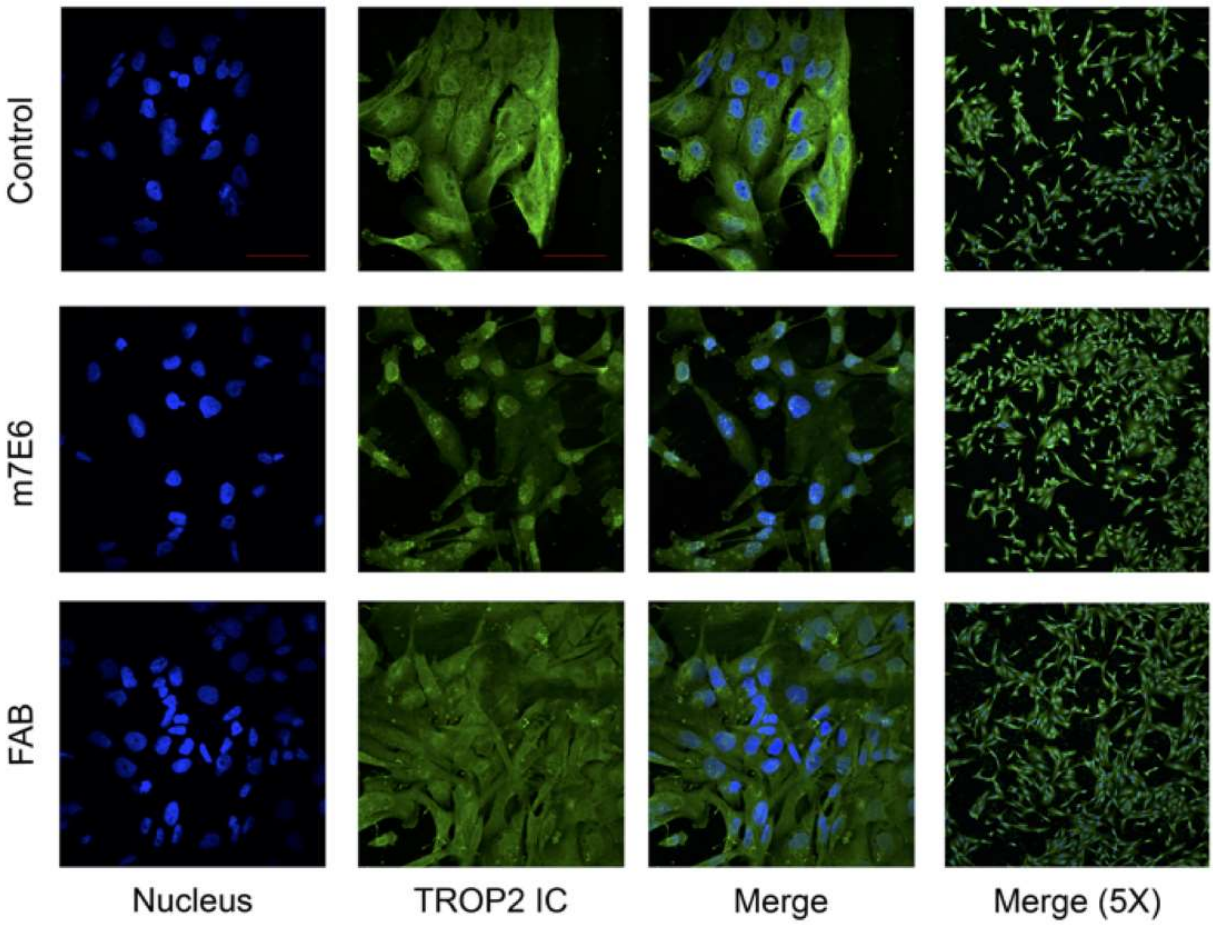



**Figure 26- Confocal images showing expression and localization of TROP2 intracellular domain.**

Confocal images showing expression and localization of TROP2 intracellular domain represented by green fluorescence in MDA-MB-231 cells treated with PBS (control), m7E6 IgG or Fab with nuclear counter-staining indicated by blue fluorescence. The individual panels were recorded at 60X magnification (scale bar = 50µm) with identical image acquisition parameters between different conditions.

# CHAPTER 5

## 5. Conformational epitope mapping of pertussis toxin antibodies and future directions

### 5.1 Introduction

Knowledge of a fine conformational epitope can provide a basis for structural vaccine design. Structural vaccine design uses protein structure to design immunogens and has promise to provide new vaccines against traditionally difficult targets[152]. One general approach of structural vaccine design is to design an immunogen that contains a stabilized neutralizing epitope of interest that may also lack any immunodominant non-protective or undesirable epitopes. Approaches like this has been used for Lyme disease[153], influenza[154, 155] and respiratory syncytial virus (RSV)[24]. In the latter case, McLellan et al.[24] designed over 150 immunogen variants containing a key neutralizing epitope, thus yielding stabilized versions of the RSV glycoprotein that maintained the antigenic site at extreme temperatures and pH. These stabilized versions were assessed for their ability to elicit protective responses and showed neutralizing activity. In this example the crystal structure used to identify the epitope of a neutralizing antibody with the RSV glycoprotein provided the basis for the rational design of a stabilized immunogen.

Whooping cough infections, caused by the bacteria *Bordella pertussis*, continue to increase in incidence in the US and other industrialized countries, despite widespread vaccination[141]. According to the CDC, pertussis cases in the United States currently persist at levels 13-fold higher than in the 1970s[156]. A more immunogenic and effective pertussis vaccine is urgently needed. We recently identified the neutralizing epitope for monoclonal antibody hu1B7[88]. Other neutralizing and non-neutralizing antibodies have been isolated from hybridomas and humanized or from vaccinated patients[142]. Nguyen et al.[157] recently identified a cocktail of humanized anti-pertussis antibodies that showed efficacy when administered as a prophylactic. Using additional neutralizing

antibodies, we seek to identify a larger neutralizing epitope on the pertussis toxin that could be used as a basis for designing a better vaccine against the pertussis toxin.

**5.2 Materials and Methods**

Conformational epitope mapping closely followed the protocol described in Kowalsky et. al.[48, 88]. Anti-PTxS1 IgGs were chemically biotinylated at a molar ratio of 20:1 Biotin to IgG using the EZ link NHS-biotin kit following the manufacturers instructions (Life Technologies, Carlsbad, CA). Observed equilibrium dissociation constants ($K_{D,obs}$) of each IgG to yeast displayed PTxS1-220 (pETCON-PTxS1-220) were determined using clonal population yeast display titrations according to Chao et al[63]. IgG concentrations tested ranged from 8 pM to 600 nM.

Yeast display sorts were conducted by fluorescence activated cell sorting on an BD Influx equipment (BD Biosciences, Franklin Lakes, NJ) using previously described PTxS1-220 site saturation mutagenesis libraries[88] at an IgG labeling concentration of half $K_{D,obs}$. 400,000 events were collected for each population (Table 9). Yeast plasmid DNA was prepared for deep sequencing following the protocol in Kowalsky et al[48]. Library DNA was sequenced on an Illumina MiSeq using a 250x2 Illumina MiSeq kit (Illumina, San Diego, CA) at the Michigan State University Sequencing Core. Sequencing data was analyzed following the procedure described in Kowalsky et al[48]. Because IgGs were used instead of Fabs as previously described significantly conserved residues cutoffs were benchmarked against a hu1B7 data set. For the IgG data significantly conserved residues were determined using a cutoff value of:

$$cutoff = \frac{SE_{\max} + SE_{\min}}{2} + 0.25\left(SE_{\max} - SE_{\min}\right) \qquad (1)$$

In addition to the cutoff value significantly conserved residues also had a solvent accessible surface area greater then 20%[158].

**Table 9- Sorting Statistics**

| | Tile Length (AA) | Sort Labeling Conditions (pM) | Events Collected for Binding Population | Percent Sorted (Display) | Percent Sorted (Binding) |
|---|---|---|---|---|---|
| **PTxS1-A8 Tile** | 72 | 64 | 400,000 | 63.02% | 7.56% |
| **PTxS1-A8 Tile** | 74 | 64 | 400,000 | 66.89% | 7.27% |
| **PTxS1-A8 Tile** | 73 | 64 | 400,000 | 63.84% | 7.63% |
| **PTxS1-A12 Tile** | 72 | 150,000 | 400,000 | 60.30% | 7.13% |
| **PTxS1-A12 Tile** | 74 | 150,000 | 400,000 | 64.58% | 7.97% |
| **PTxS1-A12 Tile** | 73 | 150,000 | 400,000 | 63.48% | 7.66% |
| **PTxS1-E12 Tile** | 72 | 140.5 | 400,000 | 57.58% | 6.63% |
| **PTxS1-E12 Tile** | 74 | 140.5 | 400,000 | 62.27% | 7.57% |
| **PTxS1-E12 Tile** | 73 | 140.5 | 400,000 | 63.11% | 7.27% |

### 5.3 Results

Antibodies A8 and E12 were obtained from the Maynard lab at the University of Texas, Austin. Following Kowalsky et. al.,[48, 88] we determined the conformational epitope for A8 and E12 against yeast displayed Pertussis Toxin subunit 1 (PTxS1). Using chemically biotinylated IgG we determined the observed dissociation constant for each interaction using yeast display clonal titrations[63]. We performed a single FACS sort on the libraries collecting 400,000 cells in each of the three populations. Positional sequence entropy was used to determine the most conserved residues at the interface using a relaxed criteria owing to use of IgG.

In previous epitope mapping experiments[88], IgGs were digested into Fabs to maintain a 1:1 antibody to antigen binding ratio. We were interested in seeing if we could eliminate the Fab

preparation step and use full IgGs instead of Fabs. To facilitate this transition we benchmarked the epitope cutoff against the previous conformational epitope for PTxS1 and hu1B7. We performed this epitope mapping experiment using hu1B7 IgG for the residues 74-147 of PTxS1 and compared this to our Fab based experiments previously performed[88]. Using the same cutoffs as before we are only able to identify 3 of 11 epitope residues (Figure 27). We next asked whether there was a better cutoff that we could use to capture more of the epitope residues. To do this we increased the sequence entropy cutoff by adding $0.25(SE_{max}\text{-}SE_{min})$ to the sequence entropy midpoint this new cutoff increase correctly captured 9 of 11 the epitope residues but also introduced 2 additional false epitope residues. To better capture the epitope residues we only considered residues that had solvent accessibility greater then 20% giving us 6 of 7 correct residues with no additional false epitope residues.

Using this new cutoff for epitope determination, the method identified seven residues at the antibody-antigen interface for A8: Gly27, Arg58, Arg76, and Thr153-Thr156. Five of these

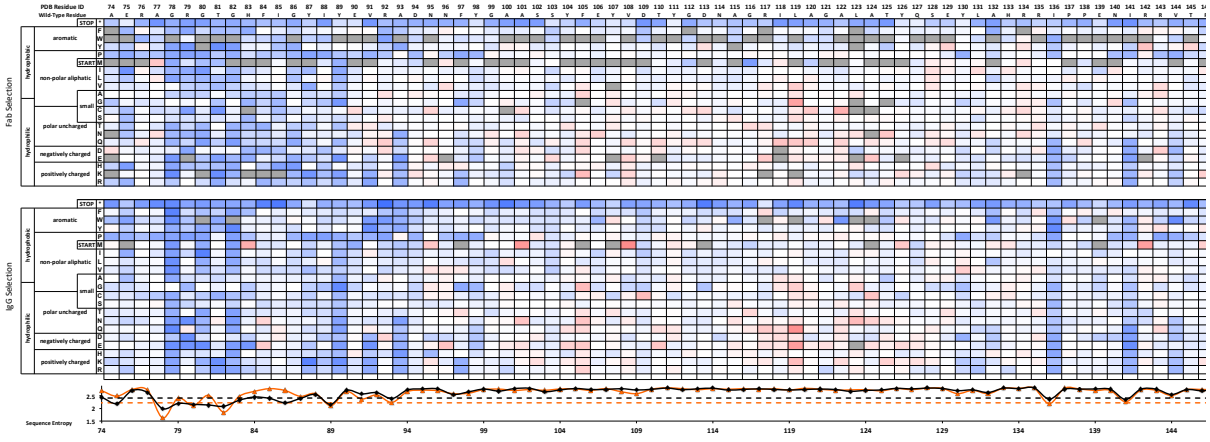



**Figure 27- Sequence-function heatmaps for PTxS1 and hu1B7 sorts with Fab and IgG**.

Heat maps are compared for epitope mapping experiments for a Fab (top) and IgG(bottom). Sequence entropy is plotted for Fab (black) and IgG (orange) with their respective cutoffs. Sorting with an IgG is less sensitive then when using a Fab. A new cutoff was introduced for the IgG sorts to accurately identify the epitope residues identified using a Fab.

six residues map to a contiguous location on the PTxS1 structure (Figure 28a) except for Gly27. The method also identified eight residues at the interface for E12 that map to a contiguous location: Thr153-Thr156, Thr158-Thr159, and Thr171 (Figure 28b). Using these residues sets we see that A8 and E12 target an overlapping but not identical epitope. These epitopes are located at a contiguous location adjacent to but distinct from the previously determined epitope for hu1B7 (Figure 28c).

### 5.4 Discussion and Future Directions

Here we present multiple neutralizing epitopes from a panel of antibodies against the pertussis toxin subunit1. By screening a panel of neutralizing antibodies we were able to identify a larger neutralizing epitope that could be used as a basis for vaccine design. Epitope mapping can be extended to other health crises such as Dengue fever[159] and Ebola[160] to determine neutralizing epitopes from isolated neutralizing antibodies.

In this project we used IgGs instead of Fabs to identify the conformational epitope. By comparison to a previously developed epitope map for huB17 Fab, huB17 IgG results in less potent discrimination of the conformational epitope. Still, we were able to determine contiguous epitope residues for neutralizing antibodies A8 and E12. Nevertheless, in future experiments it is

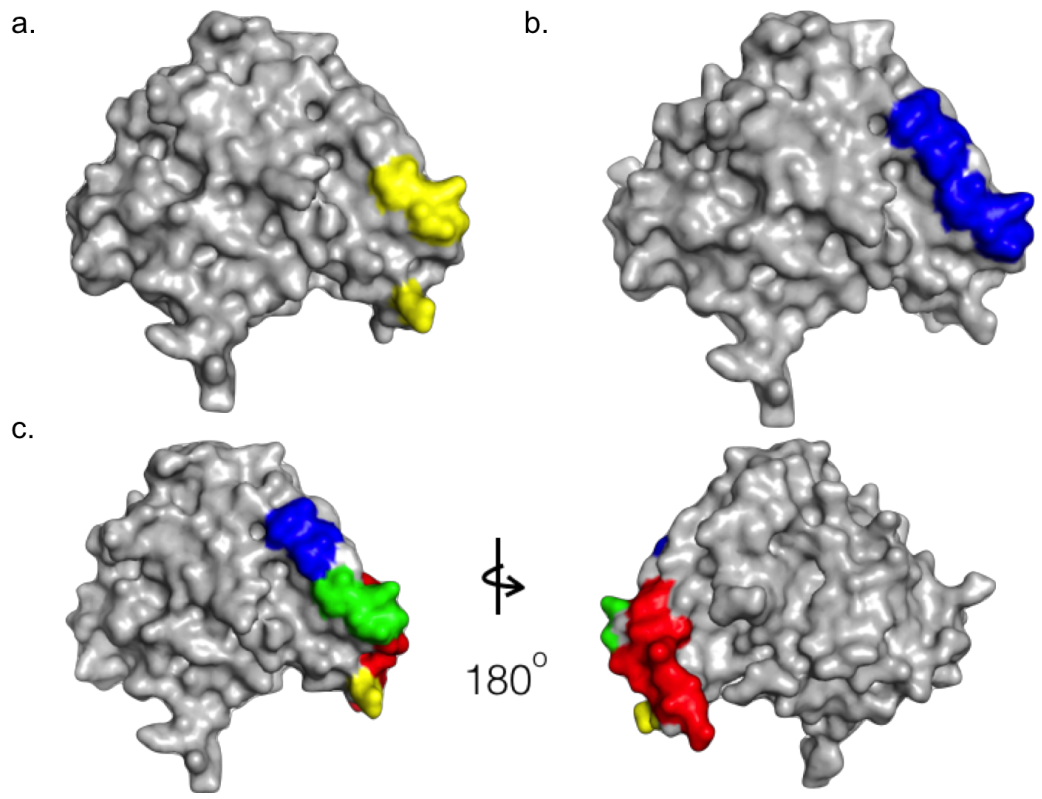



**Figure 28- Experimentally determined conformational epitopes for A8 and E12.**

a. Experimentally determined conformational epitope for A8. Significantly conserved epitope residues are indicated in orange and non-significantly conserved residues are colored purple. b. Experimentally determined conformational epitope for E12 colored like A. c. Rotated views of PTxS1 structure with hu1B7 (red), A8 (yellow), E12 (blue), and overlapping A8 and E12 (green) epitope residues indicated.

recommended to prepare Fabs instead of IgGs. Utilization of Fabs provides a higher resolution epitope map as well as maintains a 1:1 binding ratio to utilize binding affinity reconstruction from deep sequencing data.

In addition to the epitope mapping and antibody panel mapping applications that has been presented for the experimental pipeline, there are some additional applications. The wealth of mutational data obtained in the current experimental pipeline can help to identify the specificity of an antibody. Important to the specificity of an antibody is whether or not the antibody would bind to any paralogs or homologs of the antigen and also help to identify escape mutants. Currently,

only the antigen side of the interaction is mapped. By displaying an scfv or antibody on the surface of yeast instead of an antigen and using a soluble biotinylated version of the antigen, the method can be expanded to paratope mapping applications.

Because of throughput issues, high-resolution epitope mapping is introduced after lead candidate generation during the antibody discovery process. The ability to obtain epitope maps for early stage panels of antibodies would both ensure large epitope diversity in preliminary antibody panels as well as eliminate any antibodies against known protected epitopes. In addition to these considerations, epitope mapping a panel of antibodies would give an in-depth look at the immune response if isolating the antibodies directly from hybridomas or immortalized B cells. As the technology currently stands, epitope maps can be obtained in three to five weeks for less than $1500 for the first epitope map and about $500 for each additional epitope map (Table 10 and Note S6, Appendix I, for detailed break down). The majority of the time is spent in the set up of the yeast display and preparation of the mutational library and each of these steps is parallelizable. The bottleneck of the method therefore rests in the sorting technique. Unless multiple FACS can be used in parallel, which is not practical, the method is limited to 3 antibody/antigen pairs per day. While this allows the method to be scalable to tens of antibody/antigen interactions, we are unable to analyze hundreds or thousands of interactions on a high-throughput time scale. The introduction of a different selection method would allow the technology to compliments the recent advances in single sorting B-cell technology in antibody discovery platforms[120, 161].

The introduction of magnetic-bead assisted cell sorting (MACS) could help to parallelize the sorting methods. MACS is most frequently used in phage-display to sort strong binders from large libraries through many sorting steps[162]. MACS is often used because it is scalable to large

**Table 10- Cost and time for epitope mapping.**

Cost and time is determined for an antigen of 240 residues that would require three tiles. Cost does not include labor or the cost of the antigen gene.

| Step | Cost | Time |
|---|---|---|
| **Preparation yeast displayed mutational library** | $870 ($290/tile) | 2-4 weeks |
| **Sorting libraries using FACS** | $200 ($66/tile) | 1 day (2-3 interactions per day) |
| **Sequencing preparation** | $42 ($14/tile) | 2 days |
| **Sequencing and data analysis** | $42 ($14/tile) | 2 days (+2-4weeks wait time for sequencing) |
| **TOTAL** | **$1320** | **3-5 weeks** (+sequencing wait time) |

library diversity but it lacks the control over the sorted population that FACS affords. Sorting yeast-displayed libraries have using magnetic beads has previously been demonstrated[63, 163] and we hypothesize that it could substitute for the current sorting method. Effectively this would reduce the time required to sort libraries by a factor of 10 and also reduce the cost and "barrier of entry" to the method by eliminating the need for a cell sorter.

In order to implement magnetic beads as a sorting technique, many control experiments are needed. First, we will need to determine the capture probability of the yeast-displayed proteins as a function of labeling concentration using magnetic beads. Using this relation we can derive equations similar to those previously used to describe growth and FACS based selections to determine optimal sorting conditions and relations between the enrichment ratio obtained from deep sequencing and protein-protein binding affinity. It is likely that this equation is represented more closely by a stepwise function with a narrow dynamic range as opposed to a sigmoidal function that has a larger dynamic range. In this case we would need to determine what the correct

labeling concentration is such that there is discrimination between epitope and nonepitope residues. To determine this labeling concentration we can use our FACS sorted libraries as a benchmark set.

In addition to increasing the throughput of the method, we would like to generate a bound structure of the antibody/antigen interaction by incorporating the dataset into computational protein modeling software. Incorporating this capability into the process would add value to the data set. Ideally we would be able to predict the docking trajectory of the antibody/antigen complex given an antibody sequence, mutational dataset and the structure of the unbound antigen. An accurate model of the bound antibody/antigen structure would improve and aid in both therapeutic design and antibody affinity maturation.

To do this we would use the biomolecular protein modeling suite Rosetta[164, 165]. The RosettaDock[166, 167] application takes two unbound structures and predicts their bound structure. RosettaAntibody[168] is an application that predicts the structure of an antibody given the sequence. Combining these two applications we hope to obtain a properly folded antibody structure in the presence of the antigen. However, computational protein structure prediction requires sufficiently sampling the conformational space[166]. This often requires the creation of hundreds of thousands of docking trajectories leaving the researcher many predictions to sort through in order to find the correct one. We rely on the assumption that a lower energy gives a more sTable protein fold and therefore is the more likely conformation. To guide the sampling and providing a scoring metric, Rosetta includes an energy function that combines many energetic terms such as Van der Waals forces, electrostatic forces and hydrogen bonding energies among others[164].

The mutational data sets can be incorporated into the computational structure prediction. First, the mutational data set can be used to restrict or decrease the conformational space that needs

to be sampled. Ambiguous interface restraints[169, 170] can be introduced into Rosetta as a means of restricting the interface of the antibody and antigen to identified epitope residues. By restricting the interface to epitope residues the conformational space can be sampled more efficiently. The restraints would also be introduced into the scoring function to penalize docking trajectories far from the defined interface. The second way that the data sets can be used is as a metric to filter low energy structures outputted from the program. In this case careful consideration should be taken when considering which mutations will be sampled as Rosetta and similar protein modeling software can more accurately predict hydrophobic energies over electrostatic interactions. Computational methods should be benchmarked against existing benchmark sets for protein-protein interactions[171] and experimental mutational data obtained previously[85, 88]. For this application to be fully scalable, it would need to be efficient and accurate, producing of docked structures for dozens of antibody/antigen interactions in a single day.

The introduction of a parallel sorting technique to the conformational epitope mapping technique and improved computational antibody/antigen docking trajectories would further the field of structural vaccine design. The techniques would most importantly help to rapidly identify neutralizing epitopes from isolated neutralizing antibodies but could also be used to help identify mutations that remove immunodominant epitopes. We anticipate using these techniques to identify more develop sTable immunogens for vaccines.

# APPENDICES

# APPENDIX A

## Supplementary Notes

**Supplementary Note S1:** Practical Considerations for High-Resolution Sequence-Function Mapping.

**Target Selection**

*INPUT: Target protein*

*OUTPUT: Selection conditions, Protein gene sequence*

At this step the most important consideration is that, for a given protein target, there are conditions under which more active variants can be selected over less active variants. In this paper we explicitly consider growth-based selections and fluorescence activated cell sorting (FACS) screening techniques.

For growth-based selections, a common strategy is to identify conditions where growth depends on flux through a given pathway. For example, bacterial growth in the presence of beta-lactam antibiotics requires a certain enzymatic flux of beta-lactamase. To allow determination of more active proteins than the starting sequence, conditions should be screened where the starting sequence does not support maximal growth in the host background. On the other hand, the growth rate should be significant enough that the selection experiment can be completed in 1-2 days.

FACS, in conjunction with cell display techniques like yeast-surface display, can be used to evaluate protein-protein binders, protein-small molecule binders, and even certain enzymes. In all cases, higher fluorescence (after labeling) is correlated with more potent activity. Conditions should be screened such that the starting sequence does not support maximum fluorescence. In the case of protein binders, one key parameter to vary is the labeling concentration relative to the dissociation constant of the interaction for the starting sequence. At labeling concentrations well above the dissociation constant, increases in binding affinity result in minimal fluorescence

increases. The set up and validation of the selection system is key to the efficiency and results of the experiment. A poorly set up selection will yield poor sequence-function maps.

**Gene Tiling**

*INPUT: Protein gene sequence, sequencing read length*

*OUTPUT: Gene tiling scheme, "inner" primer sequences*

The gene sequence needs to be segmented into tiles for library preparation and selections to allow efficient mutation counting with Illumina sequencing reads. Genes are tiled in segments of lengths that are multiples of three and at least 20 base pairs shorter than the sequencing reads. Tiles are no longer than 126 bp for 150 bp paired-end (150bp PE) reads. Each section should be at a length of a multiple of three (i.e., 117, 120, 123, etc.) as to not split an amino acid codon over multiple sections. We have written a custom script **(Script S2;** GeneTiles.m**)** to facilitate tiling and design of the inner primers. The input into the program is the gene sequence (including 33 bp up and downstream of the gene sequence) and the read length for the next generation sequencing. The output from this program is spatial location of individual gene tiles, and the set of inner primers needed for sequencing preparation.

To use GeneTile.m:

1. Create a file named "genesequence.txt" in FASTA format with your gene sequence. Include exactly 33 bp upstream and downstream of your genesequence in this file.
2. Load "GeneTile.m" and "genesequence.txt" into MATLAB and run the function GeneTile.m
3. The output file containing the spatial location of the gene tiles and the inner primer sequences.

**Library Preparation**

*INPUT: Pfunkel Method, QuikChange Primer Sequences, Gene tiling scheme*

*OUTPUT: Libraries for selection*

SSM libraries should be prepared using the Pfunkel Method essentially as described [51]. As stated in the results section, we recommend using the Agilent QuikChange Primer Designer (http://www.genomics.agilent.com/primerDesignProgram.jsp?_DARGS=/primerDesignProgram.jsp). Although it is much more labor intensive up-front, in our hands the longer lengths of the primers allow better library coverage than the primer design script provided by Firnberg *et al.* [51]. It is important to prepare the library of each tile separately, as each tile needs to be selected independently. Library coverage statistics can be calculated following Boder and Wittrup [75]; the number of transformants from each tiled library should exceed the theoretical library size by at least 7-fold (>99.9% coverage). The library for each tile is plated onto a separate Bioassay plate, and then plasmid DNA is recovered by Qiagen Midiprep.

This library plasmid DNA can then be transformed into the strain used for selection. In the specific case of *S. cerevisiae*, high efficiency transformations can be done following Benatuil et.al. [64]. The high number of transformants obtained from this method almost always supports the degeneracy of SSM libraries. For both *S. cerevisiae* and *E. coli*- based libraries, cell stocks can be stored at -80°C until used for selections.

Double transformation artifacts do not need to be considered if transforming a low copy number plasmid (like in yeast surface display). Double transformation artifacts do need to be considered when transforming with medium or high copy number plasmids. Under the model developed in this paper, no correction needs to be made for growth-based selections using 10 or less average population doublings at double transformation percentages less than 10%. Our recommendation is that for growth-based selections you vary the transformation parameters (cell

density, amount of plasmid DNA, etc.) until the transformation efficiency can support the degeneracy of the library and the percentage of double transformants is less than 10%.

**Selections**

*INPUT: Libraries, Selection Conditions*

*OUTPUT: Selected Libraries*

Selections should be performed following the experimental conditions explained in the **Results** section. We recommend growth-based selections be run for 6-8 average population doublings. Growth-based selections should be performed on exponentially growing cells only (avoid lag phases and post-exponential growth phases). The initial inoculum density should be much larger than the number of variants in the library population. Following selection, cells can be stored in glycerol stocks at -80$^{o}$C before library DNA is prepared for deep sequencing. For FACS, collecting the top 5% of the population using a square gate (one color sorting). To generate good counting statistics, at least 100x of the theoretical library size should be collected. For a library size of 2500, this means 250,000 cells collected (5,000,000 cells sorted). In the specific case of yeast-surface display, cells should be labeled with biotinylated antigen at 50% of the dissociation constant for the wild-type interaction. Following sorting, yeast should be recovered in selective media and stored in 20 mM HEPES 150 mM NaCl pH 7.5, 20% (w/v) glycerol in 1x10$^{7}$ aliquots at −80 °C until they are prepared for deep sequencing.

**Deep Sequencing Preparation**

*INPUT: Selected Libraries, Unselected Libraries, Inner/Outer Primers*

*OUTPUT: Libraries ready for next generation sequencing*

For next-generation sequencing we utilized the Illumina MiSeq 150-bp PE reads. This technology has recently been expanded to 300 bp pe reads, as using longer reads reduces the number of

libraries and selections. Inner primers are designed using the gene tiling script (**Script S2)**. The inner primers follow the pattern:

FWD: 5'- GTTCAGAGTTCTACAGTCCGACGATC<SEGMENT OVERLAP>-3'

REV: 5'- CCTTGGCACCCGAGAATTCCA<SEGEMENT OVERLAP REV. COMP.>-3'

Outer primers are taken from the Illumina TrueSeq Small RNA kit and can be ordered using the sequences listed in **Table S2.** The outer primers are meant to be universal: while the same forward primer is used for each library member, the reverse primers contain barcodes for multiplexing. Each library (sorted and unsorted) should use its own barcode to allow demultiplexing of sequencing reads. To append the primers, DNA is first extracted from the stored libraries. Plasmid DNA can be extracted from *E. coli* using a Qiagen miniprep, while for *S. cerevisiae* we use by a modified smash and grab protocol (**Note S4).** Primers are attached to the gene tile using PCR method A found in **Table S3.** PCR reactions are purified using AMPure beads and quantified using Quant-it PicoGreen dsDNA Assay on a plate reader. Libraries are mixed in equimolar amounts to ensure even sequencing before being delivered for sequencing.

**Data Analysis**

*INPUT: Sequencing reads*

*OUTPUT: Normalized fitness metrics for the sequence-function landscape*

A modified Enrich-0.2 program is used to interpret and analyze the sequencing data [53]. Modifications to the Enrich-0.2 program were made according to **Script S1**. Currently Enrich-0.2 will only translate proteins from the first base pair of the dna sequence. Using the gene-tiling method the protein sequence is not always in the same frame as the first base pair. The

modifications to Enrich-0.2 allows for this capability. Using a new <TRANSLATE_START> command into the config file where you indicate the beginning of translation of the DNA to the protein sequence. We also found it necessary to edit the length of the initial matching protein sequence to 10 bp over the standard 20 found in the Enrich program since the flanking sequences to the tiles are not usually longer than 20 base pairs. The script for the enrich patch is called EnrichPatch.py to use the patch, move it into the main enrich directory and use the command line "python EnrichPatch.py" this should modify Enrich for your purposes. To ensure proper execution of the patch, the 'example_local_config' file should contain a new <translate_start> command line in the read_aligner section.

Enrich outputs "unlink_wild_counts" for both the selected and unselected populations, at both the dna and protein levels. These files are populated frequency matrices the rows being residue position, relative to the beginning of translation for that section, and the columns are the different amino acid residues. The frequency of each variant in the libraries are listed. These files can be used to determine the $\log_2$ enrichment ratio for each variant. Apart from the frequency matrices it is important to extract the wild-type enrichment ratio for each gene tile, from the "ratios_sel_PRO" file for the NA_NA variant to use in the normalization equations. The normalization equations, for the fitness metric, that should be use are:

$$\zeta_i = \log_2\left(\frac{\frac{\varepsilon_i}{g_p}+1}{\frac{\varepsilon_{wt}}{g_p}+1}\right) \tag{11}$$

For growth-based selections and,

$$\zeta_i = \log_2\left[e\sqrt{2}\sigma'\left(erf^{-1}\left(1-\phi 2^{\varepsilon_{wt}+1}\right)-erf^{-1}\left(1-\phi 2^{\varepsilon_i+1}\right)\right)\right] \tag{20}$$

for FACS selections. Note the need for the enrichment ratio of both the variant *i* and the wild-type along with the number of population doubling times for the growth-based selections and the log transformed standard deviation for the FACS based selections. We use custom scripts to do the normalization of the data and import the normalized fitness metrics into Excel in order to visualize the fitness landscapes using the conditional formatting options.

# APPENDIX B

## Supplementary Figures

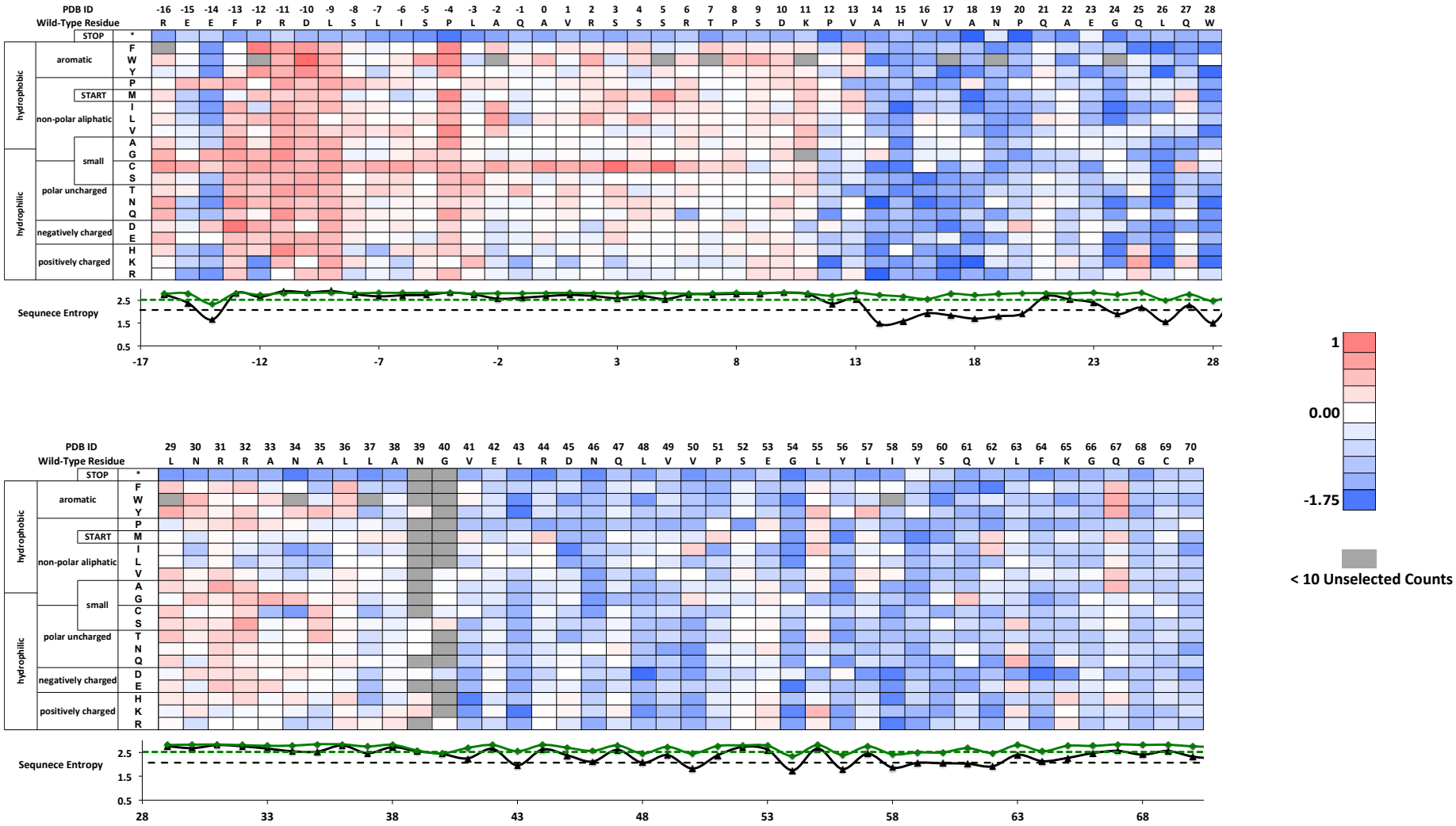
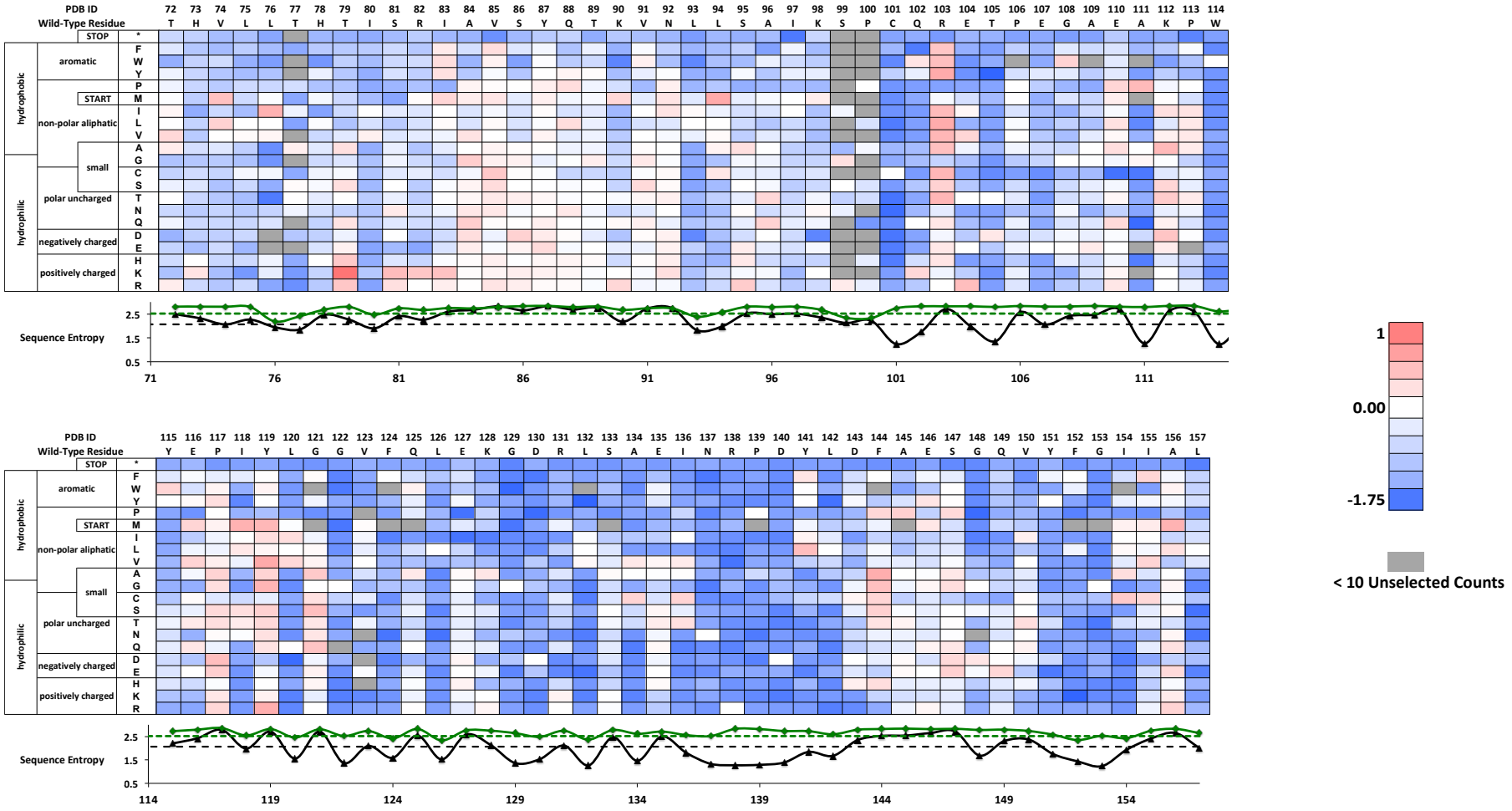



**Figure A1- TNF-Inflix_scFv Conformational Epitope Determination.**

Heatmap of fitness metric of bound vs. unselected population for all possible single non-synonymous mutations in the coding sequence for extracellular TNF. Sequence entropy for the unselected/display population (green) and unselected/bound population (black) is plotted below with their respective cut-offs (dashed lines).

## Figure A1 (cont'd)



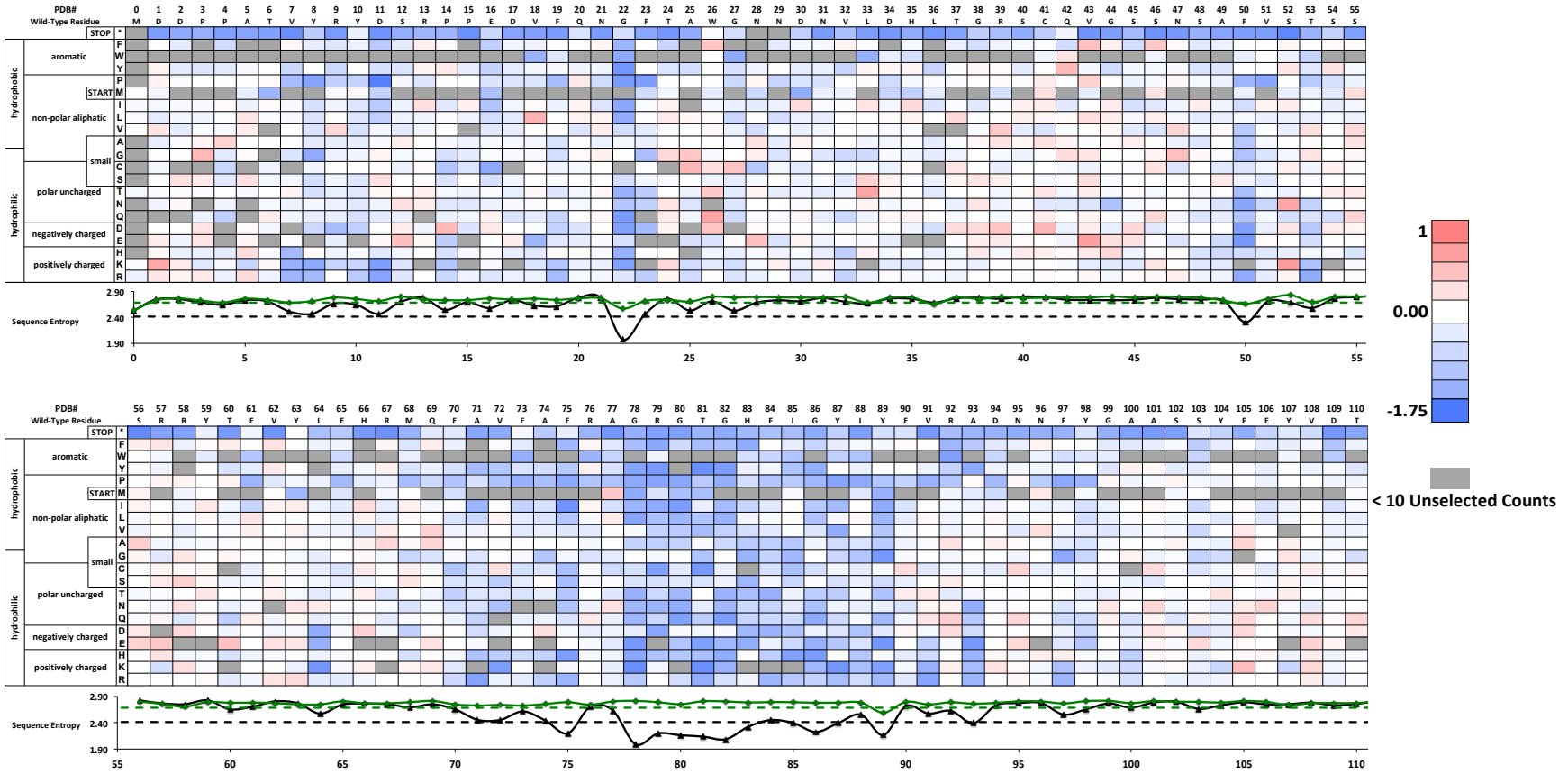
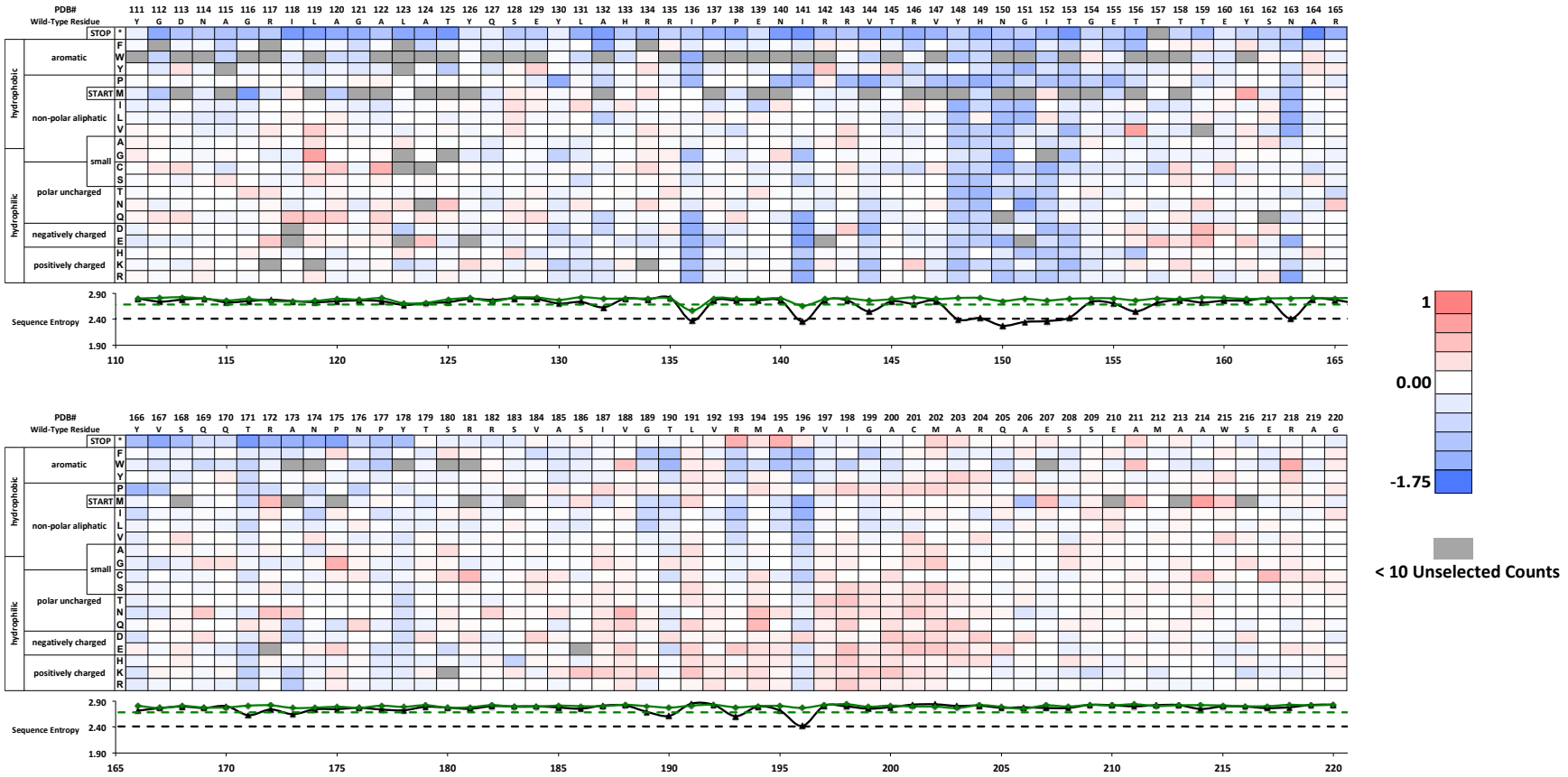



**Figure A2- PTx-S1-220-hu1B7 Conformational Epitope Determination.**

Heatmap of fitness metric of bound vs. unselected population for all possible single non-synonymous mutations in the coding sequence for Asp1-Gly220 of PTx-S1. Sequence entropy for the unselected/display population (green) and unselected/bound population (black) is plotted below with their respective cut-offs (dashed lines).

## Figure A2 (cont'd.)



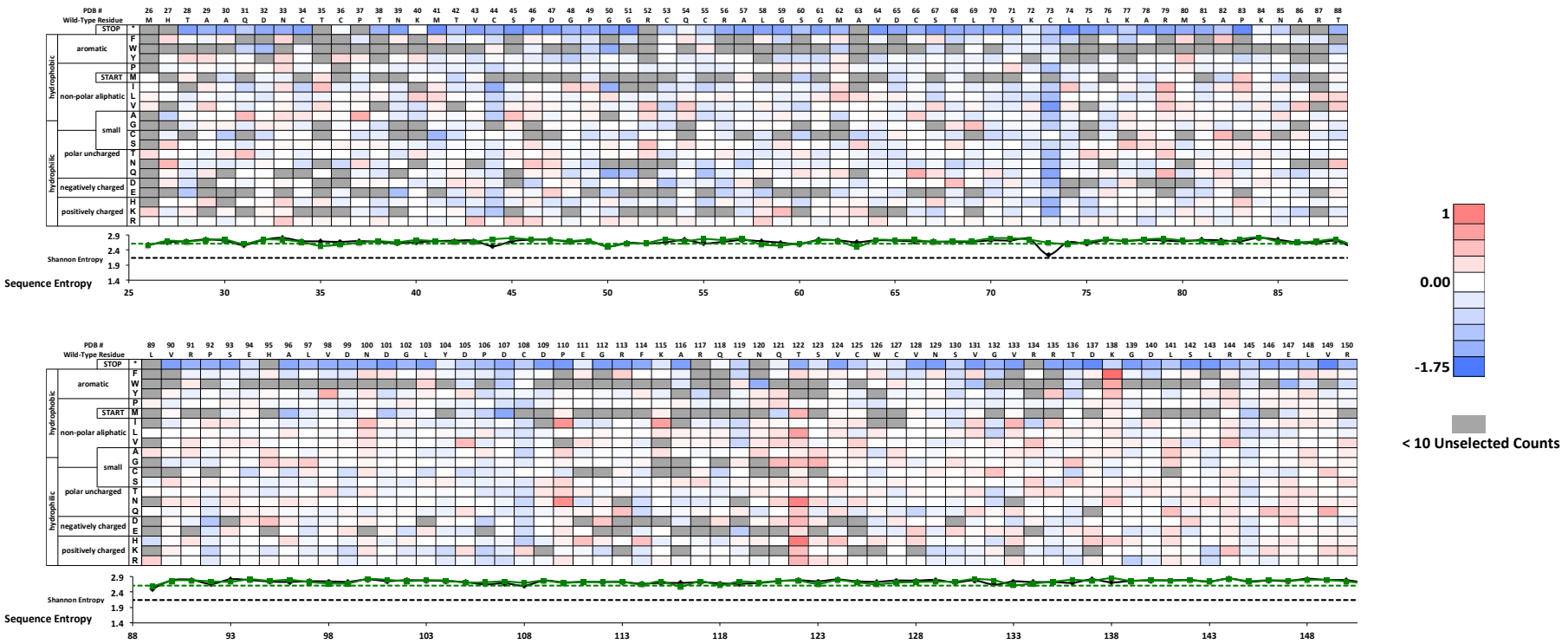
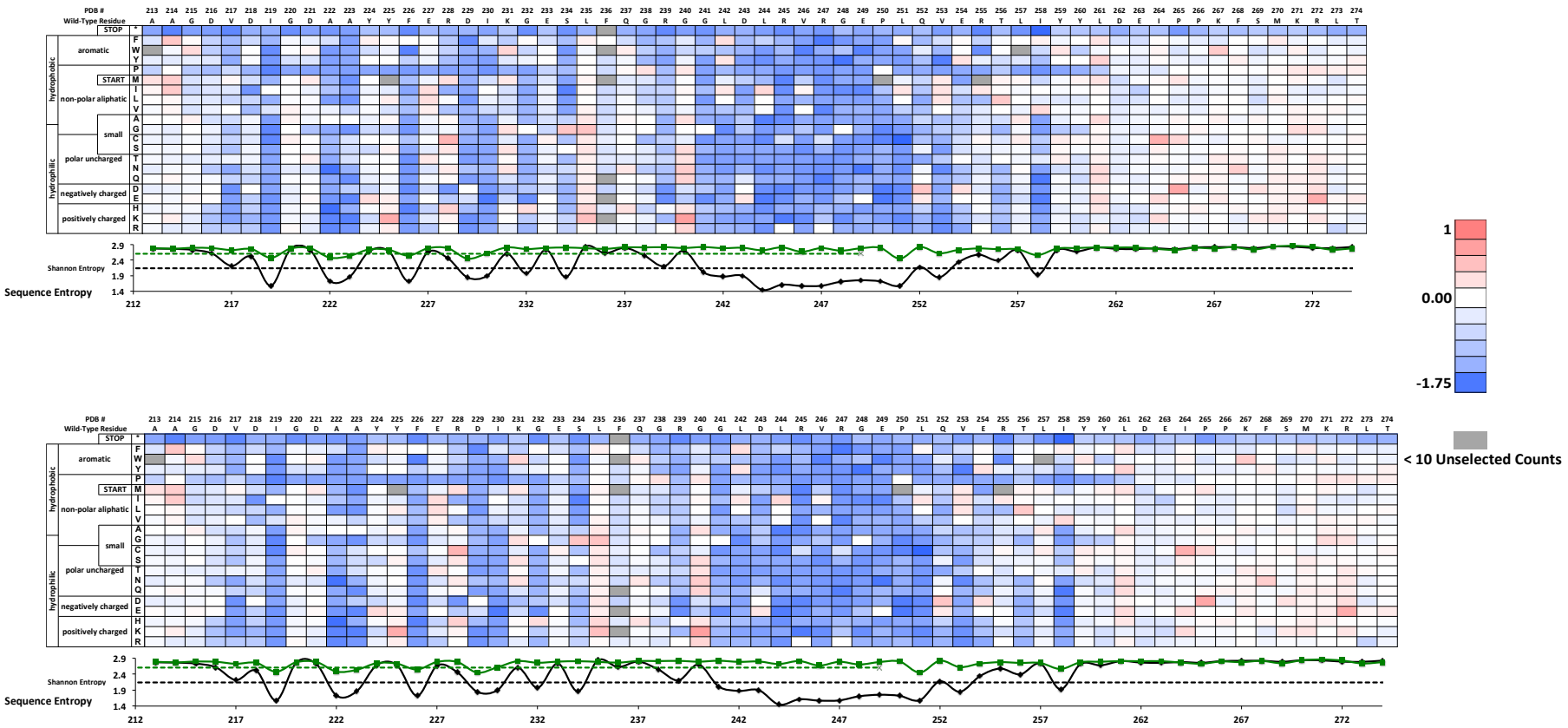



**Figure A3- TROP2-m7e6 Conformational Epitope Determination.**

Heatmap of fitness metric of bound vs. unselected population for all possible single non-synonymous mutations in the coding sequence for TROP2Ex. Sequence entropy for the unselected/display population (green) and unselected/bound population (black) is plotted below with their respective cut-offs (dashed lines).

## Figure A3 (cont'd.)



# REFERENCES

## REFERENCES


1. Ecker, D.M., Jones, S.D. & Levine, H.L. in MAbs, Vol. 7 9-14 (Taylor & Francis, 2015).

2. Cunningham, B. & Wells, J. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081-1085 (1989).

3. Gram, H. et al. In vitro selection and affinity maturation of antibodies from a naive combinatorial immunoglobulin library. *Proceedings of the National Academy of Sciences* **89**, 3576-3580 (1992).

4. Whitehead, T.A. et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature biotechnology* **30**, 543-548 (2012).

5. Levy, R. et al. Fine and Domain-level Epitope Mapping of Botulinum Neurotoxin Type A Neutralizing Antibodies by Yeast Surface Display. *Journal of molecular biology* **365**, 196-210 (2007).

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

7. Weiss, G.A., Watanabe, C.K., Zhong, A., Goddard, A. & Sidhu, S.S. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 8950-8954 (2000).

8. Smith, G.P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315-1317 (1985).

9. Pal, G., Kouadio, J.L., Artis, D.R., Kossiakoff, A.A. & Sidhu, S.S. Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *The Journal of biological chemistry* **281**, 22378-22385 (2006).

10. Rajpal, A. et al. A general method for greatly improving the affinity of antibodies by using combinatorial libraries. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 8466-8471 (2005).

11. Hill, D., Hope, I., Macke, J. & Struhl, K. Saturation mutagenesis of the yeast his3 regulatory site: requirements for transcriptional induction and for binding by GCN4 activator protein. *Science* **234**, 451-457 (1986).

12. Fowler, D.M. et al. High-resolution mapping of protein sequence-function relationships. *Nature methods* **7**, 741-746 (2010).

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

14. Tripathi, A. & Varadarajan, R. Residue specific contributions to stability and activity inferred from saturation mutagenesis and deep sequencing. *Current Opinion in Structural Biology* **24**, 63-71 (2014).

15. Matochko, W.L. et al. Deep sequencing analysis of phage libraries using Illumina platform. *Methods* **58**, 47-55 (2012).

16. Firnberg, E., Labonte, J.W., Gray, J.J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular biology and evolution* (2014).

17. Hietpas, R., Roscoe, B., Jiang, L. & Bolon, D.N. Fitness analyses of all possible point mutations for regions of genes in yeast. *Nature protocols* **7**, 1382-1396 (2012).

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

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

20. McLaughlin, R.N., Jr., Poelwijk, F.J., Raman, A., Gosal, W.S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138-142 (2012).

21. Kinney, J.B., Murugan, A., Callan, C.G. & Cox, E.C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* **107**, 9158-9163 (2010).

22. Sharon, E. et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology* **30**, 521-530 (2012).

23. Reich, L., Dutta, S. & Keating, A.E. SORTCERY—A High–Throughput Method to Affinity Rank Peptide Ligands. *Journal of molecular biology* **427**, 2135-2150 (2015).

24. McLellan, J.S. et al. Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus. *Science* **342**, 592-598 (2013).

25. McLellan, J.S. et al. Structure of RSV fusion glycoprotein trimer bound to a prefusion-specific neutralizing antibody. *Science* **340**, 1113-1117 (2013).

26. Correia, B.E. et al. Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201-206 (2014).

27. Throsby, M. et al. Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. *PloS one* **3**, e3942 (2008).

28. Sui, J. et al. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nature structural & molecular biology* **16**, 265-273 (2009).

29. Fleishman, S.J. et al. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332**, 816-821 (2011).

30. Krammer, F., Pica, N., Hai, R., Margine, I. & Palese, P. Chimeric hemagglutinin influenza virus vaccine constructs elicit broadly protective stalk-specific antibodies. *Journal of virology* **87**, 6542-6550 (2013).

31. Wei, C.-J. et al. Induction of broadly neutralizing H1N1 influenza antibodies by vaccination. *Science* **329**, 1060-1064 (2010).

32. Casina, V.C. et al. Autoantibody Epitope Mapping By Hydrogen-Deuterium Exchange Mass Spectrometry at Nearly Single Amino Acid Residue Resolution Reveals Novel Exosites on ADAMTS13 Critical for Substrate Recognition and Mechanism of Autoimmune Thrombotic Thrombocytopenic Purpura. *Blood* **124**, 108-108 (2014).

33. Pandit, D. et al. Mapping of discontinuous conformational epitopes by amide hydrogen/deuterium exchange mass spectrometry and computational docking. *Journal of Molecular Recognition* **25**, 114-124 (2012).

34. Klein, J. & Horejsi, V. Antigens, superantigens and other lymphocyteactivating substances. *Immunology*, 402 (1997).

35. Weiss, G.A., Watanabe, C.K., Zhong, A., Goddard, A. & Sidhu, S.S. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proceedings of the National Academy of Sciences* **97**, 8950-8954 (2000).

36. De Alwis, R. et al. Identification of human neutralizing antibodies that bind to complex epitopes on dengue virions. *Proceedings of the National Academy of Sciences* **109**, 7439-7444 (2012).

37. Keck, Z.-Y. et al. Mapping a region of hepatitis C virus E2 that is responsible for escape from neutralizing antibodies and a core CD81-binding region that does not tolerate neutralization escape mutations. *Journal of virology* **85**, 10451-10463 (2011).

38. Matsuzaki, Y. et al. Epitope mapping of the hemagglutinin molecule of A/(H1N1) pdm09 influenza virus by using monoclonal antibody escape mutants. *Journal of virology* **88**, 12364-12373 (2014).

39. Van Blarcom, T. et al. Precise and efficient antibody epitope determination through library design, yeast display and next generation sequencing. *Journal of molecular biology* (2014).

40. Mata-Fink, J. et al. Rapid conformational epitope mapping of anti-gp120 antibodies with a designed mutant panel displayed on yeast. *Journal of molecular biology* **425**, 444-456 (2013).

41. Doolan, K.M. & Colby, D.W. Conformation-Dependent Epitopes Recognized by Prion Protein Antibodies Probed Using Mutational Scanning and Deep Sequencing. *Journal of molecular biology* (2014).

42. Althoff, E.A. et al. Robust design and optimization of retroaldol enzymes. *Protein science : a publication of the Protein Society* **21**, 717-726 (2012).

43. Povolotskaya, I.S. & Kondrashov, F.A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465**, 922-926 (2010).

44. Cunningham, B.C. & Wells, J.A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081-1085 (1989).

45. Pál, G., Kouadio, J.-L.K., Artis, D.R., Kossiakoff, A.A. & Sidhu, S.S. Comprehensive and Quantitative Mapping of Energy Landscapes for Protein-Protein Interactions by Rapid Combinatorial Scanning. *Journal of Biological Chemistry* **281**, 22378-22385 (2006).

46. Chubiz, L.M., Lee, M.-C., Delaney, N.F. & Marx, C.J. FREQ-Seq: A Rapid, Cost-Effective, Sequencing-Based Method to Determine Allele Frequencies Directly from Mixed Populations. *PloS one* **7**, e47959 (2012).

47. Gibson, D.G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **6**, 343-345 (2009).

48. Kowalsky, C.A. et al. High-Resolution Sequence-Function Mapping of Full-Length Proteins. *PloS one* **10** (2015).

49. Bienick, M.S. et al. The interrelationship between promoter strength, gene expression, and growth rate *PLOS One, in press* (2014).

50. *Technologies, A.* (

51. Firnberg, E. & Ostermeier, M. PFunkel: efficient, expansive, user-defined mutagenesis. *PloS one* **7**, e52031 (2012).

52. Sambrook, J., Edn. 3rd ed. (ed. D.W. Russell) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. :; 2001).

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

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

55. Tinberg, C.E. et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212-216 (2013).

56. Thomas A. Kunkel, K.B., John McClary Efficient site-directed mutagenesis using uracil-containing DNA. *Methods in Enzymology* **204**, 125-139 (1991).

57. Stemmer, W.P. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389-391 (1994).

58. Hogrefe, H.H., Cline, J., Youngblood, G.L. & Allen, R.M. Creating randomized amino acid libraries with the QuikChange® multi site-directed mutagenesis kit. *Biotechniques* **33**, 1158-1165 (2002).

59. Dai, J. et al. Cloning of a novel levoglucosan kinase gene from Lipomyces starkeyi and its expression in Escherichia coli. *World Journal of Microbiology and Biotechnology* **25**, 1589-1595 (2009).

60. Bosley, A.D. & Ostermeier, M. Mathematical expressions useful in the construction, description and evaluation of protein libraries. *Biomolecular Engineering* **22**, 57-61 (2005).

61. Jain, P.C. & Varadarajan, R. A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Analytical biochemistry* **449**, 90-98 (2014).

62. Goldsmith, M., Kiss, C., Bradbury, A.R. & Tawfik, D.S. Avoiding and controlling double transformation artifacts. *Protein engineering, design & selection : PEDS* **20**, 315-318 (2007).

63. Chao, G. et al. Isolating and engineering human antibodies using yeast surface display. *Nature protocols* **1**, 755-768 (2006).

64. Benatuil, L., Perez, J.M., Belk, J. & Hsieh, C.M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein engineering, design & selection : PEDS* **23**, 155-159 (2010).

65. Clackson, T., Hoogenboom, H.R., Griffiths, A.D. & Winter, G. Making antibody fragments using phage display libraries. *Nature* **352**, 624-628 (1991).

66. Smith, G.P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315-1317 (1985).

67. Lee, S.Y., Choi, J.H. & Xu, Z. Microbial cell-surface display. *Trends in Biotechnology* **21**, 45-52 (2003).

68. Boder, E.T. & Wittrup, K.D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotech* **15**, 553-557 (1997).

69. Heewook Lee, E.P., Haixu Tang, Patricia L. Foster Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *PNAS* 1-10 (2012).

70. Tang, S.-Y., Fazelinia, H. & Cirino, P.C. AraC Regulatory Protein Mutants with Altered Effector Specificity. *Journal of the American Chemical Society* **130**, 5267-5271 (2008).

71. Perez, O.D. & Nolan, G.P. Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. *Nature biotechnology* **20**, 155-162 (2002).

72. Hammond, S.M., Bernstein, E., Beach, D. & Hannon, G.J. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature* **404**, 293-296 (2000).

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

74. Löfblom, J., Wernérus, H. & Ståhl, S. Fine affinity discrimination by normalized fluorescence activated cell sorting in staphylococcal surface display. *FEMS Microbiology Letters* **248**, 189-198 (2005).

75. Boder, E.T. & Wittrup, K.D. Optimal Screen of Surface-Displayed Polypeptide Libraries *BIotechnolgical Progress* **1998**, 55-62 (1998).

76. Moretti, R. et al. Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics* **81**, 1980-1987 (2013).

77. Fowler, D.M., Stephany, J.J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature protocols* **9**, 2267-2284 (2014).

78. Abagyan, R. & Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* **235**, 983-1002 (1994).

79. Sharabi, O., Shirian, J. & Shifman, J.M. Predicting affinity-and specificityenhancing mutations at protein-protein interfaces. *Biochem Soc Trans* **41**, 1166-1169 (2013).

80. Li, G. et al. ModuleRole: A Tool for Modulization, Role Determination and Visualization in Protein-Protein Interaction Networks. *PloS one* **9**, e94608 (2014).

81. Schymkowitz, J. et al. The FoldX web server: an online force field. *Nucleic acids research* **33**, W382-W388 (2005).

82. Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences* **99**, 14116-14121 (2002).

83. Brender, J.R. & Zhang, Y. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLoS Comput Biol* **11**, e1004494 (2015).

84. Moretti, R. et al. Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics* **81**, 1980-1987 (2013).

85. Moal, I.H. & Fernández-Recio, J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics* **28**, 2600-2607 (2012).

86. Sirin, S., Apgar, J.R., Bennett, E.M. & Keating, A.E. AB-Bind: Antibody binding mutational database for computational affinity predictions. *Protein Science* (2015).

87. Sidhu, S.S., Weiss, G.A. & Wells, J.A. High copy display of large proteins on phage for functional selections. *Journal of molecular biology* **296**, 487-495 (2000).

88. Kowalsky, C.A. et al. Rapid Fine Conformational Epitope Mapping Using Comprehensive Mutagenesis and Deep Sequencing. *Journal of Biological Chemistry*, jbc. M115. 676635 (2015).

89. Rosenfeld, L. et al. Combinatorial and Computational Approaches to Identify Interactions of Macrophage Colony-stimulating Factor (M-CSF) and Its Receptor c-FMS. *Journal of Biological Chemistry* **290**, 26180-26193 (2015).

90. Bayer, E.A., Belaich, J.-P., Shoham, Y. & Lamed, R. The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol.* **58**, 521-554 (2004).

91. Ding, S.-Y. et al. A biophysical perspective on the cellulosome: new opportunities for biomass conversion. *Current Opinion in Biotechnology* **19**, 218-227 (2008).

92. Fierobe, H.-P. et al. Degradation of cellulose substrates by cellulosome chimeras Substrate targeting versus proximity of enzyme components. *Journal of Biological Chemistry* **277**, 49621-49630 (2002).

93. Fierobe, H.-P. et al. Design and production of active cellulosome chimeras Selective incorporation of dockerin-containing enzymes into defined functional complexes. *Journal of Biological Chemistry* **276**, 21257-21261 (2001).

94. Goyal, G., Tsai, S.-L., Madan, B., DaSilva, N.A. & Chen, W. Simultaneous cell growth and ethanol production from cellulose by an engineered yeast consortium displaying a functional mini-cellulosome. *Microb Cell Fact* **10**, 89 (2011).

95. Tsai, S.-L., Oh, J., Singh, S., Chen, R. & Chen, W. Functional assembly of minicellulosomes on the Saccharomyces cerevisiae cell surface for cellulose hydrolysis and ethanol production. *Applied and environmental microbiology* **75**, 6087-6093 (2009).

96. Wen, F., Sun, J. & Zhao, H. Yeast surface display of trifunctional minicellulosomes for simultaneous saccharification and fermentation of cellulose to ethanol. *Applied and environmental microbiology* **76**, 1251-1260 (2010).

97. Carvalho, A.L. et al. Evidence for a dual binding mode of dockerin modules to cohesins. *Proceedings of the National Academy of Sciences* **104**, 3089-3094 (2007).

98. Carvalho, A.L. et al. Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 13809-13814 (2003).

99. Bogan, A.A. & Thorn, K.S. Anatomy of hot spots in protein interfaces. *Journal of molecular biology* **280**, 1-9 (1998).

100. Pinheiro, B.A. et al. The Clostridium cellulolyticum dockerin displays a dual binding mode for its cohesin partner. *Journal of Biological Chemistry* **283**, 18422-18430 (2008).

101. Stahl, S.W. et al. Single-molecule dissection of the high-affinity cohesin–dockerin complex. *Proceedings of the National Academy of Sciences* **109**, 20431-20436 (2012).

102. Haimovitz, R. et al. Cohesin-dockerin microarray: Diverse specificities between two complementary families of interacting protein modules. *Proteomics* **8**, 968-979 (2008).

103. Handelsman, T. et al. Cohesin–dockerin interaction in cellulosome assembly: a single Asp-to-Asn mutation disrupts high-affinity cohesin–dockerin binding. *FEBS letters* **572**, 195-200 (2004).

104. Mechaly, A. et al. Cohesin-Dockerin Interaction in Cellulosome Assembly A SINGLE HYDROXYL GROUP OF A DOCKERIN DOMAIN DISTINGUISHES BETWEEN NONRECOGNITION AND HIGH AFFINITY RECOGNITION. *Journal of Biological Chemistry* **276**, 9883-9888 (2001).

105. Pages, S. et al. Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain. *Proteins Structure Function and Genetics* **29**, 517-527 (1997).

106. Slutzki, M. et al. Measurements of relative binding of cohesin and dockerin mutants using an advanced ELISA technique for high-affinity interactions. *Methods Enzymol* **510**, 417-428 (2012).

107. Slutzki, M. et al. Crucial Roles of Single Residues in Binding Affinity, Specificity, and Promiscuity in the Cellulosomal Cohesin-Dockerin Interface. *Journal of Biological Chemistry* **290**, 13654-13666 (2015).

108. Studier, F.W. Protein production by auto-induction in high-density shaking cultures. *Protein expression and purification* **41**, 207-234 (2005).

109. Levy, E.D. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of molecular biology* **403**, 660-670 (2010).

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

111. Burns, M.L. et al. Directed Evolution of Brain-Derived Neurotrophic Factor for Improved Folding and Expression in Saccharomyces cerevisiae. *Applied and environmental microbiology* **80**, 5732-5742 (2014).

112. Jones, L.L. et al. Engineering and characterization of a stabilized α1/α2 module of the class I major histocompatibility complex product Ld. *Journal of Biological Chemistry* **281**, 25734-25744 (2006).

113. Miras, I., Schaeffer, F., Béguin, P. & Alzari, P.M. Mapping by site-directed mutagenesis of the region responsible for cohesin-dockerin interaction on the surface of the seventh cohesin domain of Clostridium thermocellum CipA. *Biochemistry* **41**, 2115-2119 (2002).

114. Boucher, J.I., Bolon, D.N. & Tawfik, D.S. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. *Protein Science* (2016).

115. Doolan, K.M. & Colby, D.W. Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. *Journal of molecular biology* **427**, 328-340 (2015).

116. Van Blarcom, T. et al. Precise and Efficient Antibody Epitope Determination through Library Design, Yeast Display and Next-Generation Sequencing. *Journal of molecular biology* **427**, 1513-1534 (2015).

117. Georgiou, G. et al. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology* **32**, 158-168 (2014).

118. Laserson, U. et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences* **111**, 4928-4933 (2014).

119. Benichou, J., Ben-Hamo, R., Louzoun, Y. & Efroni, S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135**, 183-191 (2012).

120. Sela-Culang, I., Ofran, Y. & Peters, B. Antibody specific epitope prediction—emergence of a new paradigm. *Current opinion in virology* **11**, 98-102 (2015).

121. Frank, R. The SPOT-synthesis technique: synthetic peptide arrays on membrane supports—principles and applications. *Journal of immunological methods* **267**, 13-26 (2002).

122. Spatola, B.N., Murray, J.A., Kagnoff, M., Kaukinen, K. & Daugherty, P.S. Antibody repertoire profiling using bacterial display identifies reactivity signatures of celiac disease. *Analytical chemistry* **85**, 1215-1222 (2012).

123. Van Blarcom, T. et al. Precise and Efficient Antibody Epitope Determination through Library Design, Yeast Display and Next-Generation Sequencing. *Journal of molecular biology* **427**, 1513-1534 (2015).

124. Doolan, K.M. & Colby, D.W. Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. *Journal of molecular biology* **427**, 328-340 (2015).

125. Sutherland, J.N. & Maynard, J.A. Characterization of a key neutralizing epitope on pertussis toxin recognized by monoclonal antibody 1B7. *Biochemistry* **48**, 11982-11993 (2009).

126. Le, J. et al. (Google Patents, 1997).

127. Chen, L.-H. et al. Expression, purification, and in vitro refolding of a humanized single-chain Fv antibody against human CTLA4 (CD152). *Protein expression and purification* **46**, 495-502 (2006).

128. Ong, Y. et al. Preparation of biologically active single-chain variable antibody fragments that target the HIV-1 GP120 v3 loop. *Cellular and molecular biology (Noisy-le-Grand, France)* **58**, 71 (2012).

129. Nguyen, A.W. et al. A binary cocktail of humanized antibodies halts whooping cough disease progression in a baboon model. *under review* (2015).

130. Liu, S.-h. et al. (US Patent 20,140,357,844, 2014).

131. Kowalsky, C.A. et al. High-Resolution Sequence-Function Mapping of Full-Length Proteins. *PloS one* **10**, e0118193 (2015).

132. Bilgin, B., Liu, L., Chan, C. & Walton, S.P. Quantitative, solution-phase profiling of multiple transcription factors in parallel. *Analytical and bioanalytical chemistry* **405**, 2461-2468 (2013).

133. Wu, M., Liu, L., Hijazi, H. & Chan, C. A multi-layer inference approach to reconstruct condition-specific genes and their regulation. *Bioinformatics* **29**, 1541-1552 (2013).

134. Liu, L. & Chan, C. IPAF inflammasome is involved in interleukin-1β production from astrocytes, induced by palmitate; implications for Alzheimer's Disease. *Neurobiology of aging* **35**, 309-321 (2014).

135. Liu, L., Martin, R. & Chan, C. Palmitate-activated astrocytes via serine palmitoyltransferase increase BACE1 in primary neurons by sphingomyelinases. *Neurobiology of aging* **34**, 540-550 (2013).

136. Patil, S.P., Tran, N., Geekiyanage, H., Liu, L. & Chan, C. Curcumin-induced upregulation of the anti-tau cochaperone BAG2 in primary rat cortical neurons. *Neuroscience letters* **554**, 121-125 (2013).

137. Wu, M., Liu, L. & Chan, C. Identification of novel targets for breast cancer by exploring gene switches on a genome scale. *BMC genomics* **12**, 547 (2011).

138. Zhang, L., Seitz, L.C., Abramczyk, A.M., Liu, L. & Chan, C. cAMP initiates early phase neuron-like morphology changes and late phase neural differentiation in mesenchymal stem cells. *Cellular and Molecular Life Sciences* **68**, 863-876 (2011).

139. Klesmith, J.R., Bacik, J.P., Michalczyk, R. & Whitehead, T.A. Comprehensive sequence-flux mapping of metabolic pathways in living cells. *under review* (2015).

140. Liang, S. et al. Structural basis for treating tumor necrosis factor α (TNFα)-associated diseases with the therapeutic antibody infliximab. *Journal of Biological Chemistry* **288**, 13799-13807 (2013).

141. Black, R.E. et al. Global, regional, and national causes of child mortality in 2008: a systematic analysis. *The lancet* **375**, 1969-1987 (2010).

142. Kim, K., Burnette, W., Sublett, R., Manclark, C. & Kenimer, J. Epitopes on the S1 subunit of pertussis toxin recognized by monoclonal antibodies. *Infection and immunity* **57**, 944-950 (1989).

143. Stein, P.E. et al. The crystal structure of pertussis toxin. *Structure* **2**, 45-57 (1994).

144. Fong, D. et al. High expression of TROP2 correlates with poor prognosis in pancreatic cancer. *British Journal of Cancer* **99**, 1290-1295 (2008).

145. Ohmachi, T. et al. Clinical significance of TROP2 expression in colorectal cancer. *Clinical Cancer Research* **12**, 3057-3063 (2006).

146. McDougall, A.R., Tolcos, M., Hooper, S.B., Cole, T.J. & Wallace, M.J. Trop2: From development to disease. *Developmental Dynamics* **244**, 99-109 (2015).

147. Maetzel, D. et al. Nuclear signalling by tumour-associated antigen EpCAM. *Nature cell biology* **11**, 162-171 (2009).

148. Stoyanova, T. et al. Regulated proteolysis of Trop2 drives epithelial hyperplasia and stem cell self-renewal via β-catenin signaling. *Genes & development* **26**, 2271-2285 (2012).

149. Vidmar, T., Pavšič, M. & Lenarčič, B. Biochemical and preliminary X-ray characterization of the tumor-associated calcium signal transducer 2 (Trop2) ectodomain. *Protein expression and purification* **91**, 69-76 (2013).

150. Pavšič, M., Gunčar, G., Djinović-Carugo, K. & Lenarčič, B. Crystal structure and its bearing towards an understanding of key biological functions of EpCAM. *Nature communications* **5** (2014).

151. Aizner, Y. et al. Mapping of the Binding Landscape for a Picomolar Protein-Protein Complex through Computation and Experiment. *Structure* **22**, 636-645 (2014).

152. Kulp, D.W. & Schief, W.R. Advances in structure-based vaccine design. *Current opinion in virology* **3**, 322-331 (2013).

153. Koide, S., Yang, X., Huang, X., Dunn, J.J. & Luft, B.J. Structure-based design of a second-generation Lyme disease vaccine based on a C-terminal fragment of Borrelia burgdorferi OspA. *Journal of molecular biology* **350**, 290-299 (2005).

154. Steel, J. et al. Influenza virus vaccine based on the conserved hemagglutinin stalk domain. *MBio* **1**, e00018-00010 (2010).

155. Bommakanti, G. et al. Design of an HA2-based Escherichia coli expressed influenza immunogen that protects mice from pathogenic challenge. *Proceedings of the National Academy of Sciences* **107**, 13701-13706 (2010).

156. From, C. Pertussis epidemic—Washington, 2012. *Morbidity and Mortality Weekly Report (MMWR)* **61**, 517-522 (2012).

157. Nguyen, A.W. et al. A cocktail of humanized anti–pertussis toxin antibodies limits disease in murine and baboon models of whooping cough. *Science translational medicine* **7**, 316ra195-316ra195 (2015).

158. Joosten, R.P. et al. A series of PDB related databases for everyday needs. *Nucleic acids research*, gkq1105 (2010).

159. Dejnirattisai, W. et al. A new class of highly potent, broadly neutralizing antibodies isolated from viremic patients infected with dengue virus. *Nature immunology* **16**, 170-177 (2015).

160. Bornholdt, Z.A. et al. Isolation of potent neutralizing antibodies from a survivor of the 2014 Ebola virus outbreak. *Science*, aad5788 (2016).

161. Georgiou, G. et al. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology* **32**, 158-168 (2014).

162. Siegel, D.L., Chang, T.Y., Russell, S.L. & Bunya, V.Y. Isolation of cell surface-specific human monoclonal antibodies using phage display and magnetically-activated cell sorting: applications in immunohematology. *Journal of immunological methods* **206**, 73-85 (1997).

163. Yeung, Y.A. & Wittrup, K.D. Quantitative Screening of Yeast Surface-Displayed Polypeptide Libraries by Magnetic Bead Capture. *Biotechnology progress* **18**, 212-220 (2002).

164. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363-382 (2008).

165. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology* **487**, 545 (2011).

166. Gray, J.J. et al. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology* **331**, 281-299 (2003).

167. Kilambi, K.P. et al. Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20–27. *Proteins: Structure, Function, and Bioinformatics* **81**, 2201-2209 (2013).

168. Weitzner, B.D., Kuroda, D., Marze, N., Xu, J. & Gray, J.J. Blind prediction performance of RosettaAntibody 3.0: grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins: Structure, Function, and Bioinformatics* **82**, 1611-1623 (2014).

169. Dominguez, C., Boelens, R. & Bonvin, A.M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **125**, 1731-1737 (2003).

170. Nilges, M., Malliavin, T. & Bardiaux, B. Protein structure calculation using ambiguous restraints. *eMagRes* (2010).

171. Kastritis, P.L. et al. A structure-based benchmark for protein–protein binding affinity. *Protein Science* **20**, 482-491 (2011).