Developing a library of recognition proteins using FimH as a protein scaffold.

Rochelle Nguyen

A thesis

submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

University of Washington

2013

Committee:

Wendy Thomas, PhD

Barry Lutz, PhD

Program Authorized to Offer Degree:

Department of Bioengineering

University of Washington

Abstract

Developing a library of recognition proteins using FimH as a protein scaffold.

By Rochelle B.L. Nguyen

High specificity recognition proteins, like antibodies, are important biological tools with many applications, including diagnostics, therapeutics, and imaging. However, there are many challenges with antibody recognition proteins, including cost of manufacturing and limited tissue penetration due to antibody size. Alternative protein scaffolds solve many of the problems associated with antibodies, but none have activatable binding affinity. Activatable binding would increase specificity by allowing tunable binding and release for therapeutic, diagnostic, and imaging applications. Bacterial adhesion protein FimH switches from high- to low-affinity binding of its ligand mannose by mechanical activation, and the binding of FimH may be altered without affecting the allosteric regulation of binding, making FimH a potential activatable protein scaffold.

In this work, the main hypothesis tested was that the binding domain of high-affinity FimH could be mutated without altering the conformation. To test this, I randomized 9 consecutive amino acids in one of the binding loops of FimH-Hi, which is locked into the high-affinity conformation by a single point mutation. I created a highly diverse bacterial display library of $4.4x10^3$-$2.0x10^4$ variants with incidence of stop codons and frame-shifts not significantly higher

than expected. A significant percentage of the library is expressed and maintained in the high-affinity conformation despite extensive randomization, suggesting that FimH may be a good protein scaffold. The results presented here demonstrate that these methods can be used to randomize all three CDR loops on the FimH-Hi lectin domain.

Chair of the Supervisory Committee:

Professor Wendy Thomas

Department of Bioengineering

Table of Contents

1 Acknowledgements

Thank you to my advisor, Wendy Thomas, for her guidance and support over these past few years. Her exceptional mentorship has an enormous positive impact on every individual in the lab, and I am so grateful for her patience and commitment to my learning.

Thank you to research scientist An-yue Tu for his molecular biology expertise and his humor and good spirits. Thank you to the rest of the Actibodies team, James Gleixner, Shivani Gupta, and Jiayi Dou, for their advice and constant willingness to help. I couldn't have asked for better teammates! Thank you to Dagmara Kisiela and the Sokurenko Lab for being an invaluable resource for all things related to microbiology.

Finally, I would like to thank my parents, Lien and Walter Nguyen, and my sisters Helene and Kim-Yen Nguyen, for their love and encouragement from afar, and I would like to thank my wonderful friends in Seattle for making these past few years so great. And of course, many thanks to Cameron Ball. His love, generosity, and intelligence know no bounds, and I am blessed to share my life with him.

2 Background and Introduction

2.1 Recognition proteins: Antibodies and alternative scaffolds

High specificity recognition proteins, like antibodies, are important biological tools with many applications, including diagnostics[1–3], therapeutics, protein separation[4,5], and imaging[6]. For most of these applications, antibodies have dominated the field because they are nature's prime example of engineering diverse specificity within a single structural scaffold.[3,7] However, there are many challenges with antibodies. Antibodies are large, with a molecular weight of about 150kDa, which results in poor tissue penetration. Furthermore, antibodies are cumbersome to produce; they are made of multiple subunits connected by disulfide bonds, which makes recombinant expression in bacteria more difficult. Any glycosylation on antibodies necessitates production in mammalian cells, making antibodies expensive to manufacture as well. And for non-therapeutic applications, where there is no need for activating an adaptive immune response, the antibody scaffold may not even be relevant. Finally, the intellectual property landscape presents unique challenges for recombinant production of new antibodies [3,7–12]. These many challenges with antibodies have driven the field towards using alternative protein scaffolds.

Alternative protein scaffolds solve many of the problems posed by antibodies [3,7–12]. Most are monomeric and are extremely thermodynamically stable in fold yet tolerant to mutation in the binding interface. Scaffolds that lack disulfide bonds or glycosylation allow ease of expression in bacteria. For therapeutics, the small size of alternative scaffolds enables better tissue penetration. For diagnostics, higher stability and expression in bacteria would increase shelf life and lower manufacturing costs. Another advantage of using small alternative scaffolds is that multivalency can easily be achieved by simply stringing monovalent domains together [8,9].

There are many alternative scaffolds with different structural characteristics: immunoglobulin-like, loop-displaying scaffolds, secondary structure binding interfaces, and oligomeric domain structures[3,7,11]. Figure 1 shows several examples of alternative scaffolds. The human Type III Fibronectin domain (Fn3) has been an attractive choice of alternative scaffold because it is monomeric and has no disulfide bonds, yet its Ig-like structure gives it structural homology to antibodies[3,7,13]. Two or three of the binding loops in Fn3 have been randomized to create diversity libraries in Fn3, and Fn3 binders have been isolated against TNF-α[14], αvβ3 integrin[15], and the SH3 domain of the human c-Src protein[16], among other targets[7,13]. From the 3-helix domain Z from staphylococcal protein A, the "affibody" was developed by randomizing helix residues on the binding area[3,7,17,18]. While the binding surface of the affibody is more ideal for protein-binding, the loop-binding of fibronectin may be more ideal for peptide or small molecule binding applications[3,7]. "Anticalins", derived from small, b-barrel proteins called lipocalins, have been re-engineered to bind small molecules, including for applications in cancer therapy, but these alternative scaffolds have disulfide bonds[19–21]. Pluckthun *et al*. used ankyrin repeat proteins to engineer modular binding[22,23], as well as picomolar binding to the HER2 receptor[24]. Ankyrin repeat proteins, like the affibody, are more suitable for protein-binding than for small molecule binding. For each of these examples, the potential scaffold was tested by creating a diversity library by randomizing sequences in the putative binding region, followed by identification of binders to new targets. A suitable protein scaffold must be able change its specificity without compromising structural stability, and the aforementioned protein scaffolds, among others, have demonstrated this quality. However, neither antibodies nor these alternative scaffolds are activatable.

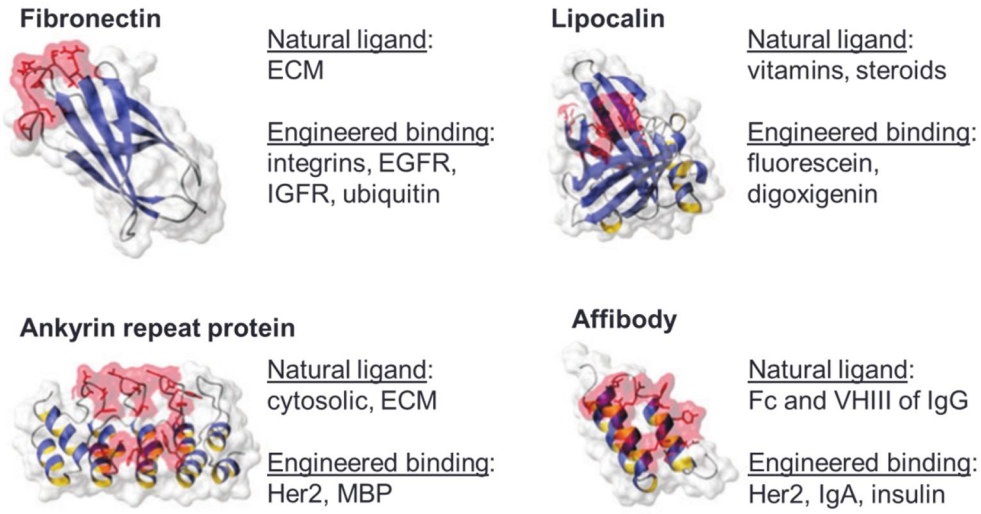

Figure 1: Examples of non-antibody alternative scaffolds, their natural ligands, and non-natural targets to which binding was engineered by library screening. Adapted from Binz *et al.*[3]

2.2 The need for activatable recognition proteins

There are many potential applications for recognition proteins with activatable or regulatable binding. For bioseparation, directed activation and deactivation of binding could more specifically release antigen for further study without contaminating the captured antigen with modifications or tagging probes. For imaging, there is a need for higher signal-to-noise ratio, which could be improved if the contrast agent could be activated to bind only at the target tissue. Targeted drug delivery may be more efficient if both a molecular target as well as an activating mechanism were required. In general, in all of these potential applications, activatable binding would increase specificity by allowing tunable binding and release.

2.3 Current strategies for activatable protein binding

In general, the most highly specific imaging probes need to combine imaging agents (fluorophores, radionuclides, etc.) with a targeting molecule (antibodies, receptor-specific ligands, etc.), and the mode of activation is specific to the imaging agent, not the targeting mechanism[25]. Current strategies for activating fluorescent optical imaging probes include using

enzymes or relying on lysosomal internalization by target cells[25,26]. However, the use of enzymes requires specific substrate, which limits broad applicability of this activation strategy, and lysosomal internalization has the difficulty of low signal amplification[26].

The monoclonal antibody HPC4 is activated to bind by Ca2+ interaction with target Protein C[27], but this mode of binding activation is specific to the binding epitope rather than the antibody, limiting the broad applicability of this strategy. There have been attempts to develop photochemical methods to activate cyclodextrin binding[28], but these methods require the use of organic solvents, which render it infeasible for any potential *in vivo* application. Inactivation and subsequent activation of antibody-antigen binding has been shown by Self *et al*. using UV irradiation[29], but this method requires chemical modification of the antibody using diphosgene, which is extremely toxic[30]. Furthermore, only 25% of antibody binding was recovered after 15minutes of UV irradiation, and 30 minutes of UV radiation resulted in antibody damage[29]. These studies were performed in plate-based assays with purified protein, and UV has limited tissue penetration, rendering this method of activation useless for deeper tissue *in vivo* applications. For diagnostic protein antigen recovery, Lewandowski *et al*. proposed a capture and release method based on enzymes and stimuli-responsive chitosan, but release methods resulted in either His-tag or residual sugar contamination of recovered protein[31,32].

The overall goal of this work is to develop a new protein scaffold for potential therapeutic application, by engineering activatability under *in vivo* conditions, and for potential diagnostic or biotechnology application, by engineering binding and release without contamination of desired protein isolates. We may be able to achieve regulatable high-affinity binding by using an allosteric protein scaffold.

2.4 FimH as a protein scaffold

FimH is an allosteric bacterial adhesion protein expressed on the tips of Type 1 fimbria in *E. coli*. Fimbriated *E. coli* display FimH by expression of genomic *fim* operon, which contains genes *fimSBEAICDFGH*. Fimbriae are made up of many subunits of FimA, with FimH expressed only at the tip of each fimbria.[33] FimA is the major subunit of fimbriae, and although *fimA* lies upstream of FimH, piligenesis is activated by FimH expression. Expression of the *fim* operon is initiated by transcription factors encoded by *fimB* and *fimE*, which control orientation of promoter sequence *fimS* upstream of *fimA*. The orientation of the invertible *fimS* sequence determines whether or not piligenesis occurs; *fimS* in the "phase-ON" position results in FimH expression and thus the fimbriated phenotype, whereas *fimS* in the "phase-OFF" inversion results in non-fimbriated bacteria. The binding of FimB to *fimS* slightly favors switching to the phase-ON position, and the binding of FimE strongly favors the phase-OFF inversion. It is important to note that even exogenous *fimB* can initiate switching of endogenous *fimS* and therefore initiate endogenous FimH expression. FimC is a chaperone protein that guides Fim proteins through the periplasm to the membrane, where usher protein FimD localizes Fim proteins in preparation for piligenesis. FimF and FimG control the elongation of the fimbria as they extend through the membrane to display FimHextracellularly.[33]

FimH consists of a ligand-binding lectin domain, which binds mannose via three complementarity-determining regions (CDRs), and a regulatory pilin domain[34,35]. FimH exhibits "catch bond" behavior, where it allosterically switches between low and high affinity with changes in shear stress under *in vivo* conditions. When the pilin domain is bound to the lectin domain, FimH is locked in the low-affinity conformation, but when the pilin and lectin domains are pulled apart by mechanical force, a β-sheet twisting mechanism opens up a binding pocket

for its ligand mannose to bind the CDRs (Figure 2). Several mutations in FimH have been identified that stabilize the protein in one conformation. The A188D variant of FimH, which from here on will be referred to as FimH-Hi, is stabilized in the high-affinity state by a single amino acid mutation in the pilin domain. Crystal structures of FimH bound to mannose show that the interactions of mannose with CDR2, specifically residues N46, D47, and N54, are critical for mannose binding due to hydrogen bonding[36–39]. Because the binding site on the lectin domain is far from the regulatory pilin domain[34,35], the specificity and regulatability of a FimH-based scaffold may be decoupled. FimH may therefore be an ideal protein scaffold from which to generate activatable recognition proteins. If we can engineer the lectin domain to bind different targets, thus changing specificity of FimH binding, while retaining the allosteric regulation of the pilin domain, this would result in a library of activatable recognition proteins, called "actibodies", using FimH as a protein scaffold. Here, I tested the hypothesis that the lectin domain of FimH could be mutated without altering the allosteric regulation of binding, by randomizing 9 consecutive amino acids in the CDR2 loop of FimH-Hi (Figure 3).

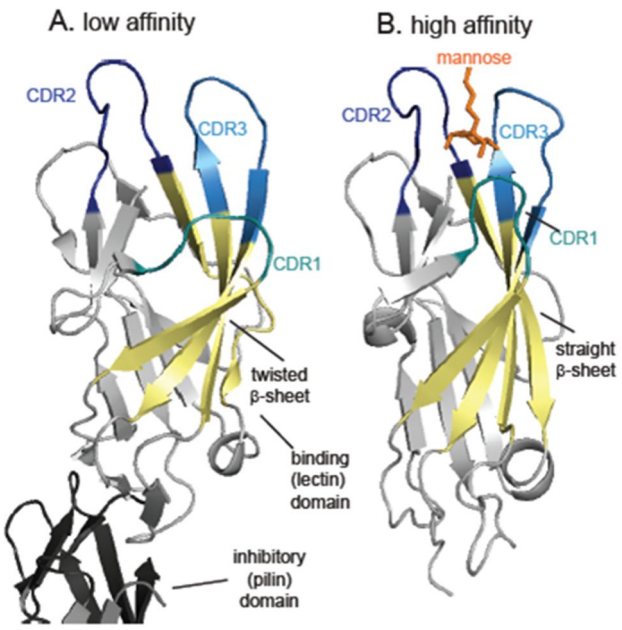

Figure 2: FimH structure. (A) FimH is locked in the low-affinity state by the inhibitory pilin domain. (B) When the pilin domain is no longer docked to the lectin domain, FimH changes conformation, forming a pocket for mannose to bind the CDR loops.

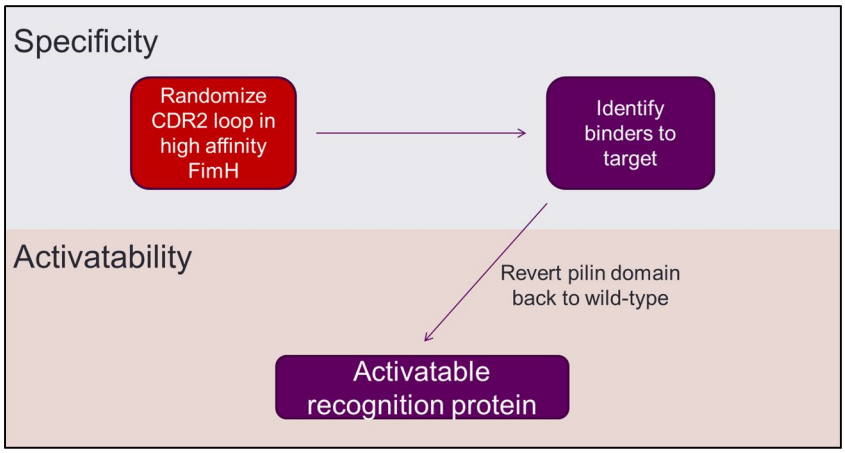

Figure 3: Flow-chart of the overall "actibody" strategy. If binding and regulation of FimH can be decoupled, screening a pre-activated FimH-A188D library against a new target, followed by reversion to activatable FimH by the single point mutation D188A, would theoretically yield activatable recognition proteins. The red high-lighted step is what is presented in this thesis.

3 Methods

3.1 Brief review of library construction methods

There are three main categories of introducing diversity into loop regions (Figure 4): randomization along an entire gene, targeted randomization, or rearrangement of existing

diversity[40,41]. Random mutagenesis along an entire gene (Figure 4A) consists of doing error-prone PCR (epPCR) with low-fidelity DNA polymerases to introduce mutations randomly along a gene[40,41]. This method is easy, fast, and appropriate for affinity maturation of a protein of interest. However, epPCR does not generate high enough diversity for a naïve library*. Recombination methods (Figure 4B) imitate homologous recombination in nature, where two related starting sequences recombine by exchanging homologous regions. But recombination requires homology, and less homologous templates may fail to recombine, limiting library diversity[41]. If information is known, either from a crystal structure or from sequence alignments, about a specific region of interest, targeted randomization (Figure 4) is ideal for library construction because it allows for more control over mutagenesis[40,41].

One method of introducing targeted randomization is by incorporating trinucleotide phosphoramidites, where each trinucleotide encodes an amino acid, allowing control over the mixture of codons for each position[41–43]. While this helps reduce library bias, the complex synthesis of trinucleotides and the high cost limit their use in the construction of diversity libraries[40,41]. Furthermore, codon bias may even be beneficial since codon degeneracy evolved to favor certain amino acids.

* There are two main types of libraries: diversity libraries and libraries for directed evolution. Diversity libraries, also called naïve libraries and primary libraries, refer to large libraries that are used for identifying binders to novel targets. Secondary libraries for directed evolution or affinity maturation are smaller and are used to identify the same target with increased activity or binding. Diversity libraries are not usually able to be fully sampled by screening methods, whereas complete sampling is often required or highly desired for directed evolution libraries[41,75,76].

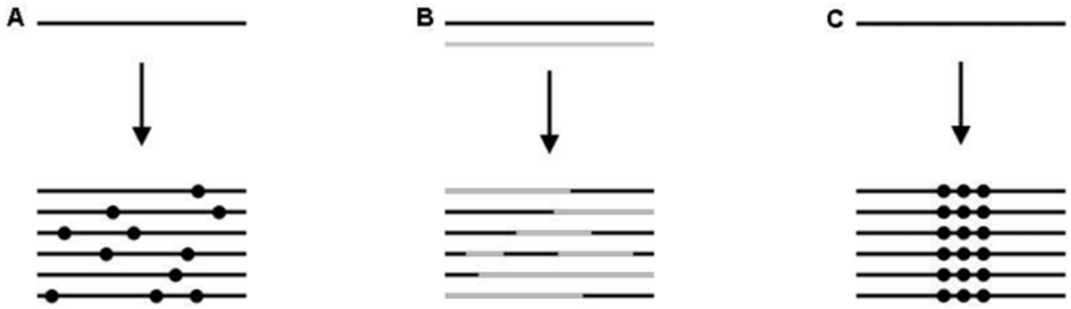

Figure 4: Schematic representation for the three main categories of library creation methods: (A) random mutagenesis along an entire gene, (B) recombination, and (C) targeted randomization. Adapted from Lipovsek *et al.*[40]

The most common method of creating libraries is oligonucleotide-based randomization with degenerate oligonucleotides (Figure 5), where a mixture of synthetic oligos with any of the 4 nucleotide bases occupy each position of codons of interest NNN[40,41,44]. Variations on this include limiting the last nucleotide of the codon to minimize the introduction of a stop codon into the mutated sequence. Whereas NNN has a 3/64 rate of stop codon addition, NNK, where K can only be a G or a T, and NNS, where S can only be C or G, reduce the rate to 1/32.[40,41,44] In all cases, there are 32^n possible combinations of codons, were *n* is the number of residues randomized. Although there is a higher risk of library bias with degenerate oligos, degenerate oligos allow for high diversity in a sequence of interest, do not rely on homologous sequences, are readily available from commercial entities such as IDT, all at a low cost, making this method ideal for the construction of naïve libraries in protein scaffolds[40,41].

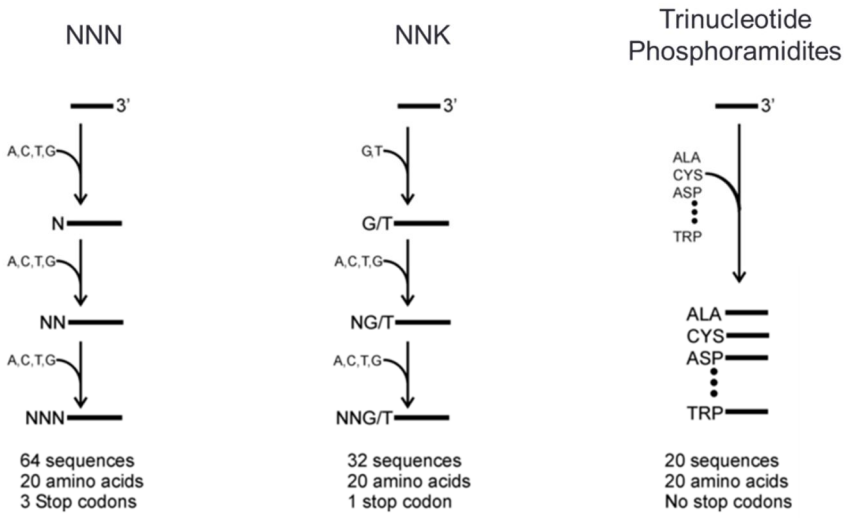

Figure 5: Strategies for degenerate oligonucleotide-based randomization. NNN randomization makes 64 possible sequences with a 3/64 chance in encoding a stop codon. NNK randomization makes 32 possible sequences with only a 1/64 chance of encoding a stop codon. Trinucleotide phosphoramidites encode codons as trinucleotides and eliminates the possibility of introducing a stop codon. Adapted from Lipovsek *et al.*[40]

Degenerate oligos have been used to create many examples of diversity libraries using methods that fall into three main categories depicted in: QuikChange®-based mutagenesis, Kunkel-based mutagenesis, or gene assembly (also known as assembly PCR or recursive PCR). Stratagene's QuikChange® Site-Directed Mutagenesis kit uses complementary primers to introduce a mutation into a plasmid vector[45], but saturation mutagenesis using degenerate oligos using this method has primarily been optimized for single-site mutagenesis[44,45], not for several consecutive residues, limiting the use of this method to creating very small libraries. Thomas Kunkel developed a method of site-specific mutagenesis utilizing DNA template containing uracil instead of thymine, followed by PCR with a primer containing the desired mutation[46]. The advantage of this method is that it only requires a single-stranded template for mutation incorporation, and the uracil-containing wild-type sequence can more easily be degraded by bacteria[46–49]. Some challenges of the Kunkel method include the potential of having wild-type sequence remaining in the final library, compromising library quality, as well as necessity of using M13 bacteriophage and certain bacterial strains, which limits the applicability of the

method. For example, the Kunkel method requires plasmids with antibiotic resistance incompatible with our FimH expression plasmids.

Gene assembly utilizes a series of overlapping oligos with degenerate sequences in the gap regions between overlaps to fully synthesize a randomized gene of interest[50–53]. This method is widely used in generating libraries in alterative scaffolds. Affibody libraries were generated by using this method to randomize a total of 13 amino acids spanning 2 helices on the protein[17,18,54]. Binz *et al*. used gene assembly to randomize 7 residues in ankyrin repeat proteins[22]. An Avimer library was created by gene assembly to randomize non-conserved residues[55].

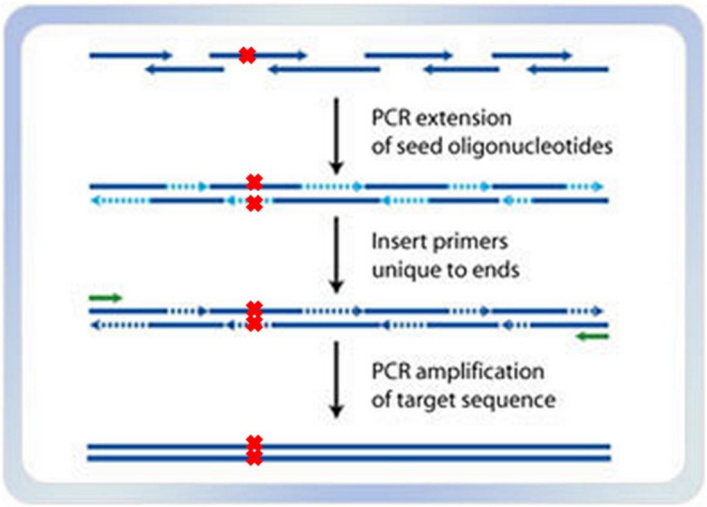

Figure 6: Gene assembly. Overlapping oligonucleotides with degenerate codons (indicated by red X's) in the gap regions are used to synthesize full-length genes of interest by PCR.

Since we encountered problems with overlapping antibiotic resistance genes with the Kunkel method, and since 9 consecutive amino acids is too long to randomize by QuikChange®-based complementary primers (data not shown), I decided to use gene assembly to generate CDR2 library in FimH-Hi.

3.2 Oligo design for gene assembly

To design our overlapping oligos, I used an online resource called DNAworks developed by Helix Systems at the NIH[53]. I designed oligos to synthesize FimH lectin domain gene from residues 2-163 (471bp total in length) while requiring that the randomized regions remain in the gaps between complementary overlapping regions of oligos. NNK randomization was used instead of NNN to reduce the probability of introducing a stop codon. Input parameters: *E. coli* class II organism; annealing temperature 58°-70°C; oligonucleotide length 45-80nt; restriction site exclusion KpnI and XcmI; fixed CDR sequences in gaps. DNAworks returned genetic sequences of dozens of possible sets of oligos, from which we chose one set that met all of our requirements. This was a set of 8 oligos of length of 48-79nt, with complementary overlapping regions of 15-22bps and the NNK regions in the gaps.

3.3 Gene assembly and library amplification

To synthesize the FimH gene, a gene assembly reaction was made using 1.6µM oligo mixture, 0.24mM dNTPs, and Pfu Turbo DNA polymerase (NEB) in 25µl total. The reaction was run with 25 cycles of 94°C for 30s, 55°C for 30s, 68°C for 1min. For our positive control, we used wild-type oligos instead of NNK oligos. The product was run on a gel, and the remainder was purified using PCR clean-up (Qiagen). Full-length product was amplified by PCR using short end primers, followed by PCR clean-up (Qiagen). To check sequences, the gene mixture was cloned into pCR®-Blunt by Zero Blunt® PCR Cloning Kit (Invitrogen) and transformed into TOP10 competent cells for blue/white selection. Six wild-type and six CDR2 mutants were sequenced. After verification of diversification of CDR2, the gene assembly product as well as wild-type control were inserted into FimH plasmid pGB224-A188D using the Gibson assembly kit (NEB). 0.12 pmol of either wild-type or CDR2 mutant insert was added to

0.06 pmol of PCR-generated pGB224-A188D vector and incubated with Gibson buffer for 50°C for 1 hour. To check for successful incorporation of insert into vector, Gibson product was transformed into XL1b chemically competent cells. Ten mutants were sequenced. The Gibson product was transformed into MegaX DH510b electrocompetent cells (Invitrogen) in order to amplify the FimH plasmid library. Again, wild-type and 10 CDR2 mutants were sequenced.

3.4 Competent cell preparation

Electrocompetent cells (ΔFliC or MegaX-derived) were prepared by growing up 1ml of overnight starter culture in 100ml of LB with appropriate antibiotic for 3-4 hours until OD600 of 0.35-0.4 was observed. Cells were then cooled on ice, pelleted, and washed with ice-cold sterile dH2O, concentrated 1000x, and stored as 50µl aliquots in 10% glycerol.

Chemically competent cells (DH5α-derived) were prepared by growing up 1ml of overnight starter culture in 100ml of LB with appropriate antibiotic for 3-4 hours until OD600 of 0.35-0.4 was observed. Cells were then cooled on ice, pelleted, and washed with ice-cold 0.1M $MgCl_2$ followed by 0.1M $CaCl_2$. Cells were concentrated 1000x in $CaCl_2$, and stored as 50µl aliquots in 0.1M $CaCl_2$, 15% glycerol.

3.5 Transformation into expression library

Constitutive FimH expression consists of a two-plasmid system: *fimH*-A188D library pGB224 plasmid, and either pPKL114 (*fimB*) or pFimΔH (*fimAICDFG*).

Heat shock: 1-2µl of DNA was added to chemically competent cells thawed on ice. Reaction was incubated on ice for 15-20min, then subjected to 42°C for 45 seconds, then immediately returned to ice for 2 minutes. Pre-warmed SOC media was added to dilute 10-fold, and then reactions were incubated for 1 hour at 37°C shaking at 200rpm. Transformations were then diluted and plated in 50µl volumes.

Electroporation: Up to 1µl of DNA was added to electrocompetent cells thawed on ice. Cells were then electroporated under the following conditions: 25µFd, 200Ω, 1.7kV (pFimΔH/ΔFliC) or 2.0kV (MegaX) using the BioRad GenePulser II in the Sokurenko lab. BioRad 0.1mm cuvettes work optimally. Immediately after electroporation, cells were resuspended in 1ml of room temperature SB media, transferred to round-bottom culture tubes, and incubated for 1 hour at 37°C shaking at 200rpm. Transformations were then diluted and plated in 50µl volumes.

3.6 Colony PCR

To amplify *fim(fgh)*,10x buffer, 1.5µl of 1:10 dilution of "Fim op" primers (James Gleixner), 1µl 2mM dNTPs, water, and Taq polymerase were combined to a total of 50µl. The cycling reaction: 95°C for 30s; 95°C for 1min, 55°C for 1min, 68°C for 1.5min for 25 cycles; 68°C for 7min, 4°C forever.

3.7 Crystal violet assay for FimH-RNAseB binding

Bacteria were grown up overnight in 5ml SB with antibiotics, at 37°C with shaking. The next day, cells were prepared by centrifuging at 4500 RPM for 10min and washing with 1ml of PBS 3x. RNAseB was immobilized to wells of 96-well plate (Immulon HBX, Thermo) at 20µg/ml in $NaHCO_3$ for 1 hour at 37°C. Wells were then washed 4x with PBS and blocked with 0.2% BSA-PBS. After washing twice with PBS, residual buffer was aspirated and bacteria were added to a final OD of 2 in 0.2%BSA-PBS at 37°C for 40min. After washing 6x with PBS, cells were fixed at 60°C for 5min, and a 1:4 dilution of crystal violet stain was added to each well and incubated for 15min at room temperature. Wells were washed 5 times with water, and 50% ethanol was added to wells. Absorbance at 600nm was measured with 1 second mixing using the Woodrow lab's TECAN plate reader.

3.8 Analysis of library quality by flow cytometry

FimH expression or conformation was verified using flow cytometry with either αPD or Mab21 primary antibody. An overnight culture of colonies were pelleted and washed with PBS twice, followed by resuspension into PBS for a final OD of 2. Cells were pelleted and washed once more in PBS and then blocked with 0.2% BSA-PBS for 30 minutes at room temperature with inversion. Cells were then pelleted and washed with PBS, followed by resuspension in 1:500 αPD or 1:200 Mab21 in 0.2% BSA-PBS for 1 hour at room temperature. Following incubation, cells were pelleted and washed 2 times in PBS, and then incubated for 1 hour in the dark at room temperature in secondary antibody (goat anti-rabbit for αPD and goat anti-mouse for Mab21) diluted to 1:2000 in 0.2% BSA-PBS. Unstained cells were not incubated with secondary antibody. Cells were pelleted and washed 3 times, and then resuspended in PBS at about 10^8-10^9 cells per ml into a BD round-bottom tube for flow cytometry. Samples were run on FACS Canto 3 at the Cell Analysis Facility in the UW Immunology Department. Laser settings: FSC sensitivity 550, SSC 597, Alexa Fluor-488 431, with threshold FSC and SSC at 10,000. 100,000 events were collected for each sample, and data was analyzed with FloJo 10.0.

4 Results

4.1 Gene assembly oligo design output

The set of oligonucleotides chosen from DNAWorks output contained 8 oligonucleotides encoding residues 2-163 of the FimH lectin domain, with CDRs completely constrained to the gap regions (Figure 7). Oligonucleotide lengths ranged from 48-79 nucleotides, the melting temperatures were 57-58°C, and the overlap lengths ranged from 14-22 bp.

All 9 amino acids of the CDR2 loop were randomized by NNK oligonucleotides. Therefore, theoretically, the maximum library size for randomizing 9 amino acids simultaneously using NNK oligonucleotides is $20^9 = 5.1x10^{11}$ variants.

Figure 7: DNAWorks overlapping oligonucleotides for gene assembled synthesis of FimH lectin domain. Highlighted sections denote the CDRs. Each of these oligonucleotides was ordered from IDT, and oligonucleotide #3 containing CDR2 was ordered with NNK codons for CDR2.

4.2 Gene assembly

A mixture of the 8 overlapping oligonucleotides was used to generate a gene fragment of the FimH lectin domain with randomized CDR2 by gene assembly. The product was run on an agarose gel and a smear was observed, which is consistent with what is expected for gene assembly reactions[50,52] (Figure 8). There was a faint band at ~470bp, indicated in Figure 8 by the white arrow, as well as a very bright band at the bottom, indicative of short, <100bp fragments of DNA.

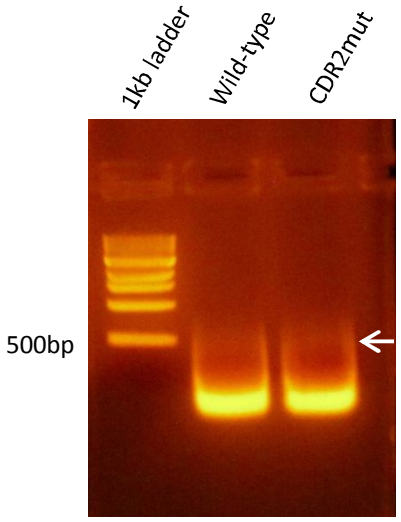

Figure 8: Gene assembly product. Eight overlapping oligonucleotides were mixed and allowed to anneal in a thermal cycler. For the wild-type sample, all oligonucleotides contained wild-type FimH CDR2 sequences. CDR2mut included a mixture of one of the oligonucleotides to randomize all 9 residues of CDR2. The white arrow is pointing to the faint band at ~470bp indicative of full-length product.

Since the oligonucleotides were the limiting reagent in the gene assembly reaction, the upper limit of the library size is 2.4×10^{13}, which is 100-fold higher than the theoretical library size based on NNK randomization. Since only a faint band at ~470bp was observed in Figure 8 (estimated to be about 5ng of DNA), the estimated library diversity at this point is 9.3×10^{9} variants.

Amplification of the gene assembly reaction is shown in Figure 9. For both the wild-type and randomized samples, a bright ~470-bp fragment was observed, and there was also slight

amplification of one band ~700bp in length in the CDR2 mutants. The library size in this step was only limited by the amount of gene assembly product used as template; therefore, the diversity represented in 0.5µl of gene assembly product was $1.9x10^8$ variants.

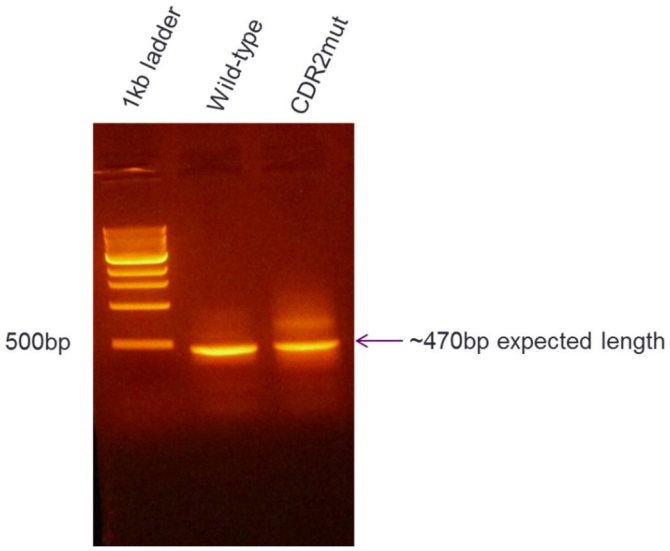

Figure 9: PCR amplification of full-length gene assembly product. Full-length (~470bp) gene assembly product was amplified using outside primers.

To verify randomization in the CDR2 region of the FimH lectin domain gene, amplified gene assembly product was cloned into pCR®-Blunt and transformed into TOP10 competent cells. Five clones were sequenced, and all 5 had non-wild type sequences in CDR2 (Table 1). One clone had mutations in the first few residues of CDR1, and another clone had mutations in CDR3. The incidence of stop codons was only 20%, which is not significantly different from the 25% expected frequency of stop codons ($p>0.1$). The incidence of frame-shift observed was 40%. According to Bessette *et al.*, the frequency of frame-shifts arising from a single nucleotide deletion is between 0.4-1.3 per kilobase. So the expected incidence of frame-shifts for this 470bp gene assembled product is 20-65%. Finally, one clone was missing CDR3, leaving 20% theoretically functional (lacks stop codons and frame-shifts). Each randomized CDR2 sequence was non-wild type, and no two sequences were identical, suggesting that this method

successfully introduced high diversity. Clones transformed with wild-type oligonucleotides retained wild-type sequences.

Variant	CDR2 sequence
Wild-type	NDYPETITD
Mut1	GGXKHGGAG
Mut2	GLVSMLGLL
Mut3	FSRWGVLSE
Mut4	HEMGFLHAQ
Mut5	LISIVVVCWG

Table 1: Sequencing results for CDR2 mutants. Amplified gene assembly product was cloned into pCR®-Blunt and transformed into TOP10 competent cells. Five clones were sequenced. The red X represents a stop codon.

4.3 Gibson assembly

The amplified gene assembly fragment was inserted into a PCR-generated plasmid vector by Gibson Assembly® Cloning Kit (New England Biolabs). To verify insertion of gene assembly product into plasmid vector, Gibson assembly product was transformed into XL1b chemically competent cells, and 10 clones were sequenced. 100% were positive for the gene assembly insert (Table 2). 90% of clones sequenced contained unique, non-wild type CDR2 sequences, 30% contained a stop codon, and 30% contained frame-shifts, leaving 40% theoretically functional. Gibson assembly did not alter the frequency of stop codons or frame-shifts arising from the prior gene assembly reaction with NNK primers. The overlap regions for Gibson assembly retained wild-type sequences with no frame-shifts. Gibson assembly product was not run on a gel, so it is not possible to estimate the efficiency of insertion into the plasmid vector.

Variant	CDR2 sequence
Wild-type	NDYPETITD
Mut1	QVAVTPFLC
Mut2	GAGGDSAWG
Mut3	NDYPETITD
Mut4	GGAISGYXF
Mut5	RSRWCRVEL
Mut6	YSLSDWXLA
Mut7	XADTGFGRR
Mut8	VRGRSGPLV
Mut9	AGTSGRHQV
Mut10	GPARGGARA

Table 2: Sequencing results after Gibson assembly. Gibson assembly product was transformed into TOP10 for sequencing. Red X's denote stop codons.

4.4 Amplification of plasmid library by transformation in MegaX DH10b

In order to amplify the plasmid library while maintaining the number of unique sequences, the Gibson assembly product was transformed into MegaX DH10b (Invitrogen). The mass of Gibson product to be used for transformation was optimized for maximum transformation efficiency, and I observed the highest efficiency on the order of 10^6 CFU/μg DNA with 43.1ng of Gibson product (Figure 10). This corresponds to 4.4×10^4 transformants/ml (95% CI = $3.9 \times 10^4 - 5.1 \times 10^4$).

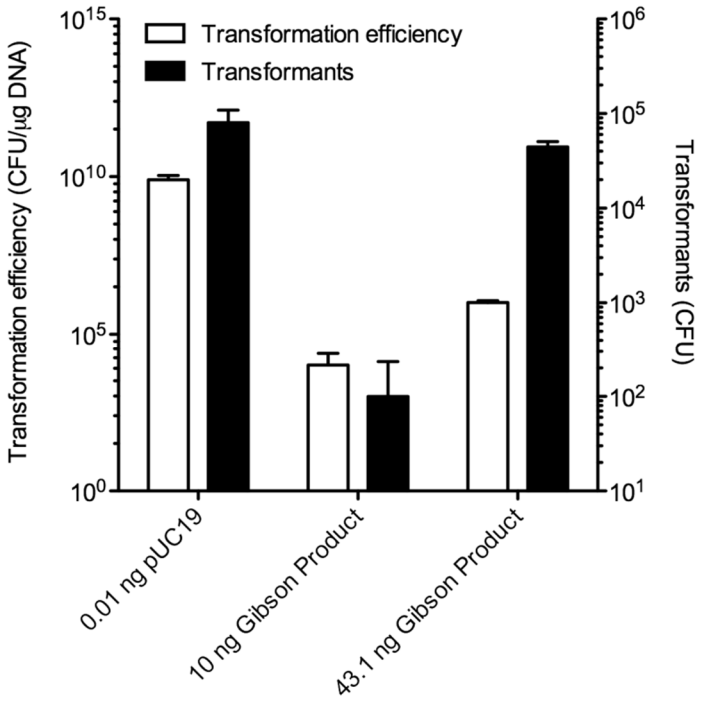

Figure 10: Optimization of transformation of plasmid library into MegaX DH10b. All transformations were done with an electric field potential of 20kV/cm, 25µF, and 200Ω. All data was taken from a single plate, and confidence intervals were calculated based on Poisson statistics.

Ten clones were sequenced to verify that diversity was maintained after amplification by MegaX DH10b cells (Table 3). 80% contained unique, non-wild-type CDR2 sequences, 0% contained frame-shifts, 20% contained a stop codon, leaving 60% of clones sequenced to encode theoretically functional protein. I did not expect any change in frequency of stop codons or frame-shifts, but due to the small sample size of only 10 clones sequenced, observing 0% with frame-shifts is not significantly different. Since I did not observe any significant deviation from the expected frequencies of stop codons or frame-shifts, I calculated the expected loss of library size from stop codons and frame-shifts to be between 45%-90%, resulting in a total functional library size after transformation to be 4.4×10^3 - 2.0×10^4.

Variant	CDR2 sequence
Wild-type	NDYPETITD
Mut1	YSGSEEAAG
Mut2	NDYPETITD
Mut3	FailedRxn
Mut4	MGACGAHRW
Mut5	GYGSSQASV
Mut6	VCXNTIESN
Mut7	PCYHVTQGG
Mut8	VMIQXAIGG
Mut9	NDYPETITD
Mut10	RLAWSGSSR

Table 3: Sequencing verification of plasmid library amplified by MegaX DH10b. Gibson assembly product was transformed into MegaX DH10b (Invitrogen), and 10 clones were sequenced to check that diversity is maintained. Red X's represent stop codons.

4.5 Transformation of plasmid library into expression library

Expression of my FimH library on the tips of bacterial fimbriae requires other components of the *fim* operon[56,57], which our lab carries on either pPKL114 (containing *fimB*) or pFimΔH (containing *fimAICDFG*), depending on the cell strain. A description of each of the *fim* components can be found in Section 2.4. The K12 derivative *E. coli* strain AAEC191A has been used by our lab and our collaborators in the Sokurenko lab (Dept. of Microbiology, University of Washington) for testing fimbrial expression systems because the strain has been engineered to completely lack the *fim* operon, allowing complete control over fimbrial expression with our two-plasmid system. Our lab modified the strain by knocking out the *fliC* gene to prevent expression of flagella, which confounds certain experiments. This improved *fim* null strain, called ΔFliC, was initially chosen as the host strain for expression of my FimH library.

After I prepared electrocompetent ΔFliC cells containing pFimΔH, I optimized the voltage conditions for transformation into pFimΔH/ΔFliC by using library template plasmid (no randomization). I found an electric field strength of 17kV/cm yielded the optimum transformation efficiency at about 9 x 10^4 CFU/ml (Figure 11).

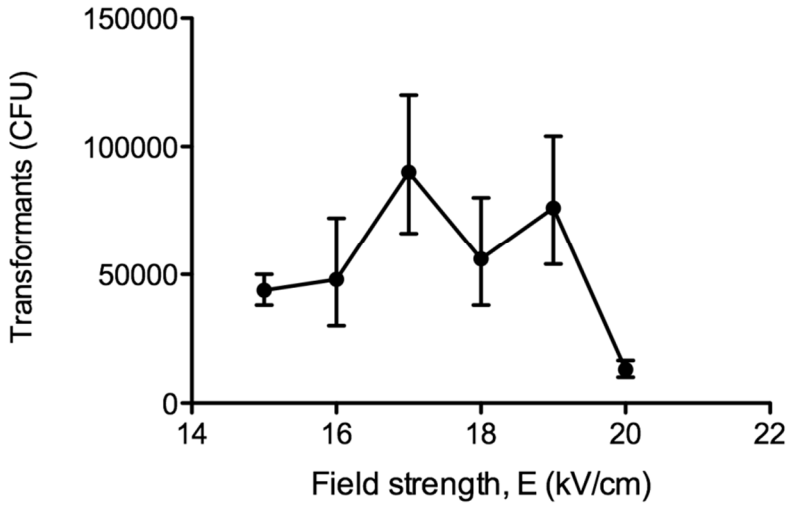

Figure 11: Optimizing voltage conditions for electroporation of pGB224-A188D control plasmid into pFimΔH/ΔFliC. Each data point was taken from a single plate, and bars indicate 95% confidence intervals calculated using Poisson statistics.

However, I observed significant batch-to-batch variation in pFimΔH/ΔFliC competent cell preps, as well as low number of transformants with my plasmid library or commercial parent plasmid pACYC184. I made many attempts to transform (by electroporation and heat shock) pACYC184 into pFimΔH/ΔFliC, but I was not able to achieve as high of CFU/ml as with commercially-derived MegaX or DH5α competent cells (Figure 12). This prompted me to consider using a strain of *E. coli* that has been commercially optimized for high DNA uptake.

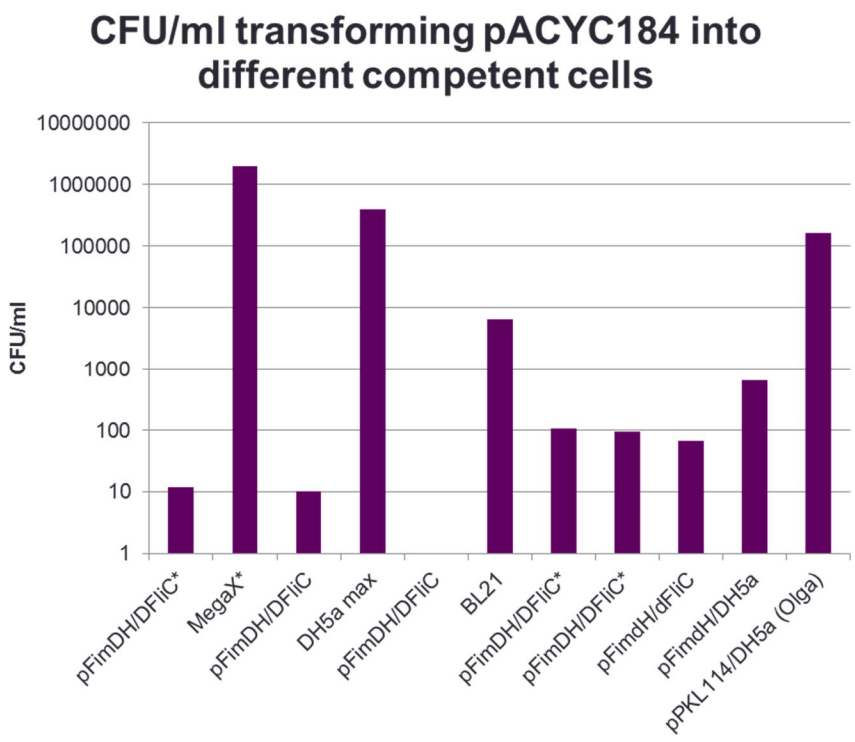

Figure 12: Transformation of control plasmid pACYC184 into different competent cells. * = transformation by electroporation. All others transformed by heat shock.

Our lab frequently uses chemically competent DH5α (Invitrogen), and when transforming the library into chemically competentized pPKL114/DH5α, I observed an increase in CFU per reaction by about 3 orders of magnitude over transformation into ΔFliC. Similarly, I obtained an increase to 102,000 transformants (95% CI = 86,500 – 120,000 transformants) in pFimΔH/MegaX, prepared from commercially electrocompetent MegaX (Invitrogen). With these numbers of transformants more than 10-fold greater than the current library size, a diversity of $4.4x10^3$ - $2.0x10^4$ total variants was maintained.

For complete control over fimbrial expression with our two-plasmid system, the host cell strain must be *fim* null. In order to determine whether DH5α and MegaX cells contain endogenous *fimH* and express it, I performed colony PCR to amplify any (if at all) genomic *fimH* present in each cell strain containing no exogenous plasmids, as well as a crystal violet assay to

detect any FimH binding to highly mannosylated RNAseB. *fimH* null K12 derivative AAEC191A was used as a negative control, and *fim*+ wild-type K12 strain MG1655 was used as a positive control. MG1655 does not require use of the two-plasmid system since it is known to already be *fim*+.

In the colony PCR, both DH5α and MG1655 exhibit a bright 1kb band (Figure 13), indicating that DH5α contains endogenous *fimH*. MegaX and AAEC191A do not show amplification of a 1kb band, suggesting that MegaX does not contain endogenous *fimH*. Additionally, comparison of MG1655 and MegaX genomes (available on NCBI), both by manual search for *fimH* on each genome and by BLAST alignment of the two strains, confirms that MegaX DH10b lacks endogenous *fimH*.

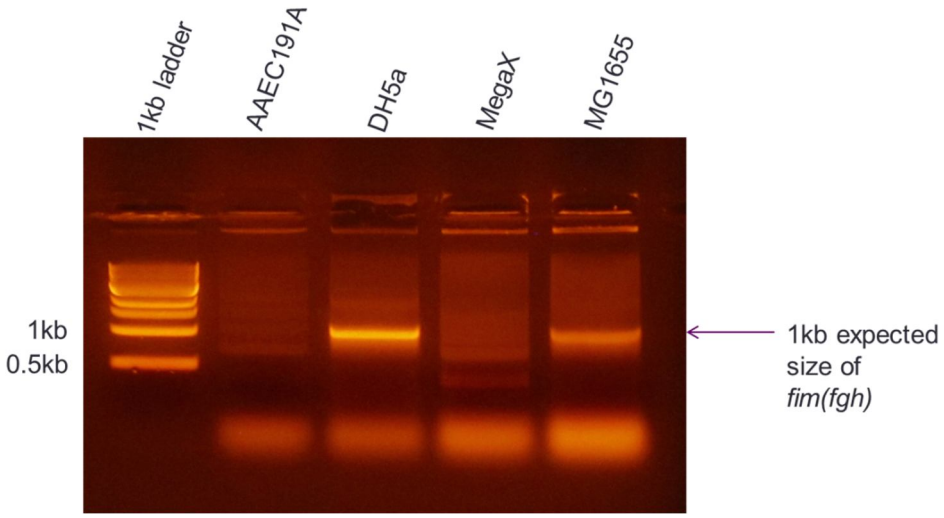

Figure 13: MegaX does not contain endogenous *fimH*. Colony PCR was done on several different cell strains using primers that amplify *fim(fgh)* in order to detect endogenous *fimh* in DH5α or MegaX.

The crystal violet assay was used as an indirect test of FimH expression by measuring cell binding to mannosylated RNAseB. Again, DH5α exhibited RNAseB binding comparable to MG1655, and addition of either pFimΔH or pPKL114 to initiate fimbrial expression significantly increased RNAseB binding (Figure 14). These results strongly suggest that DH5α express endogenous FimH. On the other hand, MegaX cells, regardless of whether they contained a

second plasmid, showed negligible RNAseB binding, comparable to negative controls and to background signal. Based on both the colony PCR and RNAseB-binding assays, I concluded that DH5α contains endogenous *fimH* and expresses FimH comparably to wild-type MG1655, and therefore DH5α cannot be used for expression of my FimH library due to contamination of my library with wild-type, endogenous FimH.

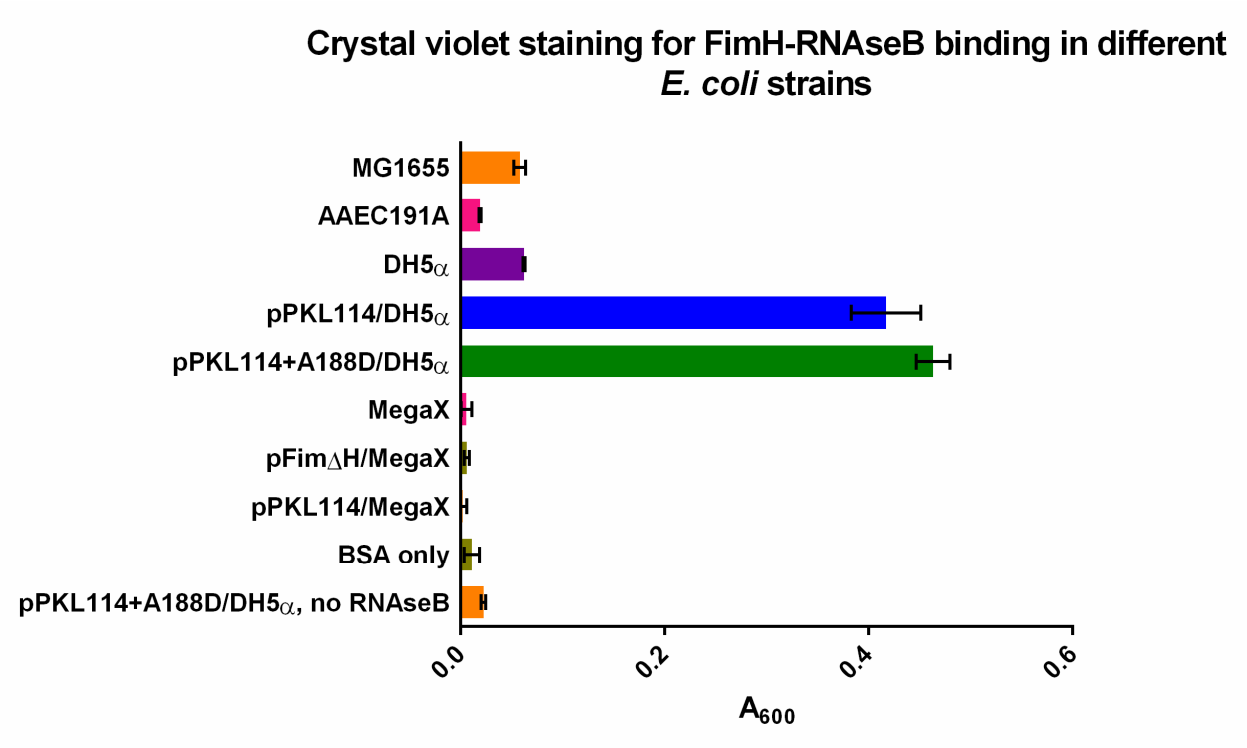

Figure 14: Crystal violet staining assay to detect FimH-RNAseB binding. Different *E. coli* strains transformed with various plasmids were incubated on plates functionalized with RNAseB and fixed for crystal violet staining. Absorbance at 600nm was measured. Error bars represent standard deviation of triplicate wells.

4.6 Analysis of library quality by flow cytometry

Since transformations into both pFimΔH/ΔFliC and pFimΔH/MegaX both maintained an estimated library size of $4.4x10^3$-$2.0x10^4$ variants (albeit pFimΔH/ΔFliC library may only be able to support the lower limit of potential library size), both libraries were carried through flow

cytometry assays to assess the degree of library expression as well as what percentage of expressors retained their high-affinity conformation.

4.6.1 Assessing library expression using αPD

To quantitate what fraction of my library was folded and expressed on the fimbrial tips of *E. coli*, I performed flow cytometry with anti-pilin domain (αPD) antibody. Since the entire CDR2 loop was randomized such that any antibody that bound to the lectin domain of FimH may no longer recognize the expressed protein, I used αPD, which is a polyclonal antibody that binds to the pilin domain of FimH, to detect FimH expression. In the formation of the fimbrial tip and the expression of FimH, the pilin domain is expressed after the lectin domain, so the use of αPD rather than an anti-lectin domain antibody gave the added advantage of not falsely detecting any truncated FimH variants. However, results of my experiments using αPD suggested that αPD was cross-reacting with proteins other than FimH, confounding the measurement of library expression. Although it is known that neither ΔFliC nor MegaX contain endogenous FimH, there was a clear αPD-positive population in cells containing only pFimΔH, suggesting some cross-reactivity of the antibody (Figure 15). Therefore, I concluded that αPD could no longer be used to measure the expression of my library.

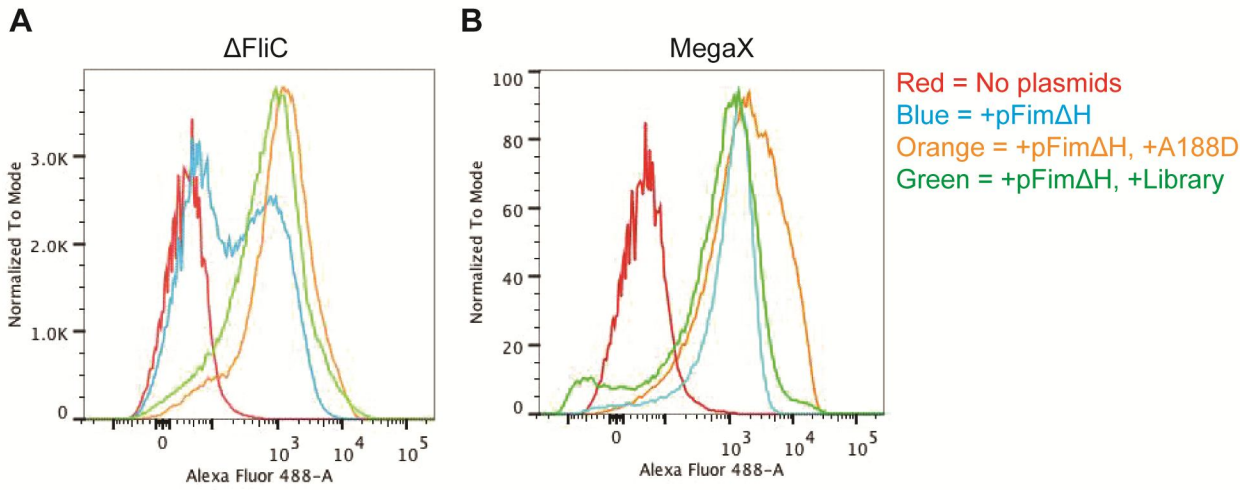

Figure 15: αPD exhibits cross-reactivity. Amplified plasmid library was transformed into (A) pFimΔH/ΔFliC or (B) pFimΔH/MegaX and stained with αPD. 100,000 events were measured per sample, and histograms were normalized to mode for easier comparison of mean fluorescence.

4.6.2 Flow cytometry for library conformation using Mab21

Although the cross-reactivity of αPD prevented accurate measurement of percent library expression, Mab21 was used to gain information about the fraction my library that expressed and retained high-affinity conformation. Mab21 is a monoclonal antibody specific for the high-affinity conformation of FimH and whose epitope is buried in the interdomain space between the lectin and pilin domains, far from the binding loops[58]. Flow cytometry results with Mab21 suggested that both libraries showed some expression of the high-affinity conformation: 26.5% of ΔFliC library transformants and 15.8% of MegaX library transformants were positive for Mab21 binding (Figure 16 and Table 4). 100,000 events were collected for each sample, which oversamples each library 50-500-fold, so there is a >95% probability that every member of each library was sampled[59,60].

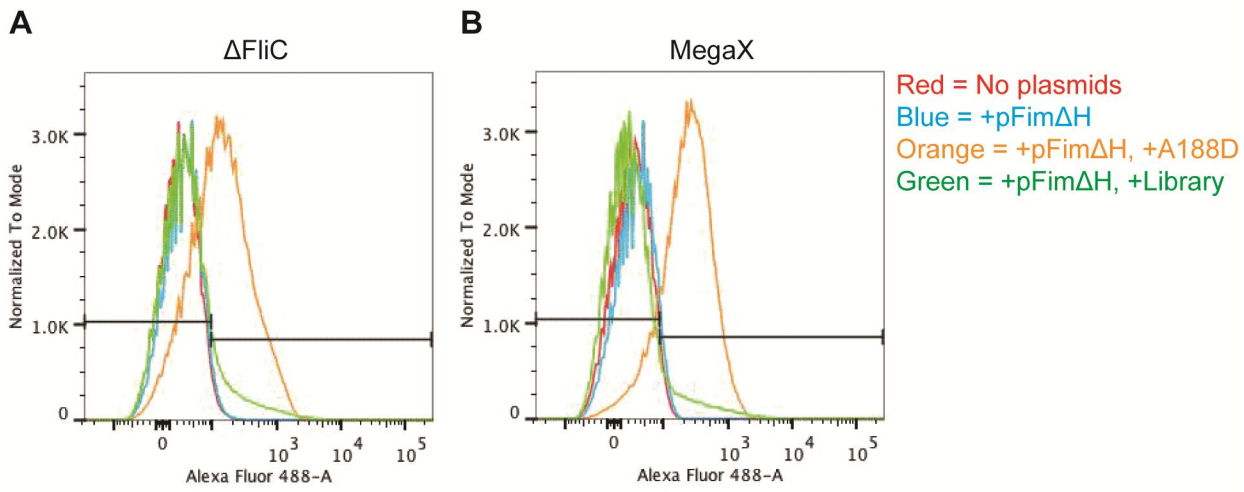

Figure 16: Mab21 binds to CDR2-randomized FimH-Hi library expressed in the high-affinity conformation. Amplified plasmid library was transformed into (A) pFimΔH/ΔFliC or (B) pFimΔH/MegaX and stained with Mab21. Brackets indicate threshold for distinguishing Mab21- and Mab21+ populations. 100,000 events were measured per sample, and histograms were normalized to mode for easier comparison of mean fluorescence.

Cell strain	% Library Mab21+	% A188D Mab21+	$\frac{\%\ \text{Library Mab21+}}{\%\ \text{A188D Mab21+}}$
ΔFliC	17.7	66.7	26.5
MegaX	13.8	87.5	15.8

Table 4: Fraction of Mab21+ library normalized to Mab21+ A188D control. Normalized percentage is indicative of the fraction of the CDR2-randomized FimH-Hi library that is expressed and is maintained in the high-affinity conformation.

5 Discussion

Overall, I achieved a functional library size of about $4.4\text{x}10^3$-$2.0\text{x}10^4$ variants in ΔFliC and MegaX DH10b cell strains, with high diversity. The critical step in optimizing library size is transformation efficiency, but once optimized, the principles behind these methods (gene assembly and Gibson assembly) could be applied to construct library randomizing all three CDR loops of FimH.

5.1 Library construction and transformation into expression library

Gene assembly with degenerate oligonucleotides successfully generated a library of high diversity with incidence of stop codons and frame-shifts not significantly higher than expected. The source of wild-type contamination is likely the PCR-amplified vector for Gibson assembly. The vector into which the CDR2 library fragment was inserted was derived from PCR amplification of the pGB224-A188D plasmid, which contains wild-type CDR2. Rather than simple PCR clean-up, which only eliminates primers and fragments of DNA <100nt, a gel purification should be done to prevent contamination of Gibson assembly product with wild-type CDR2 plasmid. Even though PCR amplification should have generated exponentially greater amounts of linear vector than wild-type CDR2 plasmid, any wild-type plasmid is likely to taken up preferentially over ligation product by bacteria at the transformation step[48,61–63]. Therefore, limiting contamination by wild-type CDR2 plasmids will further increase library diversity. That said, no other two sequences were alike, so gene assembly successfully generated high diversity in CDR2.

Gene assembly has been reported to have elevated incidence of non-targeted mutation compared to routine PCR[52], which may explain the non-wild-type CDR1/3 loops observed in two clones. KOD Hot Start Polymerase (Novagen) is recommended to decrease the incidence of non-targeted mutations[52]. In the future, it is necessary to sequence the entire vector to calculate the rate of unintended mutation. That said, according to Bessette *et al*., only a four-fold increase in degeneracy should ensure all intended randomized sequences are present in the gene assembled fragment library. Since the original gene assembly reaction used enough dNTPs to cover the potential library size 50-fold, incidence of unintended mutations should not limit my library size. Overall this method of genetic mutation to randomize all nine residues of CDR2 can

be applied to generating a plasmid library containing all three randomized CDR loops of the FimH lectin domain (see Future Work).

The main limiting steps in maintaining library size were the transformations. To amplify the plasmid library, the transformation of Gibson assembly into MegaX DH10b achieved $4.4x10^4$ transformants corresponding to transformation efficiency on the order of 10^6 CFU/µg of DNA. The positive control (pUC19) had a transformation efficiency of $8.2x10^9$, which is almost 1.5-fold lower than the lower bound of the manufacturer's reported efficiency[63], suggesting that there is some room for improvement of this transformation. Although one needs to calculate transformation efficiency (CFU/µg DNA) to compare the results to the manufacturer's claims, this is not the best measurement of success for library construction. Instead, number of transformants should be optimized, and this generally increases with increasing concentrations of DNA, which is not always the case with transformation efficiency[64–67]. For example, in one of the earliest attempts to transform Gibson product into MegaX, 43.1ng of Gibson assembly transformed with highest efficiency into electrocompetent MegaX, but 431ng of Gibson assembly product actually resulted in 5-fold higher number of transformants (data not shown). Unfortunately, at that time, it was not known to optimize number of transformants, and subsequent transformations used a maximum of 43.1ng of Gibson assembly. Therefore, for repeating these methods in creating a 3-loop diversity library, transformation efficiency, frequency (transformants/survivors), and most importantly number of transformants should be considered when evaluating optimum transformation conditions. Another possible point of improvement is to inactivate any ligases in the Gibson assembly reaction that may be interfering with transformation[61,64,68]. A bulletin from BioRad reports that heat inactivation of ligase for 15min at 65°C increases transformation efficiency 2-5-fold[61]. Although it remains to be seen

whether this would optimize our transformation, if DNA concentration and cell concentration were held constant, an increase in transformation efficiency would correspond to an equivalent increase in the number of transformants.

For the second transformation of the amplified plasmid library into electrocompetent pFimΔH/ΔFliC or pFimΔH/MegaX, I believe that the main limitation to number of transformants achieved was the competent cell preparation. I cannot evaluate my competent cell preparations with pUC19 control plasmid because pUC19 contains the same antibiotic resistance gene as pFimΔH, making it impossible to distinguish between transformants and non-transformants. But when comparing my maximum transformation efficiency to others electrocompetentizing MegaX, there may be room for improvement. Wu *et al*. reported a maximum transformation efficiency of $1.5x10^9$ CFU/µg DNA in MegaX DH10b using several optimizations: (1) starting preparation with cells grown to early log phase OD600 = 0.15 rather than 0.35-0.4; (2) concentrating cells by three washes with 10% glycerol rather than two washes with water and one wash with 10% glycero1; (3) final concentration of electrocompetent cells should be $0.5\text{-}0.6x10^{10}$ cell/ml[69]. Importantly, all of their experiments electroporated using a field strength of 12.5kV/cm rather than 20kV/cm, which is recommended by BioRad for commercially prepared MegaX DH10b, so perhaps these conditions would not be optimum for a higher electroporation field strength. Also, these conditions were optimized for maximum transformation efficiency, which may not actually improve total number of transformants, but these suggestions may provide insight on what conditions can be considered for optimization of electrocompetent MegaX cell preparation.

The BioRad Gene Pulser II Manual and should be consulted for general principles of electroporation[70], and Dower *et al*. (1988) present a thorough analysis of how electroporation

conditions affect transformation efficiency, transformation frequency, and number of transformants[64]. Based on my current knowledge and experience, I hypothesize that increasing the competent cell concentrations and DNA concentrations will increase the number of transformants (and potentially decrease the transformation efficiency), thereby increasing the library size.

One last suggestion for competent cell preparation is to prepare large batches of electrocompetent cells, because there is considerable variability between cell batches[64], so some re-optimization may be required for each batch of competent cells. Finally, it should be noted that Lucigen can prepare custom competent cells for library applications[71], price upon request.

5.2 Characterization of library expression and conformation

Based on the library construction methods used, I expected that 10-55% of the CDR2 variants generated would not contain stop codons or frame-shifts and could theoretically fold and express on the fimbrial tips. My flow cytometry experiments indicated that 26.5% of my pFimΔH/ΔFliC library and 15.8% of my pFimΔH/MegaX library were expressed in the high-affinity conformation. Since these observed percentages are within the range of expected expression based on library construction methods, it is possible that all functional variants in my library were successfully expressed and remained locked in the high-affinity conformation despite having the entire CDR2 loop randomized. However, it is impossible to know whether the high-affinity variants detected constitute a minority or a majority of the total expressed library. To parse the problem, I attempted to measure the percentage of the library expressed, regardless of conformation, using αPD. Unfortunately, polyclonal antibody αPD bound to cells containing pFimΔH without *fimH* of any form, suggesting that αPD may be cross-reacting with something encoded on pFimΔH (such as FimA), or induced to express by pFimΔH. Our lab is

investigating monoclonal antibodies that bind FimH in a conformation-independent manner to probe expression of the library.

6 Conclusions and Future Work

In summary, I created a highly diverse library of $4.4x10^3$-$2.0x10^4$ variants by randomizing 9 amino acids in the CDR2 loop of FimH-Hi. The gene assembly and Gibson methods were successful in generating high diversity with incidence of stop codons and frame-shifts not significantly higher than expected. The results presented here demonstrate that these methods can be used to randomize all three CDR loops on the FimH-Hi lectin domain (see Future Work), and a clear strategy exists for optimizing library creation through the improvement of transformations. A significant percentage of the library is expressed and maintained in the high-affinity conformation, suggesting that FimH-Hi is a good protein scaffold because it retains its high-affinity conformation despite randomization in the lectin domain. Moreover, a single point mutation can revert FimH-Hi to the wild-type, activatable conformation, generating a theoretically activatable protein scaffold. Next steps include sorting the library using FACS with Mab21 to isolate and grow up only the fraction of the library that expresses a FimH variant in the high-affinity conformation. Further conditioning of this library includes depleting non-specific binders and mannose-binders to obtain a bacterial display library of high-affinity FimH mutants that no longer bind mannose.

6.1 Proposed molecular biology strategies for randomizing all three CDR loops

To achieve randomization of all three CDR loops, there are many molecular biology options. The strategies presented here are based on the methods and results reported in this thesis. The best strategy will maximize diversity, minimize stop codons, have a high probability of success, and be cost- and time-efficient. Here, I propose several strategies that I think best fit

these requirements, and discuss the advantages and disadvantages of each method as well as alternatives.

The initial strategy for generating a library randomizing all CDR loops utilized gene assembly to fully synthesize a 471bp sequence encoding the FimH lectin domain using 8 overlapping oligonucleotides. Of these 8 oligonucleotides, 3 contained NNK regions (3 corresponding wild-type oligonucleotides are also available). The first attempt at generating a library using gene assembly with these 8 oligonucleotides yielded a high incidence of stop codons; 10 codons were sequenced, and all 10 contained at least one stop codon in any of the CDR regions (data not shown). Although 100% of clones sequenced contained a stop codon, this percentage is not significantly higher than the expected occurrence of 56% ($p>0.1$). Nonetheless, in light of this preliminary data, it may be optimum to use the existing CDR2 library developed in this thesis, after conditioning by several assays including fluorescence activated cell sorting (FACS) to eliminate non-expressing clones, as a starting point from which to generate randomization into CDRs 1 and 3. This would decrease the expected incidence of stop codons from 56% to 42%.

The first option, and the one I believe is optimal based on my experience with these methods, is to use the gene assembly oligonucleotides previously ordered to generate short fragments of randomized CDR1 and CDR3 separately, followed by insertion into FimH vector pGB224-A188D with CDR2 library fragment (subcloned from conditioned CDR2 library reported in this thesis) via Gibson assembly. Important design criteria include ensuring that the oligonucleotides used will synthesize fragments that overlap by 16-40bp for subsequent Gibson assembly. The amplification step done after gene assembly in this thesis may not be necessary here if initial gene assembly product results in a single, clear band at expected length. However,

a smear on the gel is expected with gene assembly reactions with multiple overlapping oligonucleotides, since not all oligonucleotides will anneal into full-length product with 100% efficiency[52]. Therefore, the advantage of the amplification step is (1) to amplify any full-length product formed in the initial reaction, and (2) to allow a second chance for oligonucleotides to anneal into full-length product. The reason this is accomplished with two PCR reactions rather than a single reaction with more cycles is to minimize library bias resulting from amplification of only a few variants. After generation of CDR2 fragment library by PCR, all three CDR fragment libraries can be combined with PCR-generated pGB224-A188D vector by Gibson assembly. Traditional restriction enzyme cloning may also be used if unique restriction cut sites exist between CDRs, although NNK randomization may introduce restriction cut sites impossible to predict. That said, to prepare for either strategy, amplification primers should be designed to accommodate restriction enzyme cloning (if possible) and Gibson assembly. To allow for either strategy, the Gibson assembly overlap regions should span each restriction enzyme cut site. If no cut site exists, only Gibson assembly specifications need to be considered. Gibson assembly necessitates 16-40bp overlap, with a melting point greater than 48°C to ensure specificity[62].

The advantages of using this first method is that CDRs 1 and 3 are randomized and combined with CDR2 mixture that has no stop codons, decreasing the frequency of stop codons in the total 3-loop library. This method also uses existing oligonucleotides (although a few additional amplification primers may be necessary), saving time and cost. However, there are two major differences between this method and the one used in this thesis for generating the CDR2 library: (1) subcloning the CDR2 library fragment, and (2) using Gibson to assembly four fragments together rather than just two. The latter is not expected to cause any problems since the Gibson Assembly kit is designed for ligating up to 6 fragments regardless of fragment

length[62,72,73]. But in the event that Gibson assembly does not successfully ligate all 4 fragments, there are several different GeneArt® fragment assembly kits available from Life Technologies that may serve as alternative methods for about $15-$43 per reaction[74]. Subcloning of the CDR2 library fragment from the library described in this thesis may present several challenges that could potentially decrease library size. First, the probability of any variant in the original plasmid library being present in a sub-cloned library is decreased such that ensuring propogation of total diversity may require the number of transformants to be far greater than the number of transformants in the original library[75]. A more detailed discussion on improving transformations can be found in the Discussion of this thesis. Second, if the CDR2 fragment library is PCR-generated for Gibson assembly, this additional PCR may result in library bias.

To avoid having to ligate 4 fragments simultaneously, sequential insertion of CDR1 and CDR3 into PCR-generated CDR2-containing vectors can be used to obtain fully randomized library. The potential disadvantage of this method is that multiple CDR2-containing vectors will need to be PCR-generated.

Gene shuffling is a potential alternative to avoid the need to excise or sub-clone CDR2 out of its plasmid. One possible strategy employing gene shuffling is to synthesize the full-length (471bp) sequence with CDRs 1 and 3 randomized simultaneously (leaving CDR2 wild-type), insert the mixture of 471bp sequences into pGB224-A188D, and employ gene shuffling with CDR2 plasmid library. This eliminates the need for sub-cloning sub-cloning the CDR2 fragment library out of its plasmid. Gene shuffling consists of digestion by DNAse followed by a PCR reaction to allow crossover between templates. The goal is to basically have entire CDR regions swap from plasmid to plasmid in order to obtain a plasmid library with all three CDR loops randomized. Whereas this is obtained by careful cloning or Gibson assembly in previously

mentioned methods, gene shuffling is quick and requires little library construction expertise[40,41,51]. However, gene shuffling does not allow the same extent of control over shuffling, and one of the main disadvantages of gene shuffling is the potential for no recombination to occur and therefore significant incidence of wild-type sequences. To address both of these challenges, the rate of recombination is controlled mainly by the choice of what size fragments to isolate and recombine after DNAse digestion[41].

The last potential strategy discussed here is to repeat all of the methods presented in this thesis with CDR1 and CDR3 separately, followed by conditioning of the library to achieve three separate libraries, each with a different CDR randomized, and each one conditioned such that only variants expressed in the high-affinity conformation and lack mannose- and non-specific binding are retained. At this point, each plasmid library can be isolated and combined together by gene shuffling, cloning, or any of the previously mentioned methods. The advantage of this strategy is that by the time the three randomized CDR loops are combined, none of the variants have stop codons or frame-shifts, and all of the variants can be expressed and retain their high-affinity conformation. The disadvantage of this method is that having to do all of the transformations for each library in parallel is not only time consuming and resource-intensive, but if not all transformations are of high enough efficiency, this method could also provide many more bottle-necks in library size.

7 Appendix

Table 5: Cell strains used in this thesis.

Cell strain	*fim* genotype	Antibiotic resistance	Uses in this thesis
ΔFliC	*fim*-	Kanamycin	Library expression
MegaX DH10b	*fim*-	-	Plasmid library amplification Library expression
DH5α	Wild-type *fim*+	-	Attempted library expression

XL1b	N/A	Tetracycline	Testing Gibson assembly by blue/white selection
TOP10	N/A	-	Testing gene assembly

Table 6: Plasmids used in this thesis. MCS = multiple cloning site

Plasmid name	MCS Genotype	Antibiotic resistance	Use in this thesis
pGB224-A188D	*fimH*-A188D	Chloramphenicol	Library randomization
pFimΔH	*fim(AICDFG)*	Ampicillin, kanamycin	Fimbrial expression
pPKL114	*fimB*	Ampicillin	Fimbrial expression
pCR®-Blunt	Used as a vector	Kanamycin, zeocin	Gibson check

Table 7: Antibodies used in this thesis.

Antibody	Type	Target	Raised against
αPD	Rabbit polyclonal	Pilin domain	Periplasmic pilin domain isolate
Mab21	Mouse monoclonal	High affinity conformation FimH	Purified high-affinity FimH

8 Works Cited

1. Golden, A. L. *et al.* Simple fluidic system for purifying and concentrating diagnostic biomarkers using stimuli-responsive antibody conjugates and membranes. *Bioconjugate chemistry* **21**, 1820–6 (2010).

2. Nash, M. a, Waitumbi, J. N., Hoffman, A. S., Yager, P. & Stayton, P. S. Multiplexed enrichment and detection of malarial biomarkers using a stimuli-responsive iron oxide and gold nanoparticle reagent system. *ACS nano* **6**, 6776–85 (2012).

3. Binz, H. K., Amstutz, P. & Plückthun, A. Engineering novel binding proteins from nonimmunoglobulin domains. *Nature biotechnology* **23**, 1257–68 (2005).

4. Guiochon, G. & Beaver, L. A. Separation science is the key to successful biopharmaceuticals. *Journal of chromatography. A* **1218**, 8836–58 (2011).

5. Ly, L. & Wasinger, V. C. Protein and peptide fractionation, enrichment and depletion: tools for the complex proteome. *Proteomics* **11**, 513–34 (2011).

6. Jayson, G. C. *et al.* Molecular imaging and biological evaluation of HuMV833 anti-VEGF antibody: implications for trial design of antiangiogenic antibodies. *Journal of the National Cancer Institute* **94**, 1484–93 (2002).

7. Koide, S. Design and Engineering of Synthetic Binding Proteins Using Non-Antibody Scaffolds. *Molecular Biology*

8. Gebauer, M. & Skerra, A. Engineered protein scaffolds as next-generation antibody therapeutics. *Current opinion in chemical biology* **13**, 245–55 (2009).

9. Gill, D. S. & Damle, N. K. Biopharmaceutical drug discovery using novel protein scaffolds. *Current opinion in biotechnology* **17**, 653–8 (2006).

10. Skerra, A. Engineered protein scaffolds for molecular recognition. (2000).

11. Skerra, A. Alternative non-antibody scaffolds for molecular recognition. *Current opinion in biotechnology* **18**, 295–304 (2007).

12. Binz, H. K. & Plückthun, A. Engineered proteins as specific binding reagents. *Current opinion in biotechnology* **16**, 459–69 (2005).

13. Koide, a, Bailey, C. W., Huang, X. & Koide, S. The fibronectin type III domain as a scaffold for novel binding proteins. *Journal of molecular biology* **284**, 1141–51 (1998).

14. Xu, L. *et al.* Directed Evolution of High-Affinity Antibody Mimics Using mRNA Display. *Chemistry & Biology* **9**, 933–942 (2002).

15. Richards, J. *et al.* Engineered Fibronectin Type III Domain with a RGDWXE Sequence Binds with Enhanced Affinity and Specificity to Human αvβ3 Integrin. *Journal of Molecular Biology* **326**, 1475–1488 (2003).

16. Karatan, E. *et al.* Molecular Recognition Properties of FN3 Monobodies that Bind the Src SH3 Domain. *Chemistry & Biology* **11**, 835–844 (2004).

17. Nygren, P.-A. Alternative binding proteins: affibody binding proteins developed from a small three-helix bundle scaffold. *The FEBS journal* **275**, 2668–76 (2008).

18. Nord, K. *et al.* Binding protein selected from combinatorial libraries of an a-helical bacterial receptor domain. *Nature biotechnology* **15**, 772–777 (1997).

19. Skerra, A. Alternative binding proteins: anticalins - harnessing the structural plasticity of the lipocalin ligand pocket to engineer novel binding activities. *The FEBS journal* **275**, 2677–83 (2008).

20. Beste, G., Schmidt, F. S., Stibora, T. & Skerra, a Small antibody-like proteins with prescribed ligand specificities derived from the lipocalin fold. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 1898–903 (1999).

21. Kim, H. J., Eichinger, A. & Skerra, A. High-affinity recognition of lanthanide(III) chelate complexes by a reprogrammed human lipocalin 2. *Journal of the American Chemical Society* **131**, 3565–76 (2009).

22. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, A. Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins. *Journal of Molecular Biology* **332**, 489–503 (2003).

23. Miller, M. K. *et al.* The Muscle Ankyrin Repeat Proteins: CARP, ankrd2/Arpp and DARP as a Family of Titin Filament-based Stress Response Molecules. *Journal of Molecular Biology* **333**, 951–964 (2003).

24. Zahnd, C. *et al.* A designed ankyrin repeat protein evolved to picomolar affinity to Her2. *Journal of molecular biology* **369**, 1015–28 (2007).

25. Lee, S., Xie, J. & Chen, X. Activatable Molecular Probes for Cancer Imaging. *Curr Top Med Chem* **10**, 1135–1144 (2011).

26. Kobayashi, H. & Choyke, P. L. Target-Cancer-Cell-Specific Activatable Fluorescence Imaging Probes: Rational Design and in Vivo Applications. *Accounts of chemical research* **44**, 83–90 (2011).

27. Stearns, D. J., Kurosawa, S., Sims, J., Esmon, N. L. & Esmon, C. T. The Interaction of a Ca2 + -dependent Monoclonal Antibody with the Protein C Activation Peptide Region. *Journal of Biological Chemistry* **263**, 826–832 (1988).

28. Arad-Yellin, R. & Green, B. Photochemical closing and opening of the guest-binding cavity of cyclodextrins. *Nature letters* (1994).

29. Self, C. H. & Thompson, S. Light activatable antibodies: Models for remotely activatable proteins. *Nature* (1996).

30. Sigma-Aldrich *Material Safety Data Sheet: Trichloromethyl chloroformate*. 1–7 (2013).

31. Lewandowski, A. T., Small, D. A., Chen, T., Payne, G. F. & Bentley, W. E. Tyrosine-Based "Activatable Pro-Tag": Enzyme-Catalyzed Protein Capture and Release. *Biotechnology and bioengineering* **93**, (2006).

32. Wu, H.-C. *et al.* Biofabrication of antibodies and antigens via IgG-binding domain engineered with activatable pentatyrosine pro-tag. *Biotechnology and bioengineering* **103**, 231–40 (2009).

33. Schwan, W. R. Regulation of fim genes in uropathogenic Escherichia coli. *World J Clin Infect Dis* **1**, 17–25 (2011).

34. Yakovenko, O. *et al.* FimH forms catch bonds that are enhanced by mechanical force due to allosteric regulation. *The Journal of biological chemistry* **283**, 11596–605 (2008).

35. Aprikian, P. *et al.* Interdomain interaction in the FimH adhesin of Escherichia coli regulates the affinity to mannose. *The Journal of biological chemistry* **282**, 23437–46 (2007).

36. Choudhury, D. X-ray Structure of the FimC-FimH Chaperone-Adhesin Complex from Uropathogenic Escherichia coli. *Science* **285**, 1061–1066 (1999).

37. Hung, C.-S. *et al.* Structural basis of tropism of Escherichia coli to the bladder during urinary tract infection. *Molecular microbiology* **44**, 903–15 (2002).

38. Wellens, A. *et al.* Intervening with urinary tract infections using anti-adhesives based on the crystal structure of the FimH-oligomannose-3 complex. *PloS one* **3**, e2040 (2008).

39. Schembri, M. a, Kjaergaard, K., Sokurenko, E. V & Klemm, P. Molecular characterization of the Escherichia coli FimH adhesin. *The Journal of infectious diseases* **183 Suppl** , S28–31 (2001).

40. Lipovšek, D., Mena, M., Shaun, M., Basu, S. & Baynes, B. M. Chapter 4: Library Construction for Protein Engineering. *Protein Engineering and Design* 1–27 (2010).

41. Neylon, C. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic acids research* **32**, 1448–59 (2004).

42. Braunagel, M. & Little, M. Construction of a semisynthetic antibody library using trinucleotide oligos. *Nucleic acids research* **25**, 4690–1 (1997).

43. Virnekäs, B. *et al.* Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic acids research* **22**, 5600–7 (1994).

44. Steffens, D. L. & Williams, J. G. K. Efficient Site-Directed Saturation Mutagenesis Using Degenerate Oligonucleotides. *Nucleic Acids Research* **18**, 147–149 (2007).

45. Stratagene QuikChange ® Site-Directed Mutagenesis Kit. **200518**,

46. Kunkel, T. a, Roberts, J. D. & Zakour, R. a Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods in enzymology* **154**, 367–82 (1987).

47. Handa, P., Thanedar, S. & Varshney, U. Rapid and Reliable Site-Directed Mutagenesis Using Kunkel ' s Approach. 1–6

48. Huang, R., Fang, P. & Kay, B. K. Improvements to the Kunkel mutagenesis protocol for constructing primary and secondary phage-display libraries. *Methods (San Diego, Calif.)* **58**, 10–7 (2012).

49. Pai, J. C., Entzminger, K. C. & Maynard, J. a Restriction enzyme-free construction of random gene mutagenesis libraries in Escherichia coli. *Analytical biochemistry* **421**, 640–8 (2012).

50. Prodromou, C. & Pearl, L. H. Recursive PCR: a novel technique for total gene synthesis. *Protein engineering* **5**, 827–829 (1992).

51. Stemmer, W. P., Crameri, a, Ha, K. D., Brennan, T. M. & Heyneker, H. L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164**, 49–53 (1995).

52. Bessette, P., Mena, M. A., Nguyen, A. W. & Daugherty, P. S. Construction of Designed Protein Libraries Using Gene Assembly Mutagenesis. *Directed evolution library creation: Methods and protocols* 29–37 (2003).

53. Hoover, D. M. & Lubkowski, J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic acids research* **30**, e43 (2002).

54. Grönwall, C. *et al.* Selection and characterization of Affibody ligands binding to Alzheimer amyloid beta peptides. *Journal of biotechnology* **128**, 162–83 (2007).

55. Silverman, J. *et al.* Multivalent avimer proteins evolved by exon shuffling of a family of human receptor domains. *Nature biotechnology* **23**, 1556–61 (2005).

56. Klemm, P. Two regulatory fim genes, fimB and fimE, control the phase variation of type 1 fimbriae in Escherichia coli. *EMBO Journal* **5**, 1389–1393 (1986).

57. Klemm, P. & Schembri, M. a Fimbrial surface display systems in bacteria: from vaccines to random libraries. *Microbiology (Reading, England)* **146 Pt 12**, 3025–32 (2000).

58. Tchesnokova, V. *et al.* Type 1 Fimbrial Adhesin FimH Elicits an Immune Response That Enhances Cell Adhesion of Escherichia coli Type 1 Fimbrial Adhesin FimH Elicits an Immune Response That Enhances Cell Adhesion of Escherichia coli †. *Infection and Immunity* **79**, 3895–3904 (2011).

59. Levay-young, B. *et al.* Library Format for Bioengineering Maximizing Screening Efficiency through Good Design. **33**, 0–2 (2013).

60. Ostermeier, M. Designing switchable enzymes. *Current opinion in structural biology* **19**, 442–8 (2009).

61. Zoller, P. High Electro-transformation Efficiencies Obtained With DNA From Ligation Mixtures. *BioRad* 18–19 (2000).

62. Gibson Assembly ™ Master Mix. *New England Biolabs*

63. Invitrogen MegaX DH10B ™ T1. 3–6 (2006).

64. Dower, W. J., Miller, J. F. & Ragsdale, C. W. High efficiency transformation of E. coli by high voltage electroporation. *Nucleic Acids Research* **16**, 6127–6145 (1988).

65. Dower, W. J. *Genetic Engineering: Principles and Methods*. 275–296 (Plenum Press: New York, 1990).

66. Miller, J. F., Dowert, W. J. & Tompkins, L. S. High-voltage electroporation of bacteria : Genetic transformation of Campylobacterjejuni with plasmid DNA. **85**, 856–860 (1988).

67. Calvin, N. M. & Hanawalt, P. C. High-Efficiency Transformation of Bacterial Cells by Electroporation. *Journal of bacteriology* **170**, 2796–2801 (1988).

68. Ymer, S. Heat inactivation of DNA ligase prior to electroporation increases transformation efficiency. *Nucleic acids research* **19**, 6960 (1991).

69. Wu, N., Matand, K., Kebede, B., Acquaah, G. & Williams, S. Enhancing DNA electrotransformation efficiency in Escherichia coli DH10B electrocompetent cells. *Electronic Journal of Biotechnology* **13**, 1–9 (2010).

70. Biorad Gene Pulser Xcell Electroporation System Instruction Manual.

71. Lucigen: Simplifying Genomics. at <http://lucigen.com/store/Competent-Cells-for-Cloning/>

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

73. Gibson, D. G., Smith, H. O., Iii, C. A. H., Venter, J. C. & Merryman, C. Chemical synthesis of the mouse mitochondrial genome. *Nature methods* **7**, 901–903 (2010).

74. Seamless Cloning and Genetic Assembly. *Life Technologies* at <http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Cloning/Seamless-Cloning-and-Genetic-Assembly.html?s_kwcid=AL!3652!3!23373674241!b!!g!!+gene+assembly&ef_id=UdjGdAAABEfi59wV:20130716042318:s>

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

76. Lipovsek, D. & Plückthun, A. In-vitro protein evolution by ribosome display and mRNA display. *Journal of immunological methods* **290**, 51–67 (2004).