Integrating Gene Synthesis and MITOMI for Rapid Protein Engineering

THÈSE N^{O} 6796 (2016)

PRÉSENTÉE LE 12 FÉVRIER 2016

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE CARACTÉRISATION DU RÉSEAU BIOLOGIQUE
PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Matthew Christopher BLACKBURN

acceptée sur proposition du jury:

Prof. M. Lütolf, président du jury
Prof. S. Maerkl, directeur de thèse
Prof. R. Aebersold, rapporteur
Prof. A. de Mello, rapporteur
Prof. B. Correia, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2016

Acknowledgements

Sometime during the end of my bachelor's degree, back when I was settling into academic research and considering what my next professional/educational step would be, some sage graduate student told me 'a PhD isn't a sprint, it's a marathon'. While some marathoners can finish gracefully, primarily owing to their diligence in training and self-motivation, others will cross the line entirely exhausted but floating on the support and encouragement of their cheering friends and family. Over the past few years, I have been fortunate to be surrounded by such an inspiring and comforting mix of colleagues, friends and family, and I am certain I wouldn't have started nor finished the arduous journey of obtaining a PhD had I been without them.

I would like to first thank Prof. Sebastian Maerkl for all of his effort and patience in pushing me to become a better scientist and engineer, but also for giving me the unique opportunity to live and work abroad for the first time in my life. Moving to Switzerland was a life-changing event for me, and by having the concurrent challenge of integrating within a foreign country while juggling the responsibilities of a PhD project, I feel I have matured into a better and more capable person. I would also like to specially thank Marcel Geertz, who was my first mentor when I joined the lab, but also introduced me to many cycling routes in Suisse-Romande when I first arrived. I'd also like to thank all of the past members of the lab "family" for answering all of my microfluidic and MATLAB questions, introducing me to EPFL and Swiss life, as well as for all of the social outings: Nicolas Dénervaud, Luis Miguel Fidalgo, José Luis Garcia, J-B Nobs, Sylvie Rockel, and Arun Rajkumar. The newer members of the lab have also been great friends and colleagues, and I've appreciated having such a positive lab environment to work in.

During my first trip to Switzerland for the EPFL PhD Hiring Days, which was my first voyage to Europe, I met several people, many of whom are among my closest friends. Our circle of friends has slowly expanded as we all took the leap into our PhDs at roughly the same time and formed a special bond through numerous evenings spent at Satellite, celebrating various life events at each other's apartments, and hiking/traveling around

Switzerland: Henrike Niederholtmeyer, Jeff Kasten and Ksenya Schors, Dom Monteil, Klas and Elin Kronander, Sebastian Waszak, Maryna Babayeva, Clara Moldovan, Alina Isakova, Mariia Kaliuzhna, and Kevin Leempoel. I have to express special thanks to Henrike Niederholtmeyer, who in addition to being an early friend from the Hiring Days, then a supportive and smiling lab mate, and for nearly 4 years also a fun flat mate, we traveled together on several occasions and shared many memorable events (surprise 30th birthday party for example). Also, I need to specially thank Alina Isakova for her suggestions and insight during my research project, but also for helping to facilitate my 4-month 'rental' of the Deplancke lab's microarray scanner so that I could finish collecting all of the data for my project using the same machine (after our lab's scanner crashed and never woke up again).

Through rock climbing, hiking, skiing and ski touring events with the Club Montagne and the Sports Center at EPFL, I became close friends with Benedikt Fasel, Maren Arp, Gunnar Steinberg, Douglas Watson, Mandana Sarikhani, María Ámundadóttir and Jon Agustsson. They have also helped make Switzerland feel more like a second home, and encouraged me to get outdoors more frequently.

Outside of the lab group, one of my first Swiss friends after I moved to Lausanne was Sarah El-Achachi, who began as my French language tandem, and morphed into a close friend. Jérôme Crittin has also been a great friend and lent much support and advice to me over the past few years. One person who has also had a great impact on my feeling 'settled' in Switzerland is Ian Rousseau, who I first met through the MIT Cycling club in 2008, but grew to be much closer friends with since he relocated to Europe at nearly the same moment I did. He also introduced me to a whole other group of amicable, intelligent and charming friends who live in Zurich, with whom I've shared many memories from Zurich Street Parade, skiing in Graubunden, and hiking everywhere in between: Remo Ughini, Ursin Hutter, Max and Linda Ahnen, Lea Nowack, Karel Steurs, and Frances Erb.

I've spent much time thanking friends and colleagues, who've been my second, adopted European family since I moved from the USA, but they are definitely not a substitute for real family. I'd like to thank my parents, and also my brother and sister for being supportive of my decision to live and study abroad, but also for my parents' support and love throughout my bachelors and masters degrees, prior to leaving for Switzerland. Although I'm not entirely sure they understood everything I explained to them regarding my thesis topic, they were always sympathetic and eager to listen to my triumphs and failures. I'm also thankful for the love, patience, and support of my partner, Michaël Dollone, and all of his family, who have been extremely welcoming and hospitable towards me, even at the beginning of our relationship when I was struggling to speak French with them.

Matthew Blackburn
Lausanne, December 5, 2015

Abstract

The research described within this thesis is primarily motivated by two fields of science with reversed, yet complementary, approaches to addressing the same essential objective: manipulating and understanding the complex program of life. Whereas *synthetic biology* has to do with the bottom-up construction of living systems from a library of 'parts', *systems biology* takes a top-down, reverse-engineering analysis of the same processes within functioning cells to elucidate the crucial elements and interactions. Both perspectives have led to important contributions in learning how biological systems operate. Discoveries from these studies can have far-reaching implications, notably in medicine, manufacturing, energy, and the environment (1).

One exciting outcome of the research efforts within synthetic biology is the ability to design genetic constructs encoding proteins with novel, optimized functions. Although advances have been made in predicting how a protein's amino acid sequence relates to its higher order structural form and behavior, experimental methods for protein engineering remain invaluable for evaluating computationally derived designs. Unfortunately, the design-build-test cycle for rational protein development is commonly a time, labor and resource intensive endeavor. This is primarily due to limitations in the ability to generate libraries of gene variants and bottlenecks in cell-based expression and quantitative screening of protein products.

This thesis describes the development of a solid-phase gene assembly technique and its integration with microfluidic protein analysis to accelerate the prototyping of novel protein designs. The gene assembly method utilizes commodity-scale, chemically manufactured DNA, and operates without the need for ligase or restriction enzymes. As a bench top process amenable to scale up, the modularity of the assembly process allows the efficient and economical generation of expression-ready gene variants. We directly coupled this gene assembly pipeline to a microfluidic device to enable high-throughput, on-chip, cell-free protein expression, purification, and characterization. By circumventing molecular cloning and cell-based steps, the lag time between protein design and quantitative analysis was dramatically reduced.

As a proof of concept, this protein engineering platform was applied towards the construction of over 400 artificial, engineered variants of C_2H_2 zinc finger (ZF) proteins. ZF protein domains are the most prevalent form of transcription factor (TF) in humans, and are found throughout the tree of life. They provide a convenient structure for refactoring due to their relatively small size and composability, and are an ideal model for exploring the biophysics of TF-DNA specificity. The ability to engineer ZF proteins makes them useful as programmable, DNA-targeting units for applications in biotechnology and synthetic biology. We demonstrate that although ZFPs can be readily engineered to recognize a particular DNA target, engineering the precise binding energy landscape remains a challenge. Additionally, we show that ZF-DNA binding affinity can be tuned independently of sequence specificity. Together, these results demonstrate the versatility of the coupled gene assembly and microfluidic analysis platform as a new tool for rational protein engineering.

Keywords: transcription factor; zinc finger protein; microfluidics; MITOMI; high-throughput; cell-free protein synthesis; gene assembly; protein engineering; binding affinity; synthetic biology; DNA binding domain

Résumé

La recherche décrite dans cette thèse se fonde principalement sur deux champs scientifiques. Bien qu'opposés, ils sont complémentaires dans leurs approches d'un même objectif fondamental : manipuler et comprendre le programme complexe de la vie. Alors que la biologie synthétique a une approche ascendante de la construction de systèmes vivants à partir d'un répertoire d' 'éléments', la biologie des systèmes se distingue par son approche descendante et ses analyses rétroingénieriques des mêmes procédés au sein des cellules vivantes. La biologie des systèmes permet de comprendre quels sont les éléments importants et les interactions déterminantes. Les deux approches ont grandement contribué à nous faire apprendre le fonctionnement des systèmes biologiques. Les découvertes qui en découlent pourraient avoir un impact dans différents domaines, tels que ceux de la médecine, de la manufacture, de l'énergie ou encore de l'environnement (1).

L'un des résultats intéressants des efforts de recherche dans la biologie synthétique est la capacité à concevoir des constructions génétiques codant pour des protéines avec de nouvelles fonctions, optimisées. Bien que des progrès aient été réalisés dans la manière de prévoir comment les séquences d'acides aminés d'une protéine peuvent avoir un impact sur sa forme structurelle d'ordre supérieur et son comportement, les méthodes expérimentales pour l'ingénierie des protéines restent indispensables lorsqu'il s'agit d'évaluer les prédictions dérivées de modèles computationnels. Malheureusement, le cycle conception-construction-test pour le développement de protéines conçues rationnellement requiert généralement beaucoup de temps, de travail et de ressources.
La difficulté à générer des variantes de gènes, ainsi que les obstacles majeurs dans l'expression cellulaire et dans le screening quantitatif des protéines en sont les causes principales.

Cette thèse décrit le développement d'une technique d'assemblage de gènes en phase solide et son intégration dans l'analyse microfluidique des protéines afin d'accélérer le prototypage de nouveaux modèles de protéines. Le procédé d'assemblage de gènes utilise les produits d'échelle commerciale, de l'ADN fabriqué chimiquement, et fonctionne sans l'utilisation d'une ligase ou d'enzymes de restriction. En tant que processus d'assemblage visant à être reproduit à grande échelle, la modularité de celui-ci permet la génération efficace et économique de variantes génétiques prête à s'exprimer. Nous associons directement ce pipeline d'assemblage de gènes à un dispositif microfluidique à haut débit pour permettre l'expression, la purification et la caractérisation de la protéine. En contournant le clonage moléculaire et les étapes où l'utilisation de cellule est nécessaire, le temps de latence entre la conception de protéines et l'analyse quantitative est considérablement réduit.

En tant que preuve du concept, cette plate-forme d'ingénierie de protéines a été appliquée à la construction de plus de 400 variantes artificielles, toutes conçues à partir des protéines de doigt de zinc (DZ) C_2H_2. Les domaines protéiques des DZ sont la forme la plus répandue du facteur de transcription (FT) chez l'homme et se retrouvent tout au long de l'arbre de vie. Ils fournissent une structure pratique pour leur remaniement en raison de leur petite taille et leur modularité. Ils sont aussi un modèle idéal pour explorer la biophysique de la spécificité FT-ADN. La possibilité de concevoir des protéines DZ les rend utiles pour des applications en biotechnologie et biologie synthétique puisque leurs cibles sont programmables. Nous démontrons que, bien que les DZ peuvent être facilement modifiées pour reconnaître une cible particulière d'ADN, l'ingénierie précise de l'énergie liaison ADN-DZ reste un défi. En outre, nous montrons que l'affinité de liaison d'ADN-DZ peut être régulée indépendamment de la spécificité de la séquence. Ensemble, ces résultats démontrent la polyvalence de l'assemblage de gènes, associée à la plateforme microfluidique en tant que nouvel outil pour l'ingénierie rationnelle des protéines.

Mots clés : facteurs de transcription ; protéines doigts de zinc ; la microfluidique ; MITOMI ; criblage à haut débit ; synthèse de protéines exempt de cellules ; assemblage des gènes ; l'ingénierie des protéines ; affinité de liaison ; la biologie synthétique ; domaine de liaison à l'ADN

Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 DNA Oligomers and Gene Synthesis

De novo chemical synthesis of DNA (deoxyribonucleic acid) oligomers has been possible since the late 1950s (2), and was a critical tool in the elucidation of the genetic code by Gobind Khorana and Marshall Nirenberg in the 1960s (3, 4), roughly a decade after James Watson, Francis Crick, Maurice Wilkins and Rosalind Franklin solved the structure of DNA in 1953 (5). Modern oligomer synthesis has its foundation in polynucleotide chemistry research from the laboratory of Robert Letsinger during the 1960s and 70s, which demonstrated oligonucleotide synthesis from a polymer substrate and later utilization of phosphorochloridites for faster reactions and improved yields (6, 7). Contemporary oligomer synthesis occurs via automated workflows incorporating solid-phase controlled pore glass (CPG) phosphoramidite chemistry developed in the laboratory of Marvin Caruthers(8) in 1981. The phosphoramidite approach involves a 4-step reaction cycle of deprotection, condensation/coupling, capping, and oxidation that extends an oligomer chain from a silica (CPG) column substrate (7-10).

Integrated DNA Technologies (IDT), the world's largest supplier of custom nucleic acids, still produces the majority of its oligomers via CPG columns, which allow unidirectional flow of reagents from the core to the reaction surface, where the 3'-affixed oligonucleotide grows in the 5' direction. Due to imperfect coupling efficiencies dependent on both the type of nucleotide being incorporated and its position, oligomer yields decrease with length. In spite of this shortcoming, owing to improvements in reagent quality, reaction efficiencies and purification steps, low-cost synthesis of oligomers up to

200 nucleotides (nt) with error rates as low as 0.5 per 100 nt has become commonplace (11-14). The development of oligomer microarrays through the use of photolithographic techniques coupled with light-directed synthesis reactions in the early 1990s (15), and later innovations involving inkjet printing (Agilent Technologies) or electrochemical techniques (CustomArray), have significantly reduced oligonucleotide production costs compared to column-based synthesis, albeit with considerable reductions in yield and quality (10, 11, 16-18). DNA oligomers with lengths approaching 1 kilobase (kb) have been synthesized, but these oligomers are extremely low yield and become cost-prohibitive for large library sizes. Considering the technological and scientific achievements since the construction of the first synthetic gene in 1970 (19), it is reasonable to expect continued advances in the field of oligonucleotide synthesis.

Gene synthesis, or the fabrication of gene-length DNA constructs, is commonly achieved through enzymatic assembly of short oligomer sets into longer chains, which can approach lengths of several kilobases (10, 11, 20-23). One option for enzymatic assembly, called ligase chain assembly (LCA), involves the use of thermostable DNA ligase to join overlapping oligonucleotide segments spanning the entire length of both strands of the gene at elevated temperatures to melt secondary structures (24, 25). While this technique is simple to implement, it is also the most expensive, as it requires 100% of the gene of interest to be synthesized since both strands need to be partitioned into smaller, gapless units. Another technique for assembly is polymerase cycling assembly (PCA (26) or recursive PCR (27)), which uses polymerase to extend partially overlapping (15-25 bp) oligomers, followed by PCR amplification of the full length construct. Since the oligomers are only partially overlapping, this technique is slightly less costly as less oligomers are needed for assembly. Variants of PCA are the TopDown (28, 29) or Automatic Kinetics Switch (30) methods, which require special design of oligomers such that those in the gene center anneal at high temperatures, while oligomers at the gene extremities are designed with decreasing annealing temperatures. This forces the thermodynamics of the assembly reaction such that the core sequence is stringently built at high temperatures, and through successive PCR rounds with decreasing annealing temperatures, the rest of the gene is extended to its final length. PCA also requires a final round of amplification using primers that anneal at the extremities of the desired product.

Both ligation- and PCR-based techniques can use conventional column-synthesized oligomers or oligomer microarrays for assembly material. Methods that choose

microarrays as their DNA source require additional engineering considerations to address the disadvantages of microarray oligomers, such as isolation of sequence-related oligomer pools, selection of error-free sequences, and amplification of the source oligomers to obtain concentrations conducive to gene assembly (17, 31-36). Genes synthesized from column- or microarray-sourced oligomers will inevitably contain sequence errors, and there are a variety of hybridization and enzyme mismatch cleavage techniques available to improve sequence fidelity in oligomer source pools as well as the full-length, amplified gene constructs (11, 16, 25, 37). Ligation and PCR-based assembly techniques are frequently performed in convenient “one-pot” reactions, which contain all of the oligomer segments and terminal amplification primers, but these reactions are not compatible with assemblies with high GC content or repetitive stretches. To circumvent problems arising from synthesis of genes containing repetitive stretches, solid-phase iterative capped assembly (ICA) (38) and fast ligation-based automatable solid-phase high-throughput (FLASH) assembly (39) have been demonstrated as viable ligation-based techniques.

Following gene assembly, PCR amplified, and error-corrected constructs are commonly cloned into plasmids with restriction enzyme digests, transformed into the microorganism of interest, and sequenced for verification of error-free constructs. The clever application of fluorescent or colorimetric reporter genes can be used to pre-select for errorless assemblies in transformed colonies, since deletions and insertions are the predominant error forms in gene synthesis, and only in-frame assemblies will properly express the reporter (40). The time-consuming, expensive and laborious process of generating and screening for a single perfect sequence is a major barrier to synthetic biology, gene circuit and pathway engineering, genome refactoring, and protein engineering, and thus a major obstacle to be overcome.

Table 1.1 : Comparison of common gene assembly techniques

Technique Name	Method of assembly	Advantages	Limitations	Length scale
Ligation-based assembly	Thermostable DNA ligase to join oligomers, followed by PCR amplification of full-length construct	Simple to implement, secondary structures melted at elevated ligation temperature, one-pot	Large number of oligos must be purchased to completely cover both strands and have 5'-phosphorylated ends for ligation; cloning and sequence verification necessary; high GC content incompatible	Limited by oligomer costs, <5 kb (for compatibility in final PCR amplification)
PCR-based/ Polymerase cycling assembly (PCA)	Partially overlapping oligomers are annealed and extended by polymerase (non-exponential), followed by PCR amplification of full-length construct	Assembly and amplification reactions in single-step procedure; easy to multiplex; 15-25 nt overlap requires less oligomers to span full gene; easy to introduce targeted diversity using variants of a single oligomer	Difficulties may occur for repetitive sequences or secondary structure formation; cloning and sequence verification necessary; higher error rates from less hybridization-based error checking	<5 kb (for compatibility in final PCR amplification)
Array-based assembly	Oligomers sourced from a microarray are amplified and used in ligation or PCR based assembly	High diversity oligomer pools are very inexpensive; on chip hybridization can be used to filter out erroneous sequences; oligomer subpools can be spatially/physically segregated to prevent cross-hybridization	Very low concentrations of each oligomer in the array require PCR amplification before assembly; higher error rates than column-synthesized; no spatial segregation of variants (high interference); requires high level of orthogonality between genes to prevent cross-hybridization of oligomer pools	<5 kb (for compatibility in final PCR amplification)

Larger DNA assemblies with multiple sequence-verified, gene-length parts can be created with a variety of methods including circular polymerase extension cloning (41, 42), Golden Gate method using type IIs restriction endonucleases and ligation (43, 44), sequence and ligation-independent cloning (SLIC) (45), ligase cycling reaction (46), yeast homologous recombination (47-50), bacterial recombination (51, 52), seamless ligation cloning extract (SLiCE; bacterial and λ prophage Red recombination) (53), Gibson isothermal assembly (54), and solid-phase cloning (SPC) (55).

De novo gene synthesis has many potential applications, and has been used to demonstrate the effect of codon bias, promoter design (56-58), examining inducible enhancers (59), identifying protein domain interactions (60), constructing genetic circuits and engineering recombinant proteins for optimal behavior and expression (33, 61-66),

constructing or rewriting synthetic genomes (67, 68), aptamers for therapeutic or diagnostic applications (69, 70), and material science projects in which DNA is used to construct complex, self-assembling nanostructures, also called DNA origami (71, 72). The emerging applications listed here are only meant to give a glimpse of the diversity of scientific questions that can be addressed with synthetic DNA oligomers and gene assemblies. Given the current state of the technology, and the reasonable expectation for cost-reducing and quality-enhancing advances in the DNA synthesis industry, it is evident that there will be considerable expansion in the domain of gene assembly as future applications are explored.

1.2 Protein Engineering

Protein engineering is a broad term encompassing all aspects of protein modeling and design, modification, expression, and optimization of biochemical activity. The majority of engineered proteins are inspired by structures previously observed in nature. Through the use of rational design principles based on computational and experimental observations of permutated natural variants, novel functions can be incorporated into the protein framework or existing functions can be improved. Protein engineering is directly coupled with gene synthesis, since it is invariably changes in the genetic code that lead to successful expression, folding and function of the protein of interest. The ability to rapidly write in DNA, then transcribe, translate and test the associated protein, then process multiple iterations of this workflow, would be an invaluable tool to addressing how an amino acid sequence underlies protein structure and function.

There are a variety of approaches to evolving or rationally designing a protein of interest for new applications (23, 73). Experimentally based approaches typically involve starting with a naturally occurring form of protein with a given activity, and after generating structural diversity through random or focused mutations (74-79), performing a selection screen to harvest variants with enhanced activity. For such experiments, there needs to be a direct link between the genotype and phenotype so variants that pass the screen can be sequenced and identified. After investigating the genetic changes that result in the leap in function between the wild-type (WT) protein to the new variant, these modifications can be used to inform subsequent cycles of design and characterization. Computational-

based approaches typically require crystal structures and in depth biochemical analysis of a set of variants in order to provide some predictive capability for how other hypothetical changes could augment or diminish function (80). Computational designs inevitably require experimental validation, so at best, *in silico* predictions can be used to restrict the multitude of physical forms to be created and evaluated in a functional screen (81, 82). Separately, by sampling from natural, well-modeled protein folds and structures, it is possible to computationally derive non-natural designs to circumvent the growing but limited library of recycled designs found in nature (83).

The experimental approach of performing iterations of mutagenesis followed by screening is embodied by directed evolution (77, 84-86), which can be a fruitful, but time and resource intensive method for generating biological parts with novel functions. Mutagenesis is frequently accomplished with error-prone PCR (epPCR), inverse PCR or DNA shuffling (78, 87-90). In phage-assisted continuous evolution (PACE), a mutagenesis plasmid is used to allow inducible levels of mutagenesis by proof-reading suppression and lesion bypass (91, 92).

Depending on the breadth of variation and screening technique, targeted, systematic, or random modes of mutagenesis can be used to produce libraries with increasing levels of genetic variation (62, 77). Targeted mutagenesis, which modulates a limited set of positions with a constrained collection of amino acid residues, requires less screening effort to sort through the smaller population of variants, but necessitates prior knowledge of which mutations will be relevant and informative for the regions of interest. Systematic mutagenesis, such as alanine scanning, methodically changes many individual positions to a single residue to track how singular mutations affect function, which can be useful for identifying important residues in the absence of structural data, but will omit potentially significant interactions that occur between multiple residues. Random mutagenesis produces the most diverse and numerous libraries for screening, enabling the recovery of exotic amino acid combinations with improved function, but is highly reliant on an effective selection scheme. Regardless of the method or degree of library diversification, the main bottleneck in directed evolution is functional screening of the variants.

Genetic screens used to associate a genotype with an observed phenotype are an essential part of the process of directed evolution or even for screening single-gene libraries. These screens can be broadly split into two categories, screens involving

spatially segregated variants or screens involving bulk, mixed populations of variants. Spatially separated variants can be expressed in single cells of bacteria, yeast or Chinese hamster ovary (CHO) cells and examined as colonies or cultures in multi-well culture plates, which already limits the throughput of the screening technique (93). The use of an appropriate cell-free protein expression system (discussed further in section 1.5) can offer considerable advantages compared to traditional cell-based expression in accelerating the process between gene preparation and protein translation (94). Moving from standard culture plates to microfluidic devices reduces the volume of reagents needed for screening while improving throughput, but requires photolithography facilities for developing chip designs and a microarray printer to individually place the gene variants. Spatially separated screens can be based on optical output (fluorescence, colorimetric, luminescence or turbidity) (95) or coupled with nuclear magnetic resonance (NMR), high-performance liquid chromatography, gas chromatography, or mass spectroscopy to measure substrate depletion or product formation. Other genetic screens can be linked to toxic genes or essential genes that require suppression/cleavage or activation/reassembly, respectively, via the engineered protein variant in order for the cell to be viable and survive the selection (96).

Genetic screens for mixed variant populations require some method of purifying selection to distill out variants with optimal properties. This can be accomplished using fluorescence-activated cell sorting (FACS), which requires a fluorescent-reporter gene output or fluorescently labeled antibody to sort variants. Oil-in-water emulsion droplets, agarose droplets or cross-linked polymeric beads can also be used to isolate single transformed cells or unique gene constructs within miniature reaction vesicles for measuring substrate conversion or fluorescence. Similarly, microfluidic systems can trap single cells or physically isolate arrayed gene variants within unique reaction chambers for high-throughput measurements. Isolated gene libraries that are not transformed into microorganisms rely on encapsulated cell-free transcription/translation mixtures for expression of the protein of interest (97). Selections performed for binding affinity can be used to screen a mixed population since the full library of variants can be challenged with an immobilized target, and only ‘strong’ binders will remain after several rounds of washing to remove weak or non-binding variants. This technique is used in cell surface display or biopanning with bacteriophage display in which protein variants are fused to endogenous surface proteins (98). As a solution to the transformation bottleneck observed in surface

display techniques, ribosome display can be used to chemically link the growing protein to its coding mRNA through in vitro expression using cell extract (99, 100). In order to observe which variants survived the selection and understand what genetic changes occurred to arrive at a protein with the desired activity, nearly all screening strategies require DNA sequencing to identify optimizing mutations. By using high-throughput DNA sequencing, deep mutational scanning can be used to observe how populations of gene variants fluctuate before and after selection (63). A global comparison of library diversification techniques and screening/selection strategies can be found in the review by Packer and Liu (2015) (101).

Prediction algorithms for protein folding and function have improved over the decades with the continued growth of experimentally determined three-dimensional protein structures (using NMR or X-ray crystallography) and the expanse of sequence information produced from massively parallel sequencing technology (102, 103). Using template or homology model building, accurate protein structural models can be derived by comparing amino acid sequences to proteins with known structures (104). Since protein structure and function are interrelated, the ability to infer structural form from a protein sequence would enable predictions of protein activity and biological interactions directly from the multitude of sequenced gene databases. Additionally, such structural and functional predictions would impact studies of protein mutability landscapes and facilitate improved protein design guidelines through computational evaluation of mutations (105). Until predictive softwares are capable of robustly generating accurate protein models, experimental mutagenesis and functional evaluation of protein variants will remain a necessary part of validating engineered protein designs.

1.3 Cys_2His_2 Zinc Finger Transcription Factors

Underlying the complex network of interactions governing cellular growth, development, internal signaling, and responses to external stimuli, is a broad class of proteins called transcription factors (TFs) (106). TFs bind to specific regulatory DNA sequences (TF binding sites – TFBSs (107, 108)) and modulate the process and rate of transcription of genetic information from DNA to messenger RNA (mRNA). TFs can induce these changes in transcription through several mechanisms, by binding a given

DNA sequence alone or in concert with other proteins, and recruiting (gene activation) or blocking (gene repression) transcription machinery from assembling at the gene location. In combination with other proteins, TFs can also direct chromatin remodeling by regulating histone proteins (via post-translational modification by acetylation or deactylation (109)), thereby altering the degree of condensation of the DNA coding for a gene, changing the accessibility of the gene to transcription machinery (110). TF proteins are structurally modular with a DNA-binding domain (DBD) (111) that attaches to a specific sequence of DNA and closely-related variations. They also typically include a signal-sensing domain (SSD) that binds ligands and a trans-activating domain (TAD) that provides a docking site for other proteins. Both the SSD and TAD add levels of complexity to the transcriptional regulation of a particular gene. There also exists DNA-binding proteins that are also capable of binding to RNA, and in the context of TFs, this may facilitate still other modes of transcriptional regulation through competitive binding of DNA and RNA targets, altering mRNA stability and translation efficiency, and recruiting other proteins to a DNA target site.

There are many types of DBDs in TFs, the most prominent being helix-loop-helix (HLH), leucine zippers, homeodomains, and zinc finger (ZF) proteins (112-115). By far, the ZF protein family is the most prevalent form of DBD and can be found across the tree of life (76, 116-118). The Cys_2His_2 ZFs (C_2H_2-ZF) comprise the largest class of DBDs in metazoans, and exist as solitary domains, but are more commonly seen as an array of multiple C_2H_2-ZF domains in tandem (76, 118-120). The C_2H_2-ZF domains are relatively small, 30 amino acid (aa) units, organized around a $\beta\beta\alpha$ configuration (Figure 1.3.1). Each C_2H_2-ZF tetrahedrally coordinates around a single zinc ion through interactions between two cysteine (Cys) residues within the two-stranded antiparallel β-sheet and two histidine (His) residues in the α-helix. DNA sequence specificity can be tuned primarily by changing residues -1, 2, 3, and 6 in the α- or 'recognition' helix (76, 121, 122).

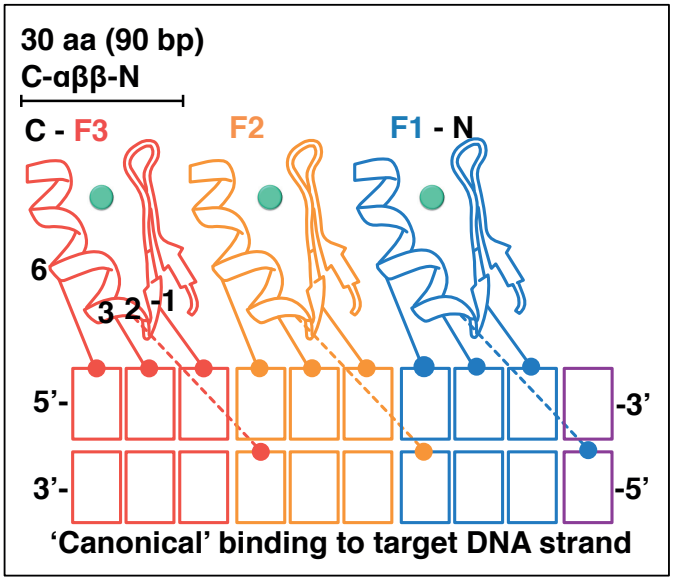


Figure 1.3.1 : Cartoon model of canonical Cys_2His_2 ZF TF binding to DNA with residues -1, 2, 3 and 6 of the recognition helix primarily encoding DNA specificity. Residue 2 makes a cross-strand contact, which creates 'context dependent' effects. Zinc ions are colored in light blue.

For a single ZF module, these residues are directly involved in targeting a 3-4 base-pair (bp) DNA address, while tandem ZF arrays bind to longer DNA sequences essentially made up of the concatenation of the individual target sites of each module in the array. Due to a cross-strand base interaction from residue 2 in the recognition helix with the first nucleotide in the complementary strand of the 3'-adjacent target triplet, some DNA targets of ZF arrays are not simply concatenations of the different DNA target sites. This target site overlap is a characteristic of tandem ZF arrays, creating an effect called 'context dependence' (119), is the primary reason for the lack of a 'recognition code' or a set of general guidelines for designing an optimal ZF array for any desired target site (120, 123). The context dependent interactions between adjacent C_2H_2-ZFs complicates predictive modeling of their DNA-binding specificities, and poses a considerable challenge to engineering C_2H_2-ZFs with precise targeting capabilities.

C_2H_2-ZF domains are unique among TFs in that by altering the DNA specificity residues in the α-helix, they could hypothetically be used to target every possible DNA triplet of interest. Their versatility as a modular DNA targeting unit is evident from their abundance and diversity in eukaryotes, but also in the limited number of conserved residues within the C_2H_2-ZF structure, which relies on stabilization through zinc ion coordination and a conserved hydrophobic core. The prevalence of ZF TFs, particularly tandem arrays, in eukaryotes has been traced to a gene family that expanded through

repeated duplications concurrent with functional binding divergence (123). Arrays of C_2H_2-ZF in humans can range from 4 tandem units to more than 30, with an average of 8.5, which in theory would bind a 25 bp target (123-125). A random target of 16 bp (5 C_2H_2-ZF tandem units) is sufficient to encode a unique site within the human genome (~3×10^9 bp), so longer target sites seem superfluous even with consideration of binding site degeneracy within C_2H_2-ZF modules. It is believed that longer tandem arrays operate not to bind a single, specific target but to bind multiple targets with varying sets of C_2H_2-ZF units in different genetic contexts (126). This multifarious behavior can also be inferred from experiments involving artificially created C_2H_2-ZF arrays, wherein binding affinity does not significantly improve beyond 4 tandem units. The C_2H_2-ZF domain is not limited to binding DNA targets, and natural variants have been observed binding hybrid DNA/RNA strands, proteins, and both single-stranded and double-stranded RNA (124, 127, 128). Due to the small size, versatility and modularity of C_2H_2-ZFs, their structure has also been used as a scaffold to create combinatorial peptide libraries for the evolution of novel binding properties (129, 130).

In spite of the engineering challenges in creating artificial C_2H_2-ZF arrays, the utility of a modular DBD with tailored binding specificities was too alluring to be overlooked. The creation of molecules capable of binding DNA to specifically regulate genes and thereby influence cellular development and physiology remains a major ambition of personalized medicine and synthetic biology. Shortly after the crystal structure of a peptide containing the natural triple C_2H_2-ZF array from Zif268 (a human and murine TF protein, also called 'early growth response' EGR1) bound to DNA was solved (131, 132), there was a surge of interest in engineered ZF applications. Initial synthetic ZF research efforts dealt with searching for a DNA-binding recognition code (121,133-136), using phage display (137-142) or other techniques as a screening mechanism to find variants that bound each possible DNA target triplet. Engineered C_2H_2-ZF arrays are typically based on the Zif268 peptide or Sp1 (taken from the consensus of human TF Sp1, another triple C_2H_2-ZF) (143, 144) framework. Either of these naturally occurring proteins is taken as a template structure or 'framework' and a library of recognition helices from available databases or binding models (75, 76, 145-148) with different DNA target specificities are inserted. Variants with differing linker sequences and lengths between individual C_2H_2-ZF modules have also been tested (149). The refactored protein is then tested against a set of DNA targets to evaluate the actual binding energy landscape of the new protein construct.

Many techniques for constructing ZF libraries from synthetic oligomers or plasmid libraries have been proposed (144, 147, 148, 150-152) and implemented with different selection schemes.

Engineered ZF domains have been constructed as artificial transcriptional activators or repressors (153-157), as cell-membrane penetrating carrier domains for transporting proteins (158, 159), and as dsDNA sensors in diagnostic applications (160-162). Their full potential was realized by simply using them as DNA-targeting moieties for a plethora of protein fusions with different effector/enzymatic domains such as integrases, recombinases, invertases, transposases, chromatin regulators (157), and most commonly nucleases (153, 163-173). Considerable effort was applied in engineering ZF nucleases (ZFNs) which are typically designed as a dimerizing pair of FokI nuclease subunits tethered to two different 3-4 C_2H_2-ZF tandem units that bind on opposite strands of the target cut-site. ZFNs are in competition with TAL (transcription activator-like) effector nucleases (TALENs), which have a solved recognition code, but consist of considerably larger DBDs, and also easily implemented CRISPR-Cas (clustered regulatory interspaced short palindromic repeat) RNA guided DNA endonucleases (174).

1.4 Methods for measuring TF-DNA interactions

There exists a diverse collection of experimental methods for both qualitatively and quantitatively detecting and measuring the DNA binding specificity of TFs *in vitro* and *in vivo*. These methods are often used in conjunction with computational approaches to enable binding-site or motif-discovery predictions and infer the network architecture of genetic regulatory systems (175-178). Since TFs must be able to locate, distinguish and bind their specific regulatory elements within a genome containing an excess of competing binding sites, the binding specificity of TFs is crucial for understanding how regulatory networks function (108, 176). By mapping which DNA elements are important in transcriptional regulation, it is possible to understand how TFBSs influence mRNA translation as well as generate the phenotypes created by the proteins they encode. Additionally, by resolving the regulatory networks within cells, it is possible to determine the source of different disease states caused by genetic variations that disrupt normal gene expression. Whereas *in vitro* experimental methods are primarily useful for

identification of TF consensus binding sites, determination of binding energy landscapes or evaluating biophysical parameters, *in vivo* methods provide genome-wide information within the context of different environmental conditions and biological settings (ie cell type or developmental stage).

1.4.1 In vitro Methods

EMSA and Nuclease Footprinting (Classical techniques)

One of the classical methods for *in vitro* characterization of protein-DNA interactions is electrophoretic mobility shift assays (EMSA) (179) and related methods (180). This technique is relatively inexpensive, but low-throughput and time/labor intensive, and can provide a quantitative measure of TF-DNA binding. In these assays, a TF or protein mixture is incubated with a potential DNA target. If protein binding occurs, the migration of the protein-DNA complex through an agarose gel within an electric field is impeded compared to unbound DNA, causing it to run more slowly through the gel resulting in a spatial shift of the band. This technique is usually reserved for qualitative analysis of TF binding activity, and is commonly bypassed in favor of other more comprehensive, data-rich, yet elaborate, characterization techniques.

Nuclease footprinting analysis is another classical technique developed in 1978 for determining the target sites of DNA-binding proteins (181-183). This technique is based on the fact that TFs bound to DNA will protect the phosphate backbone from cleavage during incubation with DNase-I. A DNA target sequence, between 50-200 bp long, with a fluorescent or radioactive label at one end is incubated with the TF, then digested. The same DNA is digested in the absence of the TF, which will produce random cuts. The labeled DNA products from both reactions is run on a gel to visualize the resulting DNA fragment pattern (footprint), and by comparing the DNA products, binding site and kinetic binding constants can be determined.

SELEX

SELEX (systematic evolution of ligands by exponential enrichment) (176, 184, 185) involves exposing purified TF to a large oligonucleotide library containing randomly generated or genomic sequences of uniform length but flanked by universal primer binding

sites. Sequences bound by the TFs are amplified by PCR whereas unbound sequences are washed away or separated from TF-DNA complexes via gel filtration. The amplified DNA is used for additional rounds of selection and amplification and in the end, the remaining bound targets are cloned and sequenced or used directly for massive, parallel sequencing (SELEX-seq) to identify the highest affinity target sequences (186, 187). A single round of oligomer binding, capture, amplification paired with high-throughput sequencing (Bind-n-Seq) was found to be sufficient to correctly identify consensus binding sites and give approximate target affinities based on the frequency of sequence reads (188).

SELEX was one of the first techniques capable of *de novo* TFBS consensus determination without any prior knowledge of TF binding site preferences. High-throughput (HT-SELEX) can be accomplished by using DNA barcodes included within the sequence of the oligomer library to uniquely identify individual SELEX samples (different TFs) after pooling bound DNA elutes and performing massively parallel sequencing (187). Since SELEX requires purified TF for binding, non-natural concentrations of TF relative to the oligomer library may result in the detection of binding artifacts, and the TF of interest may be missing relevant post-translational modifications (phosphorylation, hydroxylation, acetylation) (189, 190), binding cofactors, or ligands that alter the binding behavior due to expression in foreign systems. Additionally, because SELEX requires a washing or gel filtration step to separate the TF-DNA complexes from unbound sequences, it may fail to detect weak-affinity interactions, reducing the sensitivity of binding detection. Finally, the need to amplify TF-captured DNA with universal primers may introduce amplification bias or lead to the creation of chimeric sequences which result in inaccurate reporting of the abundance of sequences, skewing the approximation of relative binding affinities for different targets (191, 192).

PBM

Protein binding microarrays (PBMs) (141, 185, 193-195) can also be used for high-throughput, *de novo* identification of TF consensus sequences. This method involves incubating an epitope-tagged or fluorophore-labeled TF, either purified or within cell nuclear extract, with a dsDNA microarray. The dsDNA array starts as ssDNA and the complementary strand is produced via primer extension prior to introducing the TF. After washing to remove nonspecifically bound TF, locations of TF binding are detected with

either direct fluorescent signal from the TF or using a fluorescently labeled antibody targeting the presented epitope tag. The specificity of the TF can be determined by comparing the relative fluorescence levels across target spots. Although PBMs offer a rapid method for identifying TFBSs, they may not necessarily provide accurate predictions of endogenous TF binding due to missing cofactors or chromatin structure, and the microarrays themselves usually have DNA length and array density limitations that may require additional experiments to map a large sequence space. PBMs also require multiple washing and blocking steps to remove nonspecific binding events from being detected, but this simultaneously reduces the sensitivity of the screen since weak-affinity interactions will be disrupted and overlooked.

By utilizing phage display and DNA microarrays, a procedure similar to SELEX can be used to repeatedly select and amplify in liquid culture phage bound to a target DNA, followed by sequencing of the phage inserts to determine the identity of the protein (138, 196). Another technique similar to PBMs is CSI (cognate site identifier) (122) which uses a ssDNA microarray where each strand folds back on itself to form a dsDNA hairpin target. To examine TF binding, rather than requiring fluorescently tagged protein or labeled antibody for detection, a DNA intercalating dye is used to compete with TFs for DNA binding (FID, fluorescent intercalator displacement) (197).

A technique related to promoter disruption analysis and PBMs, termed synthetic saturation mutagenesis, utilized a DNA microarray to systematically create a library of barcoded, single- or double-nucleotide mutations and point deletions within promoter regions for known phage RNA polymerases (198). Using *in vitro* transcription extract and reverse-transcriptase PCR with short read sequencing to map important DNA elements by counting the distribution of barcodes retrieved, the technique was able to determine residues within the 'footprint' of polymerase binding, but also identify some position- and nucleotide specific variants that enhanced transcription compared to the native promoter. The technique was also applied to mammalian promoters using HeLa nuclear extract. Although this approach was successful at identifying important promoter elements, it required prior knowledge of the binding sequence to generate a library within the length-restrictions of synthetic DNA microarrays.

Finally, a variation of PBMs integrated with second generation, high-throughput sequencing, called HiTS-FLIP ('high-throughput sequencing'-'fluorescent ligand interaction profiling') (107, 199), was developed and validated by studying DNA binding interactions of

the yeast protein Gcn4. The process requires the use of a sequencing instrument, wherein millions of different DNA targets, localized in distinct clusters of identical sequence, are sequenced by incorporation of fluorescently labeled nucleotides. Once the sequence identity of each cluster is determined, the synthesized strand is melted away and the substrate-bound target strand is used in a Klenow reaction to generate dsDNA clusters. The purified, fluorescently labeled protein of interest is then washed across the flow cell of the sequencing instrument at increasing concentrations, and binding events are captured by imaging. By superimposing the fluorescent images generated during sequencing with those from protein-binding steps, the DNA clusters that are bound by the protein can be identified. Although this technique offers higher throughput and sensitivity than PBMs and can screen longer target sequence libraries, it requires expensive sequencing instruments and can only study one protein at a time.

SPR

SPR (surface plasmon resonance) (179, 200, 201) functions on the principle that light is reflected from a thin layer of gold at different angles depending on the composition of molecules adsorbed to the opposite (non-incident) surface of the gold layer. The light that hits the gold creates a plasmon wave in the plane of the gold layer, and the oscillations of the wave are highly sensitive to irregularities in the gold surface such as adsorbed proteins. The angle shift of reflected light compared to a reference (before protein is introduced) can be used to measure the amount of adsorbed material on the surface. BIAcore presented the first commercial SPR system for detecting biomolecular interactions and measuring binding affinity, and hence the company name is commonly associated with the technique. The method requires a protein or DNA target to be immobilized on the sensor chip surface, then the binding partner is introduced at different concentrations and interactions with the surface-bound analyte are recorded via fluctuations in the angle of reflection. The technology can be used to monitor both association and dissociation kinetics of any biomolecular binding-pair by washing the surface until equilibrium is reestablished. For studying TF-DNA interactions, a high-throughput variation of SPR can be realized by arraying multiple DNA sequences to the sensor surface, enabling parallel detection of binding events using a single TF. SPR technology provides a label-free alternative to detecting and measuring TF-DNA interactions with low sample amounts, and although measurements can be made rapidly

with high sensitivity, the technique has several disadvantages. Firstly, without a method for signal amplification, SPR requires at least 0.1% of substrate-linked molecules to attach to their binding partner in order to detect the interaction (202), which limits its efficacy to detect low affinity interactions. Secondly, the surface must be functionalized and sufficiently blocked to prevent non-specific adsorption, which can lead to the production of false signals. Finally, as with other *in vitro* techniques discussed previously, protein behavior *in vitro* may not accurately represent *in vivo* activity.

MITOMI

MITOMI (mechanically induced trapping of molecular interactions) is an *in vitro* technique combining aspects of PBMs with microfluidics to achieve high-throughput detection of low-affinity and transient TF-DNA binding events (203, 204). The realization of microfluidic large scale integration (MLSI) is the primary technological advance enabling the MITOMI principle (205, 206). The MITOMI microfluidic device is composed of hundreds to thousands of physically separated unit cells (Figure 1.4.1 and 1.4.2), which are sealed via pressure-driven actuation of silicone elastomer (polydimethylsiloxane, PDMS) valves.

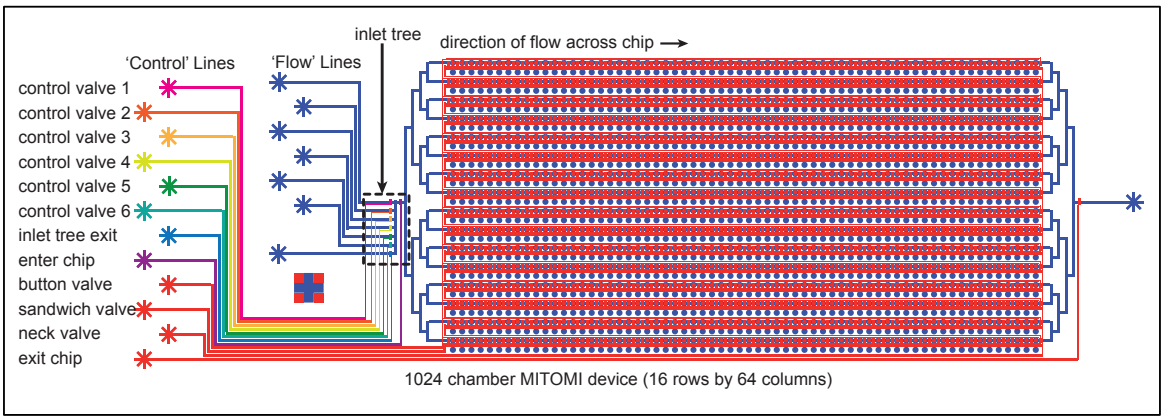


Figure 1.4.1: A schematic of the 1024 chamber MITOMI device used for all of the experiments described in this publication. On the far left are several ports where PBS-filled control lines are inserted to actuate microfluidic valves on the device with compressed air. The last 4 lines (in red) control the button valve, sandwich valve, neck valve and chip exit valve, from top-to-bottom, respectively. The flow lines are where experimental buffers/reagents are inserted to flow across the chip. The operation of the device is detailed in the methods section.

PDMS begins as two components, a liquid polymer base and a curing agent, and after mixing and baking, crosslinks to form an elastomeric, transparent, gas- and water-

permeable, biologically-compatible material (207). By casting the uncured mixture over a micropatterned surface, a microfluidic device, composed of micron-sized features, can be produced. The device is composed of two separate overlaid layers (termed the *control* and *flow* layers; Figure 1.4.3) created using standard multilayer soft photolithography techniques.

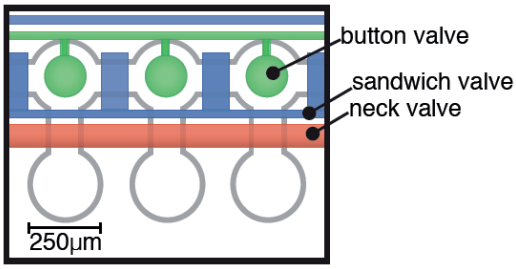


Figure 1.4.2: A cartoon of a three MITOMI unit cells with labels for the button, sandwich and neck valves.

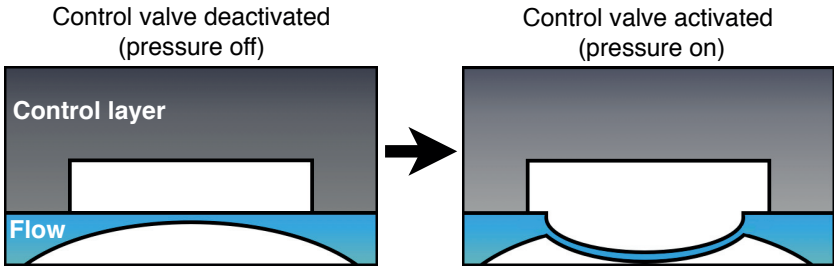


Figure 1.4.3: A cartoon image of the control layer (gray) positioned overtop the flow layer (blue). By pressurizing water or another fluid in the channels above the flow layer, the thin flow layer membrane can be deflected and pressed against the surface substrate.

The control layer containing channels for actuating the flexible valves is placed on top of the flow layer, which contains channels and chambers for flowing experimental reagents and conducting biological assays (Figure 1.4.3). Each unit cell (Figure 1.4.2) is composed of two compartments, the reaction chamber and the DNA chamber, and is controlled by three valves termed the *sandwich*, *neck* and *button* valves. The sandwich valves function to seal each unit cell from neighboring chambers to prevent mixing and contamination between reaction chambers. The neck valves function to protect the DNA chamber from rehydration until after the experimental surface is appropriately blocked and patterned for the assay of interest. A single button valve per unit cell (centrally placed in the reaction chamber) is used for surface derivatization and molecular interaction trapping and protection during washing steps. The device is aligned over a DNA microarray printed onto an epoxy-functionalized glass microscope slide such that each DNA chamber

encloses a single spot of deposited target DNA. Through successive rounds of flowing biotinylated-BSA (bovine serum albumin), neutravidin (a deglycosylated form of avidin), then activation of the button valve to protect a circular area within the reaction chamber and flowing biotinylated-BSA to block the remaining area in the reaction chamber, an island of neutravidin is produced. The surface can then be functionalized with a biotinylated antibody for precipitation of tagged protein. In the case of TFBS analysis with MITOMI, each DNA chamber contains a different fluorophore-labeled DNA target sequence or concentration. Protein expression can be performed off-chip then the assembled protein can be flowed across the device and captured by the surface-bound antibody, or the protein DNA template can be arrayed with the DNA targets and *in situ* expression can be performed by incubating the templates with cell-free extract (in vitro transcription/translation mixtures). In this way, the MITOMI device functions to compartmentalize a TF array with different DNA targets to evaluate binding interactions, in a manner similar to nucleic acid programmable protein arrays. Protein-DNA interactions are captured by activating the button membrane and washing away unbound material in each reaction chamber. An experimental variation of MITOMI that does not require antibodies involves surface-bound biotinylated and fluorescently-labeled target DNA and trapping proteins interactions after incubation. The protein is expressed with fluorescently-labeled amino acids (ie using the commercially available FluoroTect™ BODIPY®- FL labeled tRNA-lysine from Promega) or as a fusion protein with a fluorescent domain so that trapped protein and bound target DNA amounts can be quantified with fluorescent calibration curves. The MITOMI principle has been used to analyze TF binding preferences (204, 208-211), protein-protein interactions (212-214), immunodiagnostics (215), and protein-RNA interactions (216).

1.4.2 In vivo Methods

In vivo methods are a necessary complement to *in vitro* TFBS interrogation techniques, largely due to the need to evaluate how well *in vitro* results translate to living systems, but in general it appears TF binding models based on analysis of *in vivo* or *in vitro* data are equally accurate (217). As with *in vitro* techniques, there exists a selection of *in vivo* techniques to evaluate TF binding within a biologically relevant system. One *in*

vivo time and labor intensive approach to TFBS discovery is gene disruption/promoter deletion analysis (218, 219) paired with a genetic reporter to give a visual read-out of the transcriptional effect of removing or altering a given DNA regulatory element. Molecular barcoding can be used for sequencing or microarray hybridization to identify which sequences are valuable for expression. Although screening of the expression level of the reporter gene can be used to provide a quantitative output, subtle differences in transcription are difficult to detect.

A technique related to gene disruption utilizes a set of synthetic promoter libraries built with random combinations of TFBS to form 'building blocks', examined their effect on fluorescent protein expression in yeast, and used these datasets to train a thermodynamic model for prediction of gene expression from promoter sequence (220). Although the technique was useful to successfully predict gene expression controlled by a single TF for genes with related promoters, the experimental throughput was limited to examination of a single TF, with considerable resource and labor requirements. Other in vivo techniques for TFBS characterization, described below, can be broadly categorized as population-level or single-cell techniques (221).

ChIP

ChIP (chromatin immunoprecipitation) is a technique for the discovery of TFBSs *in vivo* which has its basis in a method developed in 1978 to study histone-DNA interactions (222) and has undergone significant modifications and improvements since (223, 224). This population-level technique involves the chemical fixation of all protein-DNA interactions within a large pool of cells (~10^7) (225), which covalently crosslinks the DNA-binding proteins to their targets. After cell lysis, DNA extraction and shearing the DNA via sonication, protein-bound DNA fragments are selectively purified with an antibody specific to the TF of interest or with an antibody against an epitope-tag that is presented by a modified TF. The crosslinking is reversed and the DNA fragments were analyzed, in the classical version of the technique, by PCR, hybridization with labeled probes, or molecular cloning and sequencing. More recent developments for analyzing the captured DNA involve the use of DNA microarrays in a technique called ChIP-on-chip (or ChIP-chip) (194), or ChIP-sequencing (ChIP-Seq) which uses high-throughput sequencing (226). ChIP-chip begins by first amplifying the captured DNA fragments, labeling them, then

incubating the fragments with a ssDNA microarray for hybridization. After a series of washes, the hybridized array is scanned and image processing software is used to evaluate the signal locations and intensities to determine sequence identities and relative enrichment levels. ChIP-Seq directly sequences all of the purified DNA fragments, offering higher resolution, direct quantification of enriched sequences, better genome coverage and reduced levels of artifacts compared to ChIP-chip (190). A refinement of ChIP-Seq, called ChIP-exo, involves the use of DNA nucleases (lambda exonuclease or DNase-I) to improve the resolution of a given TF binding site (229, 230). For ChIP-exo, following the DNA fragmentation step, the DNA outside of the TF-protected region is digested and the length of the DNA fragment that is sequenced is reduced from ~100-500 bp to 25-50 bp, improving the precision of TFBS discovery (224).

Though ChIP and related variants can be used to provide population-averaged, genome-wide association studies for single TFs, these results are highly dependent on TF abundance, the efficiency of crosslinking, and the availability of an antibody for purification. ChIP techniques are also restricted to capturing direct, high affinity TF-DNA interactions that do no require cofactors for binding in accessible genomic regions (free of nucleosomes), which may not include the complete DNA sequence space and can vary in different cellular settings. The majority of these limitations can be overcome using *in vitro* TFBS mapping methods.

DamID

A second *in vivo* technique, called DamID (194, 231, 232), is based on the methylation of nearby DNA when bound by a TF fused to the DNA adenine methyltransferase (Dam) protein. Dam methylates the adenine within the sequence 'GATC' which occurs on the order of several hundred times depending on the length of the genome being examined. Restriction enzymes specific to methylated DNA (DpnI) cleave the methylated regions, which are amplified, labeled and hybridized to a DNA microarray or the methylated DNA can captured with an antibody after shearing and sequenced. Although this technique is not limited by requiring an antibody specific to the TF of interest, the binding site mapping resolution of this technique is relatively coarse due to the frequency of GATC sites. Another detracting feature of this technique is that the gene

coding for the TF-Dam fusion is typically overexpressed from a plasmid, which may lead to binding-site artifacts as a result of non-natural cellular TF concentrations.

Y1H/B1H

The yeast or bacterial one-hybrid (Y1H or B1H) technique is a powerful *in vivo* genetic tool for identifying TF variants that bind a DNA element that was based on the Y2H (yeast 2-hybrid) technique (227, 228, 233, 234). Both one-hybrid techniques function on the basis that some modular TFs consist of both a DBD and a transcription activation domain (TAD). Any TF that can be cloned and expressed in *S. cerevisiae* or *E. coli* can be examined using this technique. Generally, two plasmids are required for transformation: one containing the gene encoding the TF of interest fused to a subunit of RNA polymerase (frequently the omega subunit for B1H) or an activation domain from a yeast TF (such as the TAD of GAL4 for Y1H), and a second plasmid containing either a fixed or randomized binding site located upstream of a weak promoter controlling expression of essential/selectable reporter genes (235-238). The DNA binding site is commonly referred to as the 'bait', while the TF-TAD fusion is called the 'prey'. The plasmids are usually transformed into mutated cell lines that are auxotrophic for tryptophan, leucine, histidine, or uracil (lacking TRP1, LEU2, HIS3 or URA3, respectively), meaning they cannot grow on media lacking these compounds are require recovery of the gene to survive. In the event that the TF-TAD fusion is capable of binding the bait site with sufficient affinity to activate the promoter and trigger essential gene expression, the auxotrophy is rescued, those cell survive to form colonies on minimal media, and the prey-bait combinations can be sequenced to determine the identity of each TF variant and its binding sequence. As an alternative to auxotrophic rescue selection, the bacterial gene LacZ can be used in both B1H and Y1H with media containing X-gal as a selectable, colorimetric (blue colony) reporter in combination with one of the essential genes. This technique can be used in two different modes, either by taking a TF of interest and screening a binding site library for baits that it captures, or by testing variants of TF and selecting for those that bind a specific target sequence of interest. In several C_2H_2-ZF variant screening studies employing the B1H method (75, 76, 152, 233, 239), the bait plasmid contains both HIS3 and URA3 which enables positive and negative selection strategies. By introducing 3-amino-triazole (3-AT) a competitive inhibitor of HIS3, only strongly-activating interactions

(associated with higher expression levels of HIS3) are captured in colonies which survive the inhibitory selection. Separately, by introducing 5-fluoro-orotic acid (5-FOA), which becomes toxic when metabolized in the uracil biosynthesis pathway, active promoters that function regardless of the bait sequence (TF-TADs that function by binding outside the bait site) can be eliminated from the screen. The use of URA3 as a selectable marker comes with the caveat that it is characterized by having a high background since cells with mutations in URA3 may survive on media with 5-FOA. In general, the use of double reporters for B1H and Y1H assays reduces the frequency of false positives and is easily implemented in most selection schemes.

Other technical improvements for B1H and Y1H are outlined by Reece-Hoyes and Walhout (235). Since both B1H and Y1H are reliant on DNA sequencing to retrieve the DNA bait sequence or the TF variant sequence, this method can be limited to the number of colonies that can be picked, unless massively parallel sequencing and DNA barcoding is used to multiplex many different TF variants for unique target sites from different selections. The primary limitation in B1H/Y1H is the creation of a diverse enough library to properly assess a broad sequence space of potential TF designs, but also the transformation efficiency of the strains used. Additionally, since these assays are carried out in vivo, TF variants that bind strongly to the endogenous DNA of the strains used may impart a fitness cost to the organism by disrupting normal gene expression, while weak-affinity variants will not survive the selection steps, and thereby misrepresent the detection of functional variants (190).

1.5 Cell-free Protein Expression

Crude cell-free extracts have been an important tool utilized by biologists for over a century, since Eduard Buchner first metabolized sugar to ethanol and carbon dioxide using yeast ‘press juice’ in 1897 (240), to deciphering of the genetic code by Nirenberg and Matthei during the 1960s (3, 241), to the multitude of recent scientific and industrial applications (242). Cell-free transcription and translation (TX-TL) systems are indispensible tools for accelerating synthetic biology and protein engineering research applications (93, 242-244). Expressing genetic constructs within an experimentally isolated, cell-free system enables the analysis of novel proteins and gene networks

detached from the biological noise of essential processes in living organisms. By circumventing labor- and reagent-intensive traditional molecular cloning, transformation and cell culture steps, cell-free extracts offer a versatile means of rapidly and economically obtaining practical titers of protein from recombinant DNA templates, freed from the myriad constraints of maintaining viable cells. Cell-free extracts have been demonstrated to drastically reduce the time necessary to test synthetic gene parts and networks (245), and recent applications have even proven the viability of using cell-free systems for rapidly prototyping engineered circuits in paper-based platforms (246).

Cell-free protein expression systems can potentially be made from any organism. A wide variety of cell types can be used depending on the complexity of the protein to be synthesized, ranging from prokaryotic and protozoan, to plant, insect and mammalian cells (247). In addition to conventional cell types, cell extracts can also be produced from genetically modified cell strains with special mutations or synthetic genes to enhance protein production, folding, stability, or modification. The choice of which cell-free system to use is dependent on cost, the quantity of protein desired, and whether specific protein folding chaperones, cofactors, or post-translational modifications are needed. The most commonly used, and commercially available, cell-free extracts are made from *E. coli*, wheat germ, and rabbit reticulocyte (243, 248). A crude cell extract, while time and labor-intensive to prepare, is generally sufficient for most protein synthesis applications. Crude extracts are produced by harvesting a large volume of cultured cells via lysis (bead beating, sonication or high-pressure), centrifugation to remove cell wall fragments and genomic DNA, a 'run-off' reaction to release and degrade residual mRNA from captured ribosomes, and dialysis to transfer the TX-TL constituents into a suitable storage buffer (244, 249). These extracts are also referred to as "S30" fractions since the soluble cell components and ribosomes are found in the supernatant after centrifugation at 3×10^4g. The final extract contains all of the essential cellular components for protein synthesis: ribosomes, tRNAs, tRNA synthetases, amino acids, translation factors, nucleotides, nucleotide recycling enzymes, metabolic enzymes, energy substrates, cofactors, and salts (243, 248).

Since most TX-TL systems are operated as batch or batch-fed reactions, which exhibit reduced yields over time due to increasing levels of inhibitory byproducts and declining substrate and energy sources. Significant changes in the preparation and use of cell-free extracts have been made to improve the activity of protein synthesis reactions.

By incorporating ATP/energy-regenerating systems, cell-free extracts can metabolize substrate supplements to survive for extended reaction periods, offering improved protein yields (247, 250-252). Continuous exchange or bilayer systems that enable the diffusion of substrates into and byproducts out of a synthesis reaction boosts reaction productivity and lifetime (247, 253, 254). By supplementing crude cell extracts with foldases, disulfide bond-forming enzymes, and chaperones, more complex proteins can be synthesized and properly folded with high yields. Systematic optimization of the cell growth media, cell-type, lysis method, buffer composition and other processing conditions can also lead to higher yields and improved extract activity.

In contrast to the crude cell extracts which, although functional, contain a relatively unknown mixture of cellular parts that may vary from batch-to-batch, the PURE ("protein synthesis using recombinant elements") cell-free expression system is a special protein synthesis platform made entirely of individually purified and reconstituted TX-TL components at defined concentrations (248, 255, 256). The PURE system consists of 32 polyhistidine-tagged translation factors, tRNA synthetases and other enzymes, each of which are over-expressed in separate *E. coli* cell lines, individuvally purified, then remixed at known quantities with ribosomes, buffer, tRNAs, NTPs, T7 RNA polymerase, and energy regeneration components (detailed composition given in (248, 256)). This system is offered commercially, but at a considerably higher cost per reaction than other commercially available cell extract platforms. Although the PURE system has a lower protein yield, it is still useful in many applications because it is a purified, fully characterized mixture and does not contain endogenous nucleases or proteases, which can degrade linear DNA templates, mRNA and protein products. By implementing a continuous exchange platform, it was demonstrated that the protein yield of the PURE system could be markedly improved (254). For cell-free expression reactions that require the formation of disulfide bonds to derive properly folded protein, simply by exchanging the PURE system's reducing components for oxidized reagents and adding appropriate enzymes, antibodies can be successfully produced (256). The versatility of the PURE system compared to crude cell extracts is also apparent by the substitution of fluorescently or radiolabelled tRNAs for visualization of product yield, the insertion of unnatural amino acids at specific positions by including modified tRNAs in the reaction mix, and the addition of various chaperones (ie DnaK, GroEL/GroES, DnaJ, GrpE, HrpA, Orn, TrxC, Tig, SlyD and PhnH) to enhance protein production, folding and activity (248, 257).

The second most important consideration after selecting which cell-free system to use, is the design and construction of an appropriate DNA template for expression. DNA templates can be exist as circular plasmids, linearized plasmids or linear PCR products, with preference towards using circular plasmids since they are resistant to exonucleases. By shielding linear DNA with DNA-binding proteins, protective sequence ends (245), or circularization, protein expression levels can be improved when expressing linear templates with crude cell extracts containing endogenous exonucleases. Depending on whether a prokaryotic or eukaryotic cell extract is being used, specific sequence elements should be included in the DNA template. DNA templates typically contain an IRES (internal ribosome entry site, found in viral mRNA that allows eukaryotic ribosomes to bind without a 5'cap) or Shine-Dalgarno (prokaryotic) sequence (both are ribosome binding sites or RBSs), Kozak (eukaryotic) sequence, and other appropriate promoters or enhancers in the 5'UTR (untranlated region, or leader seqeuence), affinity tags (258) at the N or C-terminus of the open reading frame (ORF), a poly-adenine stretch at the C-terminus to increase mRNA stability, and transcription terminator sequences in the 3'UTR. Species-independent translational sequences (SITS) have been developed to overcome translational inefficiency in eukaryotic-based cell extracts by including leader sequences which enable ribosome interactions even in the absence of 5'-capping (259). Depending on which phage RNA polymerase is used (commonly T3, T7, or SP6) (260), the appropriate promoter sequence is necessary upstream of the start codon to enable transcription. Phage RNA polymerases are frequently used in place of the extract-endogenous version because of their high processivity, low error rates, and stringent promoter-specific activity. In spite of their efficiency, bacteriophage RNA polymerases offer limited modularity or operator-gated regulation, but an *E. coli* RNA polymerase with sigma factor 70 system was found to provide yields comparable to phage-based transcription systems, while harnessing the larger repertoire of endogenous *E. coli* promoter/operator elements to enable construction of more complex synthetic genetic networks (261, 262).

Chapter 2:
Asymmetric Polymerase Extension (APE) Gene Assembly

2.1 Introduction

To address the significant disconnect between commodity-scale DNA oligomer synthesis, gene assembly, and protein engineering, we began the development of a novel technique for constructing genes. Originally, the gene assembly technique was subject to a set of requirements to facilitate its eventual application towards generating gene variants within a programmable microfluidic device that would also enable *in vitro* TX-TL, protein purification and characterization. All of these experimental steps would be carried out within a single device for rapid, streamlined prototyping of new biological designs. The original requirements for the gene assembly technique were the following:

1) the 'source' DNA oligomers should be purchased at the most economical scale, with minimal purification steps, and with minimal genetic redundancy (sequence overlap) to minimize costs for generating multiple variants
2) restriction or ligation enzymes should be avoided

3) the technique should be modular so that variants could be generated by substituting oligomer parts with distinct coding sequences, and also permit the production of repetitive sequences or gene parts with internal sequence homologies
4) the assembly technique should display a low error rate, so that final gene products could be used directly, without error-correction processing steps, to generate functional, properly-folded proteins using the T7 RNA polymerase-based PURE TX-TL system (255)
5) the final gene assembly product should be compatible with microarray printing with minimal column or agarose gel purification steps

These guidelines were used initially to direct the development of a gene assembly technique towards the construction of a complete, expression ready linear template encoding yeast enhanced GFP (yEGFP), and later used in the construction of zinc finger protein variants.

2.2 Results

As mentioned in Chapter 1.1, there exists a variety of techniques for constructing gene-length sequences from chemically synthesized DNA oligomers. Since one of the primary requirements listed above obviates the use of ligase or restriction enzymes, we originally focused on developing a technique that used mesophilic strand-displacing polymerase (such as phi29 DNA polymerase, which functions at 30°C), or a mesophilic, non-strand displacing polymerase (such as DNA Pol I, Large Klenow Fragment from *E. coli*, which functions at 37°C, or T4 DNA polymerase, which functions at 12°C) with DNA chew-back to run multiple cycles of oligomer assembly. Ideally, a substrate-bound primer would be used to anneal a longer oligomer gene part, extension would occur using a mesophilic DNA polymerase, the oligomer strand would be displaced/chewed-back at the 5'-end with T7 exonuclease, and the 3'-end of the next oligomer in the assembly cycle would anneal to the exposed sequence, priming the next extension step.

As a proof of concept, early method development focused on the use of oligomers making up the full linear expression template (including the 5' and 3'UTR) for yEGFP (yeast enhanced green fluorescent protein; F64L, S65T), with the fluorescent protein coding sequence taken from the plasmid pKT127. After comparing the fluorescence

obtained from four different linear templates using the PURExpress® In Vitro Protein synthesis kit (E6800, NEB), a design with a T7 terminator sequence, but without a poly-adenine (polyA) tail, was selected (see Figure 2.2.1).

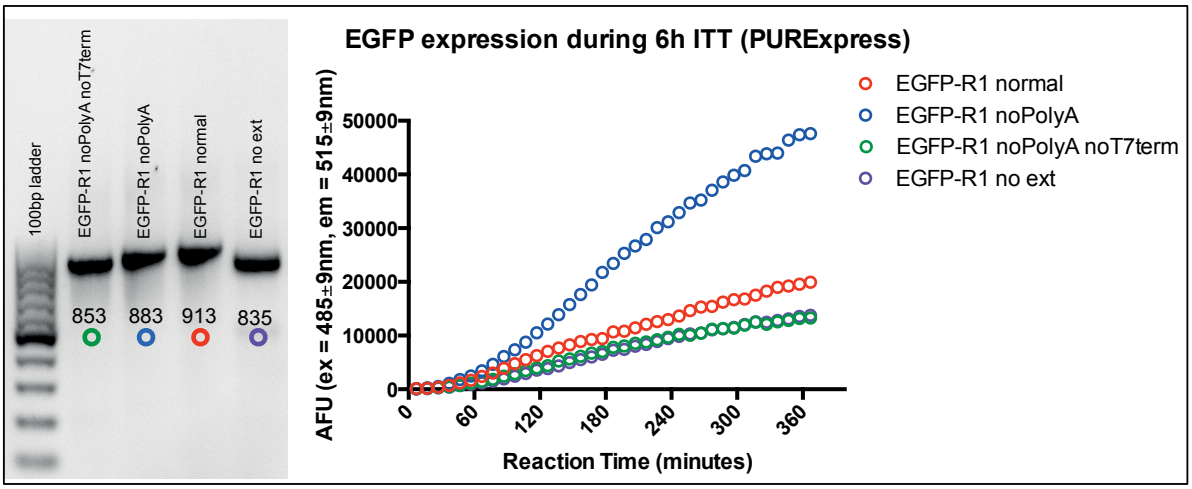


Figure 2.2.1: (left) Gel image of PCR amplified yEGFP linear template variants prepared for cell-free expression using the PURExpress system. (right) Graph of measured fluorescent output over time from batch, cell-free expression of each of the linear templates. The variant without a polyA tail (EGFP-R1 noPolyA, in blue) produced the highest level of fluorescence.

This sequence (883 bp; full linear template sequence given in Figure 2.5.1) was partitioned into 13 oligomers (sequences given in Table 2.5.1). The first 12 have a length of 90 nt, where 25 nt at the 3'-end of each is identical sequence to the 5'-end of the previous oligomer for annealing, and the last oligomer is 102 nt. The 25 nt of sequence overlap for annealing was not adjusted to homogenize annealing temperatures, and so a range of annealing temperatures (from 47 to 59°C) existed. If the overlapping regions had been modified using codon degeneracy to raise the annealing temperatures between oligomers while maintaining the proper protein coding sequence, the extension efficiency may have been improved.

With 25 nt reserved for annealing, the incorporation of an oligomer brought 65 nt of new sequence in each assembly step, which satisfied the 'minimal redundancy' requirement for the assembly technique. Since APE is not an exponential amplification process, each annealing step consisted of warming a solution containing the bead-bound priming sequence with the appropriate oligomer to 75°C, then cooling down to room temperature prior to adding in polymerase for extension. Shorter annealing regions could be used (20, 18, 16, 15, 10 and 5 nt overlaps were tested, with annealing temperatures of 50, 47, 41, 38, 30 and 14°C, respectively), but since these lengths result in less stringent

annealing, higher error rates, and lower efficiency of oligomer incorporation, they were not used. Overlaps below 15 nt were not capable of extending the annealed oligomer, which was also observed in a previously reported PCR assembly technique (263).

A range of oligomer lengths were tested as well, including 155mer and 200mers (purchased as Ultramers® from Integrated DNA Technologies), again maintaining 25 nt of overlap with neighboring oligomers. Although these oligomers offer a remarkably low level of sequence redundancy between addition steps, they were between 3-5 times more expensive (depending on synthesis scale) than 90mers because they are PAGE purified before being shipped. Additionally, Ultramer® orders contained extremely low amounts of the full-length product (see Figure 2.2.2). Due to these characteristics of Ultramers®, the gene assembly technique continued development with 90mers despite the less advantageous ratio of new sequence added in each assembly step. The 90mers were purchased with only standard desalting, and so each full-length oligomer is also contaminated with truncated products that result during column-based synthesis (see Figure 2.2.2). Due to the amount of truncated products found even in synthesized 90mer products, by utilizing a PCR purification kit to selectively remove oligomers shorter than 40 nt, the efficiency of annealing and extension reactions was improved (Figure 2.2.3).

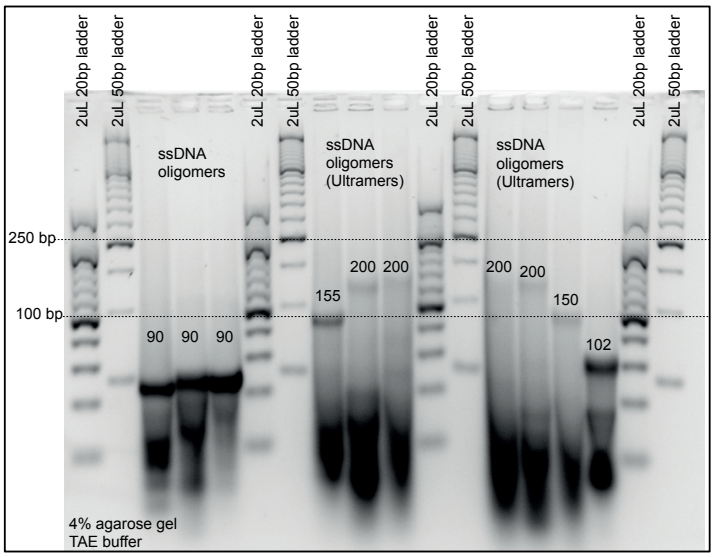


Figure 2.2.2 : Agarose gel image of single stranded DNA oligomers (ssDNA) of varying lengths. Ultramer® synthesis products offer very little full-length product (barely visible bands above 100bp ladder marker) compared to shorter oligomers. Even with PAGE purification, there is a large streak of truncated oligomers present in Ultramer® products.

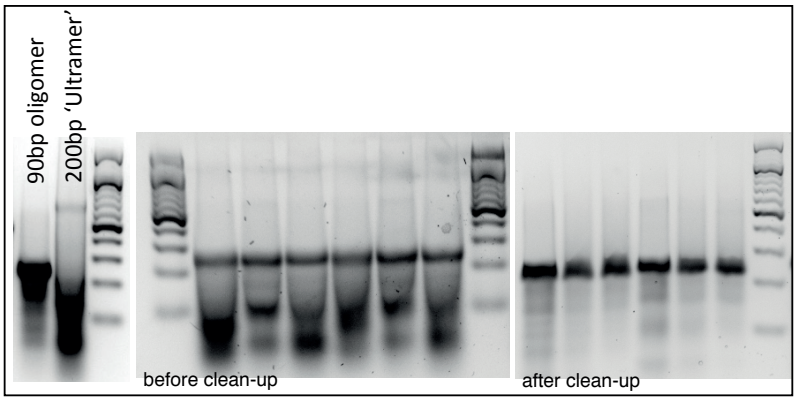


Figure 2.2.3 : Agarose gel image of 90 nt, single stranded DNA oligomers (ssDNA) before and after column purification to eliminate truncated synthesis products.

The T7 exonuclease chew-back reaction of the assembly cycle was difficult to optimize, and as a result, was replaced by an isothermal, ambient temperature, alkaline denaturation step (264). The excess of hydroxide ions in basic solutions (pH > 7) shields hydrogen bond formation between complementary strands of DNA, resulting in strand separation. This method was found to be particularly robust for dissociating DNA strands, and was easily incorporated into the gene assembly technique. A sodium hydroxide solution (NaOH, 0.15M) in water was used for all dissociation steps.

After comparing the activity of T4 DNA Polymerase and DNA Polymerase I, Large Klenow Fragment (5'→3' exo-, *E. coli*; NEB) for extension, we decided to continue only with DNA Polymerase I since it appeared to have better success at producing product. The original extension process was run for 30 minutes at room temperature or 30°C with one unit of DNA Pol I. Due to this long incubation time, the extension reaction limited the number of extension steps that could be performed in a day. Although DNA Pol I has a relatively high error rate compared to other high-fidelity polymerases, we were originally interested in developing an isothermal assembly process, and so we continued using DNA Pol I.

To facilitate a modular gene assembly technique, the process requires a solid substrate from which to 'grow' a gene by sequential addition of oligomers. The use of a solid-phase technique would also enable the *in situ* assembly of unique gene products within separated reaction chambers on a microfluidic device like MITOMI, and also provide the option of preparing large libraries of gene variants using automated liquid-handling

robotic systems. After experimenting with four varieties of streptavidin-coated magnetic beads (Dynabeads®, Invitrogen), we chose to use the MyOne™ Streptavidin T1 beads on the basis that they have the smallest diameter (1 µm), are specifically used for protein/nucleic acid applications, and due to their small diameter, have an increased binding capacity and lower sedimentation rate compared to the other varieties available. For the gene assembly technique, the first step in the process is incubating a volume of the T1 beads with a solution containing biotinylated 'initiator' oligomer, which anneals to the first oligomer in the assembly sequence. The initiator oligomer contains sequence from the 5' terminus of the antisense strand, and the first extension oligomer codes for the 3' terminus of the sense strand. In this way, the gene is assembled from what will become the C-terminus of the protein.

Since column synthesized oligomers are produced from 3' to 5', if truncated products are formed, they lack the 5'-terminal nucleotides. Since our assembly technique assembles the sense strand from the 3'-terminus, each extension step preferentially incorporates the next oligomer only if the previous step used a full-length oligomer, and therefore provides a full 25 nt of overlap for annealing (see Figure 2.2.4).

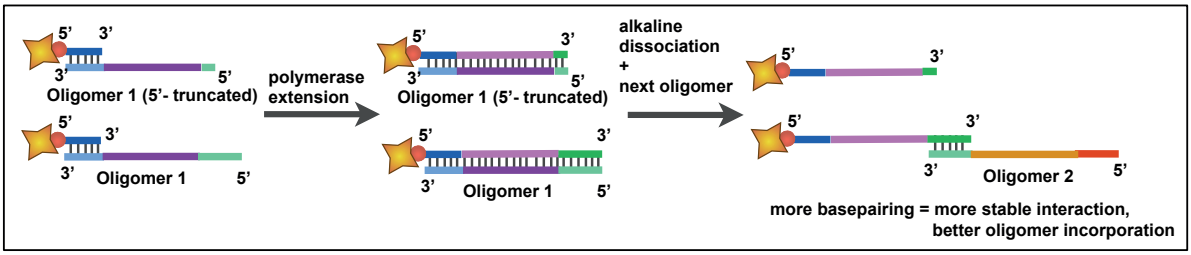


Figure 2.2.4 : Cartoon illustrating the selective incorporation and extension of full-length oligomers compared to truncated synthesis products.

To determine the limits of DNA crowding or extent of steric hindrance on the bead surface, a variety of bead volumes were incubated with the ssDNA, biotinylated initiator oligomer, but also with a range of dsDNA PCR products with one 5'-biotinylated strand (see Figure 2.2.5). The concentrations of the DNA products being tested were measured using a NanoDrop, and equal volumes of the DNA were incubated with different volumes of beads. After incubation, the beads were pelleted by magnetism, the supernatant was collected, and the DNA concentration was measured again to determine the amount of DNA remaining in the sample volume. The DNA bound to the beads was eluted by boiling the beads in water with 0.1% (v/v) SDS (sodium dodecyl sulfate), which disrupts the

streptavidin-biotin interaction, freeing the DNA from the bead surface. This eluate was then also measured to determine the DNA concentration. Due to limitations in the detection sensitivity of the NanoDrop, very small variations in DNA concentration were not detectable, but in general, the DNA mass balance was maintained (black lines in top part of Figure 2.2.5): pmol DNA_{total} = pmol $DNA_{bound\ to\ beads}$ + pmol $DNA_{in\ supernatant}$

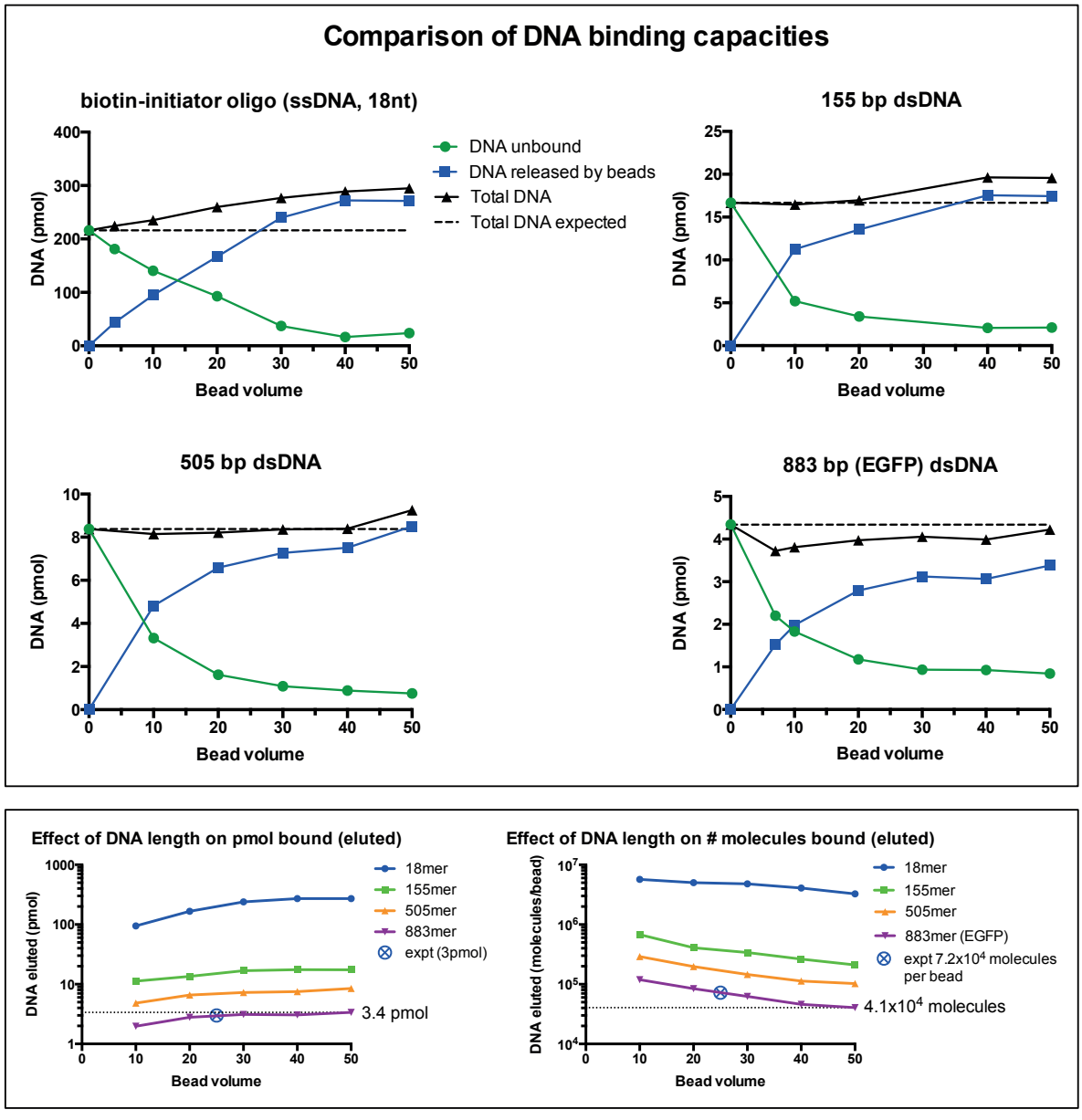


Figure 2.2.5 : Evaluating steric hindrance effects on streptavidin-coated magnetic beads as a function of DNA length. (Top) DNA saturation experiments using different bead volumes to determine the maximum amount of DNA that be bound as a function of DNA lengths. (Bottom) Summary of the amount of DNA eluted from beads after boiling.

By examining the effect of DNA length on the maximum amount of DNA bound per bead volume, it was determined that in order to provide sufficient space on 25 µL of beads to assemble a yEGFP gene, the initiator oligomer needed to be present at the same density, which was determined to be 3.4 pmol, which corresponds to roughly than 1.3% bead coverage using the initiator oligomer (Figure 2.2.5). This starting density was used for all subsequent gene assembly experiments (0.6 µL of 5 µM biotinylated initiator oligomer per 25 µL of beads).

To evaluate the successful annealing and extension of the yEGFP linear template, a set of 'checking' primers (see Table 2.5.1 for sequences; amplified products in Figure 2.2.6) were designed such that bead-bound extensions or SDS-eluted products could be amplified by PCR and visualized by gel electrophoresis. Each 'check' PCR utilized a primer with sequence identical to the biotinylated-initiator oligomer (but without biotin group) in combination with the primer needed to anneal at the extension step of interest. The efficiency of multiple extension reactions was improved, and amplification ('check' PCR) of yEGFP gene assembly was found to work for up to 9 rounds of extension (Figure 2.2.6).

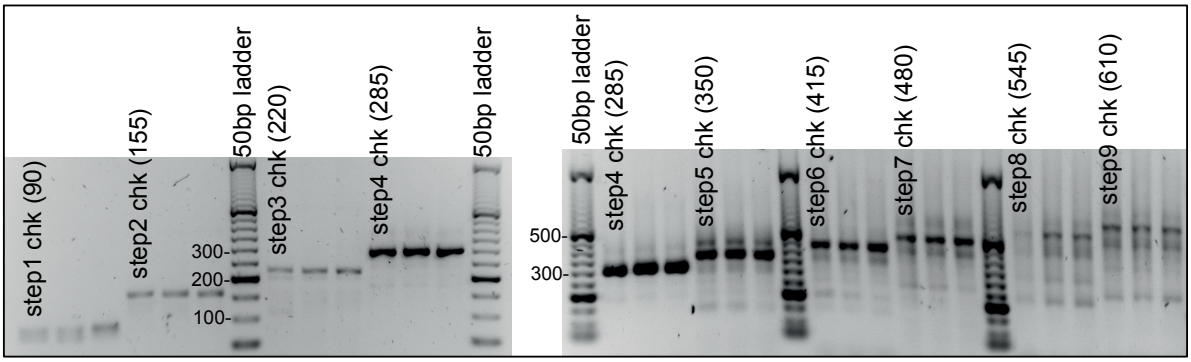


Figure 2.2.6 : Agarose gel image demonstrating that up to 9 consecutive steps of APE assembly can be achieved. Each band is the 'checking' PCR amplification product from a different APE reaction pool that was periodically sampled to evaluate successful extension.

It was not clear why the assembly process was not capable of functioning beyond 9 extension steps. It is possible that the incorporation of truncated oligomer synthesis products may have depleted the available annealing sites on the bead surface, or that steric hindrance effects were reducing the efficiency of the remaining assembly reactions. Additionally, each extension step was performed at 30°C, so it is possible that secondary structure of oligomers in later steps prevented them from being incorporated.

The assembly of a linear expression template for yEGFP was useful as a proof-of-concept for development of the asymmetric polymerase extension (APE) method. Since the intention was to eventually establish a pipeline for rapid protein engineering, and the set of protein variants that could be made from EGFP (265, 266) were not particularly interesting from a synthetic biology or genetic network perspective, we shifted our focus from fluorescent proteins to zinc finger proteins. This shift was primarily motivated by the fact that we had established our solid-phase assembly technique was functional for up to nine consecutive extension steps (610 bp), using DNA Pol I (Large Klenow Fragment) for extension, 90 nt long oligomers with 25 nt overlaps, and NaOH strand dissociation.

As mentioned in Chapter 1.3, C_2H_2 zinc finger (ZF) proteins are relatively compact protein domains, consisting of about 30 amino acids, which can be easily encoded within a single 90 nt oligomer. Zinc finger arrays (ZFAs), which recognize longer DNA target sites, can be produced by concatenating several ZF domains using artificial or naturally derived peptide linkers. Databases of ZF domain sequences and their consensus targets (146), crystal structures, as well as target prediction algorithms (76), are readily available after two decades of ZF research. Additionally, the MITOMI microfluidic device is an ideal tool for characterizing protein-DNA interactions. From these conclusions, we decided to pursue ZFA engineering as a proof-of-concept for our rapid protein engineering pipeline, and began to adapt the gene assembly protocol towards this application.

Since canonical ZF binding relies on the amino acid residues in positions -1 to 6 of the α-helix to dictate DNA target specificity, only these residues were changed to create ZFA variants that recognize specific DNA target sites. Other parts of the protein that are indirectly involved in DNA sequence recognition, called the framework, were taken directly from the three-finger murine transcription factor Zif268 (RCSB PBD 1AAY). The Zif268 coding sequence (90 aa) was converted to an *E. coli* (strain K12) codon optimized nucleotide sequence using JCat (Java Codon Adaptation Tool (267), www.jcat.de). The resulting 270 nt sequence was then partitioned into 5 oligomers: 3 'finger' oligomers, each containing the sequence coding for α-helix residues with flanking regions, and 2 'linking' oligomers containing sequences bridging the three 'finger' oligomers (Figure 2.2.7). Each oligomer has a 25 or 28 nt overlap with the oligomer preceding it for annealing. This way, all oligomers with the appropriate flanking sequence can be interchanged with each other since they contain the necessary complementary sequence for annealing. This allows single oligomers to be used in multiple zinc finger assemblies, but limits them to the same

position in the assembly process. Linking oligomers 'Link3-2' and 'Link2-1' are used for all assemblies, whereas libraries of O1F3 (Oligo1 Finger3), O3F2 (Oligo3 Finger2) and O5F1 (Oligo5 Finger1) oligos are used to generate different 3-finger assemblies (21 nt located in the colored regions of Figure 2.2.7). Oligomers were ordered from IDT (sequences given in Table 2.5.2) with standard desalting only and were rehydrated to 500 µM in 1x Tris-EDTA (TE) buffer for stock solutions, and diluted to 50 µM with PCR grade water for working solutions.

Only the three oligomers coding for the sequence variants must be ordered for each unique gene assembly, and since these are below the 90 nt pricing and synthesis scale threshold for oligomers ordered through IDT, they contain higher amounts of full-length product, and do not require column purification for APE assembly. Unlike the EGFP assembly oligomers, the five oligomers needed for ZFA assembly do not code for the entire linear, expression-ready template. APE assembly was used to generate the ZFA variants only, while other DNA parts needed for expression (5' and 3' UTR, affinity tag) are introduced downstream in PCR reactions.

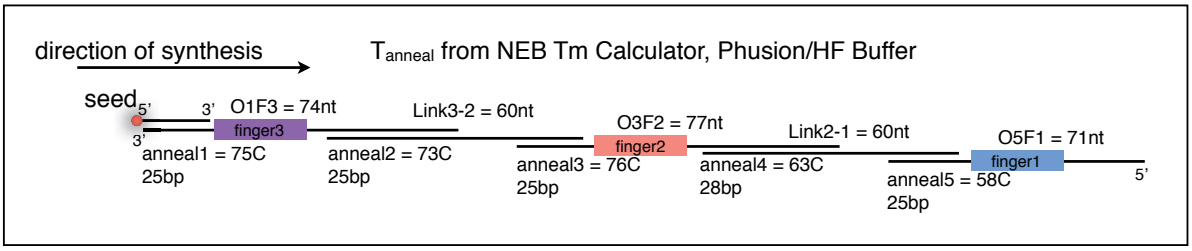


Figure 2.2.7 : Schematic of the oligomers used in APE assembly of tridactyl zinc finger transcription factors. Generic sequences are given in Supplementary Table 1. Synthesis occurs from the 3'-end of the sense strand (Finger3) to the 5'-end (Finger1). O1F3, O3F2, and O5F1 are unique oligos with 21 nt in the colored regions coding for recognition helix variants that target different DNA triplets. Link3-2 and Link2-1 are universal, and used in all APE assemblies.

To evaluate if functional ZFAs could be produced with APE assembly, oligomers were designed (Table 2.5.3) to express the unique specificity residues from wildtype Zif268 protein in addition to three artificial ZFA proteins (37-12, 92-1, 158-2) from a recent publication (156). These ZFAs were selected because their DNA-binding specificity had already been evaluated experimentally and would serve as ideal controls for our first experiments. We began by preparing gene assemblies for each of these ZFA variants using APE, and then developed a set PCRs to add on the remaining sequence needed to prepare linear templates for expression using the PURExpress system.

Prior to beginning cell-free expression experiments, the error-rate of the 5-step APE assembly was evaluated. APE assembly was carried out using either DNA Polymerase I, (Large Klenow Fragment; NEB) for reactions carried out entirely at room temperature, or Phusion High-Fidelity Polymerase (NEB) with brief annealing and extension steps on a thermal cycler. Both approaches used unpurified oligomers for the construction of ZFA variant 92-1 within a Zif268 backbone (framework). The assembly from each APE reaction was amplified by PCR using a high-fidelity polymerase.

Due to the formation of incomplete extension products during APE assembly rounds, PCR amplification can lead to the formation of multiple non-specific bands. To overcome this problem, we developed a modified band-stab technique (268) to isolate the band of interest, and reamplifyed it to reduce the amount of nonspecific products in downstream steps. This technique was performed on both the Klenow/room-temperature assembly and Phusion/thermocycled assembly PCR products. The resulting linear templates from the re-amplified band-stab products, in addition to the products from the Phusion assembly without band-stab, were cloned via Gibson assembly and transformed.

Colony PCR was used to verify clones with the expected size insert, and 32 clones from each assembly (Klenow +band-stab, Phusion +band-stab, and Phusion –band-stab) were sent for Sanger sequencing. The error rate of APE synthesis was determined by summing the frequency of all deletions, insertions and substitutions, then dividing the total number of error events by the total kilobases of interest (amplified PCR insert = 239 bp × # of clones with successful sequencing). The results of this analysis (Table 2.2.1) prompted the continued use of the high-fidelity Phusion polymerase with short annealing steps for all future APE assembly reaction. The band-stab step preparation did not appear to have any improvement on the error-rate when Phusion polymerase was used (0.785 errors/kb). We evaluated how the APE assembly error-rate compares with previously published gene-assembly techniques (Figure 2.2.8; Table 2.5.4), and it performs as well or better than the majority of other methods without the need for error-correction steps.

Table 2.2.1 : Error rate analysis for APE assembly

ZF synthesis method (polymerase used for extension)	# of full-length sequence reads (out of 32)	Total kb of interest sequenced	# deletion events	# insertion events	# substitution events	Total error events	Error/kb
Phusion -band stab	32	7.648	4	2	0	6	0.785
Phusion +band stab	32	7.648	5	0	1	6	0.785
Klenow +band stab	28	6.692	15	1	1	17	2.54

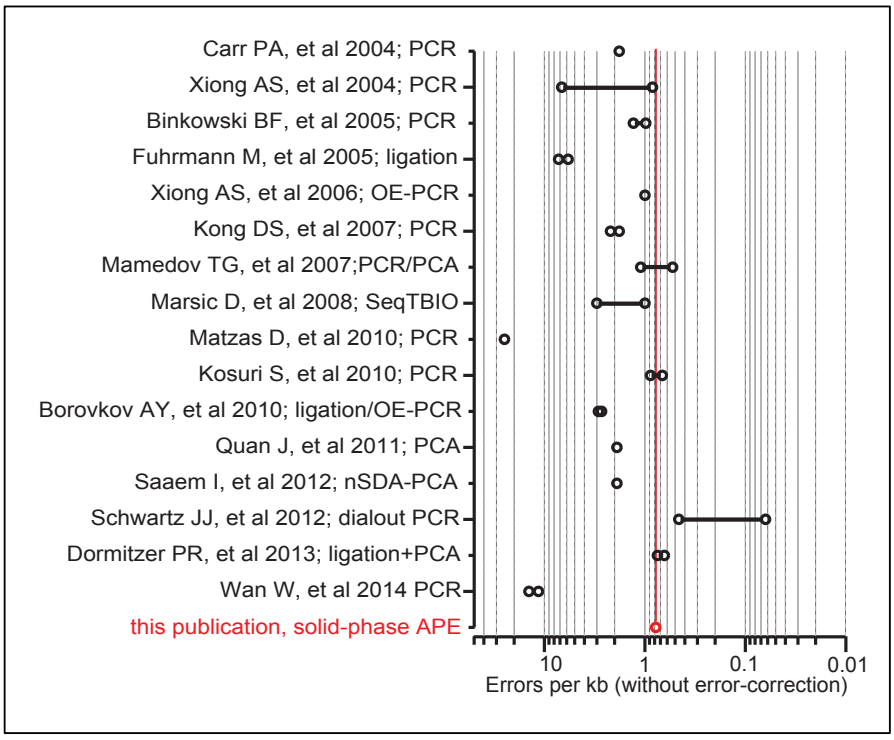


Figure 2.2.8: Comparison of APE error rate with values from previously published gene assembly techniques. A line between two points indicates a range of error rates from different experimental conditions.

To optimize the ZFA templates for on-chip expression, we fused the ZFA assemblies to a C-terminal EGFP domain using a short, but rigid, proline-alanine linker by PCR (full sequence of Zif268 linear template in Figure 2.5.2; oligomer sequences in Table 2.5.5). The EGFP domain is useful for enabling fluorescent confirmation of protein production during cell-free expression, but was also used as an affinity tag for selective purification of properly transcribed and translated product, and as a tool for quantifying the amount of protein trapped during microfluidic analysis with MITOMI. The final APE assembly process for ZFA engineering is illustrated in Figure 2.2.9, along with a timeline for array printing and operation of the MITOMI assays.

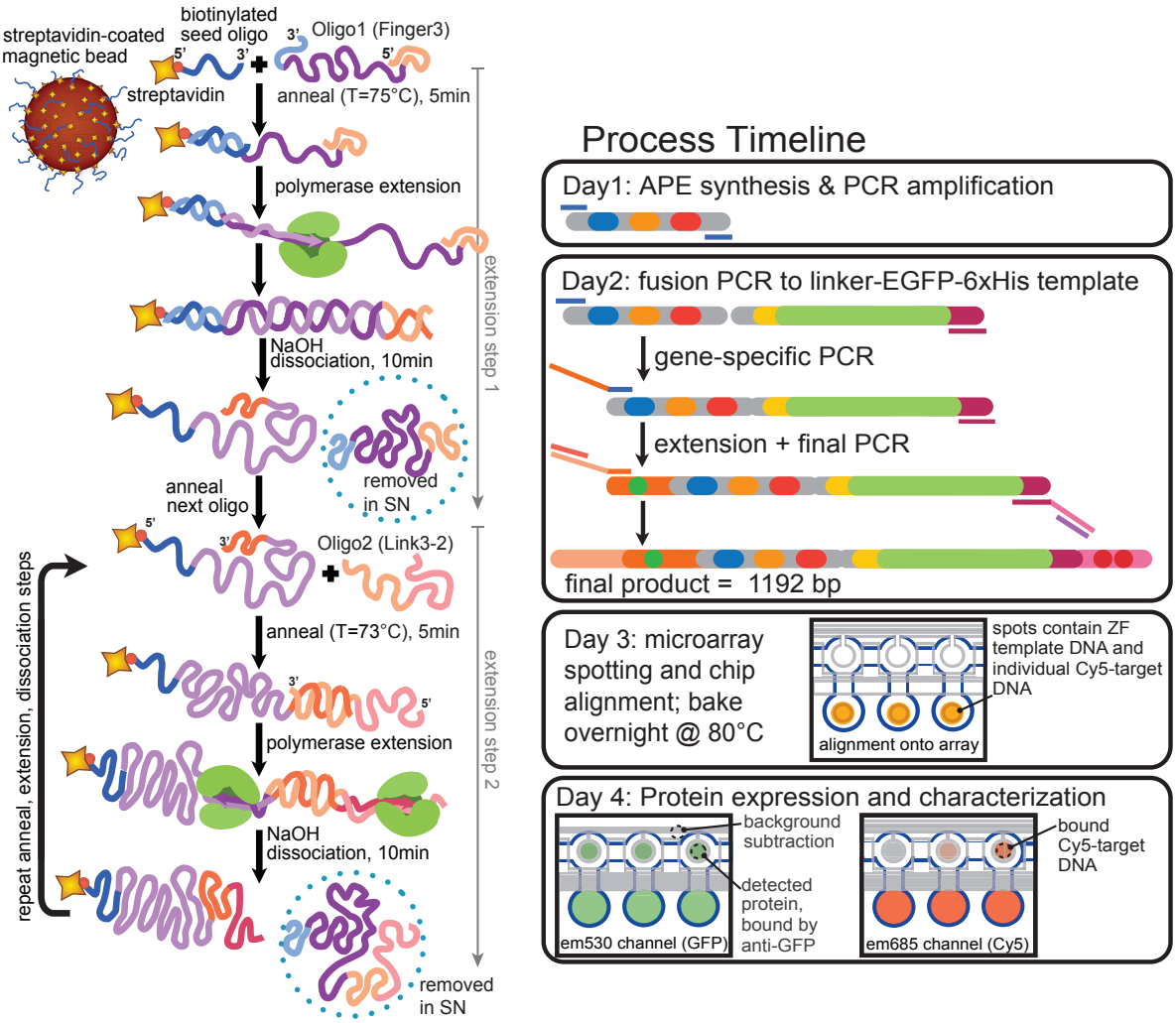


Figure 2.2.9 : (Left) Schematic of the APE solid-phase gene assembly technique, showing assembly through the first two extension steps. (Right) Process timeline from gene assembly to protein characterization.

2.3 Discussion

In Chapter 2, we discuss the development of a novel, solid-phase gene assembly technique, first for the construction of yEGFP, and later applied towards the construction of ZFA variants. By evaluating the error rate of assembly using both DNA Pol I (Large Klenow Fragment) and Phusion polymerase, we were able to optimize the assembly process. The highest error rate was observed in the colonies cloned with the Klenow APE assembly method. Since the EGFP affinity tag is appended to the assembly product by PCR, the majority of these products would have resulted in nonfunctional ZFAs, since they code for early stop codons or introduce frame-shifts which would prevent the EGFP tag

from folding and being immunoprecipitated in a MITOMI experiment. In this manner, even truncated/non-specific PCR products that are formed during the final steps of linear template preparation are not problematic, since those products will not be included in the DNA target binding analysis. The Phusion APE assembly products yielded significantly lower erroneous sequences, with no effect seen from the band-stab PCR. The high error rate seen in the DNA Pol I (Large Klenow Fragment) reaction product can be partly explained by comparing the relative error rates documented for Klenow (0.018 errors/kb, due to removal of 5'→3' exonuclease) and Phusion (0.00044 errors/kb) polymerases. Additionally, since the Phusion-based assembly technique required brief annealing steps at elevated temperatures while the Klenow-based technique was performed at room-temperature, one can reasonably expect that the high-temperature annealing improved the stringency of oligomer incorporation.

Although APE assembly was found to be limited to 9 consecutive extension steps (610 bp), this length of sequence is more than adequate for the assembly of ZFA variants. To overcome this length limitation of APE assembly and construct longer genetic parts, it would be possible to begin multiple APE assembly reactions in parallel, then use PCR to assemble the products from each of these reactions towards the construction of gene parts larger than 1 kb. The true power of the APE assembly method lies in its ability to generate genetic parts with high variability or repetitive motifs, followed by conventional PCR to introduce the constant regions of the gene, which are unchanged across variants. Since the oligomers are added sequentially and unincorporated DNA is separated from the bead-bound strands during buffer exchange steps, the APE assembly technique offers a lower risk of chimeric product formation compared to one-pot ligase or PCA-based assembly techniques. Column-synthesized oligomers can be used directly without purification, and the final, PCR amplified linear template can be used directly for microarray printing and on-chip expression. Since the assembly technique is magnetic-bead based, it is amenable to scale-up with automated liquid handling platforms.

2.4 Methods

2.4.1 Asymmetric primer extension (APE) assembly

Prior to gene-assembly, an aliquot (800-1200 µL) of MyOne™ Streptavidin T1 beads (Life Technologies) are placed in a 1.5 mL Eppendorf tube, pelleted using a magnetic stand, and resuspended in an equivalent volume of 0.2M NaOH in water. The beads are preconditioned for at least 1 hour at room temperature before use, then stored at 4°C for longer conditioning times. These conditioned beads can be used for up to one month after being suspended in the 0.2M NaOH solution. It was previously found that preconditioning of beads in NaOH releases labile streptavidin monomers (269), and in our hands this translated to a reduction in non-specific PCR bands during intermediate quality control PCR steps and during amplification of the final assembly. Each individual assembly reaction requires 25 µL of preconditioned beads. Lower bead quantities may work as well, but to account for losses during washing, buffer exchanges and transfer steps, we have continued using 25 µL with consistent success. Larger reactions are also possible by scaling all volumes accordingly. This is particularly useful during the creation of zinc finger combinatoric array variants. By starting with a large pool of beads, beginning the assembly together with the same Oligo1Fin3 and Link3-2, followed by partitioning the pool into smaller volumes and continuing assembly with different Oligo3Fin2 parts in separated reactions, followed by a final partitioning for Oligo5Fin1 parts, where the final volume in each Oligo5Fin1 assembly is 25 µL, many different genes can be assembled within the same workflow.

For a single APE reaction, 25 µL of preconditioned beads are pelleted using a magnetic stand (Invitrogen DynaMag™-Spin) for 30-60s until the solution is clear. The supernatant (0.2M NaOH) is carefully aspirated. The beads are then washed twice with 25 µL of 1x binding and washing buffer (B&W; 2x contains 10mM Tris-HCl pH 7.5, 1mM EDTA, 2M NaCl) containing 0.01% (v/v) Tween20 (BW+Tween; to reduce non-specific binding (270)). Each washing step involves adding the wash solution, mixing the solution by aspiration until the beads are resuspended, then pelleting the beads and removing the supernatant. Then the beads are pelleted again and resuspended in 25 µL of 2x B&W Buffer (without Tween20), to which 25 µL of 'seed' oligomer solution (0.12 µM biotinylated seed oligomer in PCR grade water; Supplementary Table 1) is added and mixed. This

mixture is incubated at room temperature for at least 15 minutes on a lab rotisserie. Following incubation, the beads are pelleted against the magnetic stand, washed twice with 25 µL of 1x HF Buffer without detergent (Phusion® HF Buffer Detergent-free (5x), New England Biolabs) to prevent bubble formation during resuspension, and finally resuspended in 25 µL Oligo1Fin3 extension mix (final concentrations: 1x HF Buffer with detergent, 0.2 mM dNTPs, 5% DMSO, 8 µM Oligo1Fin3, 0.3 units Phusion High-Fidelity Polymerase (NEB)).

This mixture is then placed on a thermocycler and run through a brief annealing and extension routine (5.5min at T_{anneal}, 2min at 72°C, then hold at 25°C; T_{anneal} for each oligomer is given in Fig. 2.2.7). The tube is removed from the thermocycler, the beads are pelleted, and the supernatant is removed and discarded. The beads are then washed twice with 50 µL 1x SSC buffer (saline sodium citrate, Sigma), resuspended in 50 µL 0.15M NaOH and incubated at room temperature on a rotisserie for 10 minutes to facilitate strand dissociation. The beads are then pelleted, the supernatant is removed, and the beads are washed once with 50 µL 0.15M NaOH, once with 50 µL 1x BW+Tween, and once with 50 µL 1x HF buffer without detergent. The beads are then resuspended in Oligo2 (Link3-2) extension mix (same recipe as for Oligo1, except Link3-2 is used), then placed back on the thermocycler, and run through the annealing and extension routine, where T_{anneal} has been adjusted to the temperature required for this annealing reaction. This procedure of extension, strand dissociation via 0.15 M NaOH, and buffer exchanges is repeated for each oligomer in the assembly. After the final extension reaction (Oligo5Fin1), there is no NaOH dissociation step. Instead, the beads are pelleted, washed twice with 50 µL 1x SSC buffer, and resuspended in a final volume of 20 µL 10 mM Tris-Cl pH 8.5 buffer.

The beads in Tris-Cl buffer from the final extension step are used directly as template for a PCR amplification of the complete 5-step assembly product. PCR primers were designed to amplify the 239 bp product (Table 2.5.2). For a 20 µL PCR reaction, final concentrations are as follows: 1x HF Buffer with detergent (New England Biolabs), 0.2mM dNTPs, 5% DMSO, 0.5 µM each primer (assembly check-f and –r), 0.6 µL suspension of beads in Tris-Cl (template), 0.3 units Phusion High-Fidelity Polymerase; touchdown PCR: 98°C, 30s; 74>72°C, 30s then 17 cycles at 71°C, 30s ; 72°C, 30s). 4 µL of this PCR is then run on a 2% agarose gel with 0.4x GelGreen (Biotium) at 110V for 1h. Due to non-specific primer interactions with the template and interactions with truncated

assembly products, some PCR amplifications can result in the formation of multiple truncated bands, the highest of which is the complete assembly product. To overcome this problem, we have taken advantage of a modified band-stab technique a(268) to isolate the band of interest and re-amplifying it to reduce the amount of nonspecific products in downstream steps. Briefly, the 2% gel is imaged using a blue-light transilluminator and the band of interest is captured using a 200 µL pipette tip, with the end cut off about 1 cm from the tip, by stabbing into the gel at the location of the band. The pipette tip is then placed into a 1 mL Eppendorf tube, and the agarose gel core inside the pipette tip is pushed out using a second sterile pipette tip. 20 µL of Qiagen EB buffer (10 mM Tris-Cl, pH 8.5) is added to the agarose gel sample, briefly vortexed, centrifuged, and incubated at 80°C for at least 10 minutes with the tube cap closed. The sample is then vortexed and centrifuged again, before a 0.25 µL sample of the buffer is taken as template for a second PCR amplification. This PCR is prepared and thermocycled following the same recipe from the first PCR (assembly check PCR; Figure 2.5.3).

2.4.2 Expression-ready linear template preparation

Following the second assembly check PCR, the core region coding for the three linked a-helices is complete, but the final template will consist of a C-terminal proline-linker and EGFP fusion, a 6x histidine tag, and 5' and 3' UTRs for expression within a cell-free, transcription/translation mixture. All of these parts are added to the zinc-finger assembly via 4 different PCRs: a fusion PCR (for adding the proline-linker, EGFP domain, and 6x-histidine tag; results in a 1019bp product), a gene-specific PCR (for adding part of the 5' and 3' UTRs; results in a 1084bp product), and an extension+final 2-step PCR (which completes the template construction and amplifies the full-length product of 1192bp; oligomer sequences in Table 2.5.5). The fusion PCR requires two templates: the 239bp zinc-finger construct and a previously amplified EGFP domain from the pKT127 plasmid, including a 5'-proline linker and 3'-6x histidine tag. Briefly, in a 20 µL PCR containing 1x HF Buffer with detergent, 0.2 mM dNTPs, 5% DMSO, 0.5 µM each primer (Prolinker-EGFP-f and EGFP-6His-r), 1 ng of pKT127, and 0.3 units Phusion High-Fidelity Polymerase are thermocycled for 25 cycles (98°C, 30s; 61.7°C, 30s; 72°C, 1min).

This product (Prolink-EGFP-6His, 805nt) is used without purification in the fusion PCR. The fusion PCR is carried out in 2 steps, the first reaction contains all the necessary ingredients for PCR except the nucleotide mix and polymerase. A 15 µL reaction is prepared containing 1x HF Buffer with detergent, 5% DMSO, 0.5 µM of each primer (assembly check-f and EGFP-6His-r), and 0.25 µL each of the assembly PCR from band-stab and Prolink-EGFP-6His PCR. This mixture is placed on a thermocycler and heated to 98°C for 4min, then cooled down (10% ramp) to 25°C for annealing. Then 5 µL of an extension mixture (1x HF Buffer with detergent, 0.2 mM dNTPs, 0.3 units Phusion) is spiked into the annealing mixture (20 µL total) and cycled 20 times (98°C, 30s; 72°C, 30s; 72°C, 1min). The gene-specific PCR uses the product generated in the fusion PCR as template. A 20 µL reaction is prepared: 1x HF Buffer with detergent, 0.2 mM dNTPs, 5% DMSO, 0.5 µM each primer (genespecific-f and EGFP-6His-r), 0.25 µL of the fusion PCR, and 0.3 units Phusion polymerase. This reaction is cycled using a short touchdown PCR (98°C, 30s; 75>72°C, 30s; 72°C, 1min), followed by 16 cycles (98°C, 30s; 72°C, 30s; 72°C, 1min). The 2-step extension+final PCR uses the product generated in the gene-specific PCR as template. In this reaction, it is very important to use the HF Buffer without detergent, since this product will be used directly for microarray spotting and the presence of detergent will result in large spots. A 20 µL reaction is prepared: 1x HF Buffer without detergent, 0.2 mM dNTPs, 5% DMSO, 2.5 nM each primer (extension-f and -r), 0.25 µL of 1:10 diluted gene-specific PCR (in Tris-Cl or water), and 0.3 units Phusion polymerase. This mixture is thermocycled 10 times (98°C, 30s; 61°C, 30s; 72°C, 1min). Then the reaction is kept at 72°C for 2 minutes, and cooled to 25°C. At this point, 0.1 µL of each final_highTm primer (50 µM stock; final 0.25 µM in 20.2 µL) are spiked into the mixture, and it is thermocycled again via a short touchdown PCR, 98°C, 30s; 75>72°C, 30s; 72°C, 1min, then 20 cycles 98°C, 30s; 71°C, 30s; 72°C, 1min. Successful amplification is determined by running 1.5 µL of the product on a 1% agarose gel, and checking for the 1192bp product (see Figure 2.5.3).

2.4.3 Error Rate Analysis

Gene synthesis reaction on beads was carried out using either DNA Polymerase I, Large (Klenow) Fragment (NEB) for APE assembly reactions entirely carried out at room

temperature, or Phusion High-Fidelity Polymerase with brief annealing and extension steps on a thermal cycler. Both approaches used unpurified oligomers for the construction of zinc finger array 92-1 within a Zif268 backbone. For Klenow assembly error analysis, following 20 cycles of PCR using Phusion polymerase for amplification of template detached from beads using an SDS-boiling and reannealing protocol, the reaction was run on an agarose gel and the product band was gel-stabbed, and a second PCR (20 cycles with Phusion polymerase) was run. For Phusion assembly error analysis, following 20 cycles of PCR using Phusion polymerase for amplification of the template attached to beads in 1x SSC buffer, the reaction was run on an agarose gel and the product band was gel-stabbed, and a second PCR (20 cycles) was run. The PCR product (239bp) from each of the band-stab PCRs was purified and cloned via Gibson assembly into the pUC19 plasmid with assembly-check overhangs. In addition, the PCR from the Phusion assembly was used without the band-stab procedure, to determine whether the band-stab has an effect on error rate.

Chemically competent DH5α *E. coli* cell aliquots (30 µL) were transformed with 1.5 µL of each Gibson assembly product via heat shock (30s at 42°C), recovered in 300µl SOC medium for 1 hour at 37°C, and plated on ampicillin plates for overnight growth at 37°C. Colonies from each plate (Klenow+band stab, Phusion+band stab, Phusion no stab) were picked with a sterile 200 µL pipette tip, briefly stirred in 20 µL PCR-grade water, boiled for 15min, and centrifuged. Water from the colony boils was used as template for an insert-check PCR using the primers pUC19-f and pUC19-r with Phusion polymerase. All of these PCR reactions were run on 1.5% agarose gels with GelRed to determine which colonies had the correct sized insert. Colonies were picked and analyzed in this way until 32 colony PCRs for each assembly method were identified with a single band corresponding to the correct sized insert. The insert-check PCRs were submitted for Sanger sequencing in a 96-well plate without PCR cleanup (Microsynth AG), and the resulting sequencing reads were aligned with the expected sequence to analyze the error-rate and identify which types of errors were prevalent.

2.5 Supplementary Figures and Tables

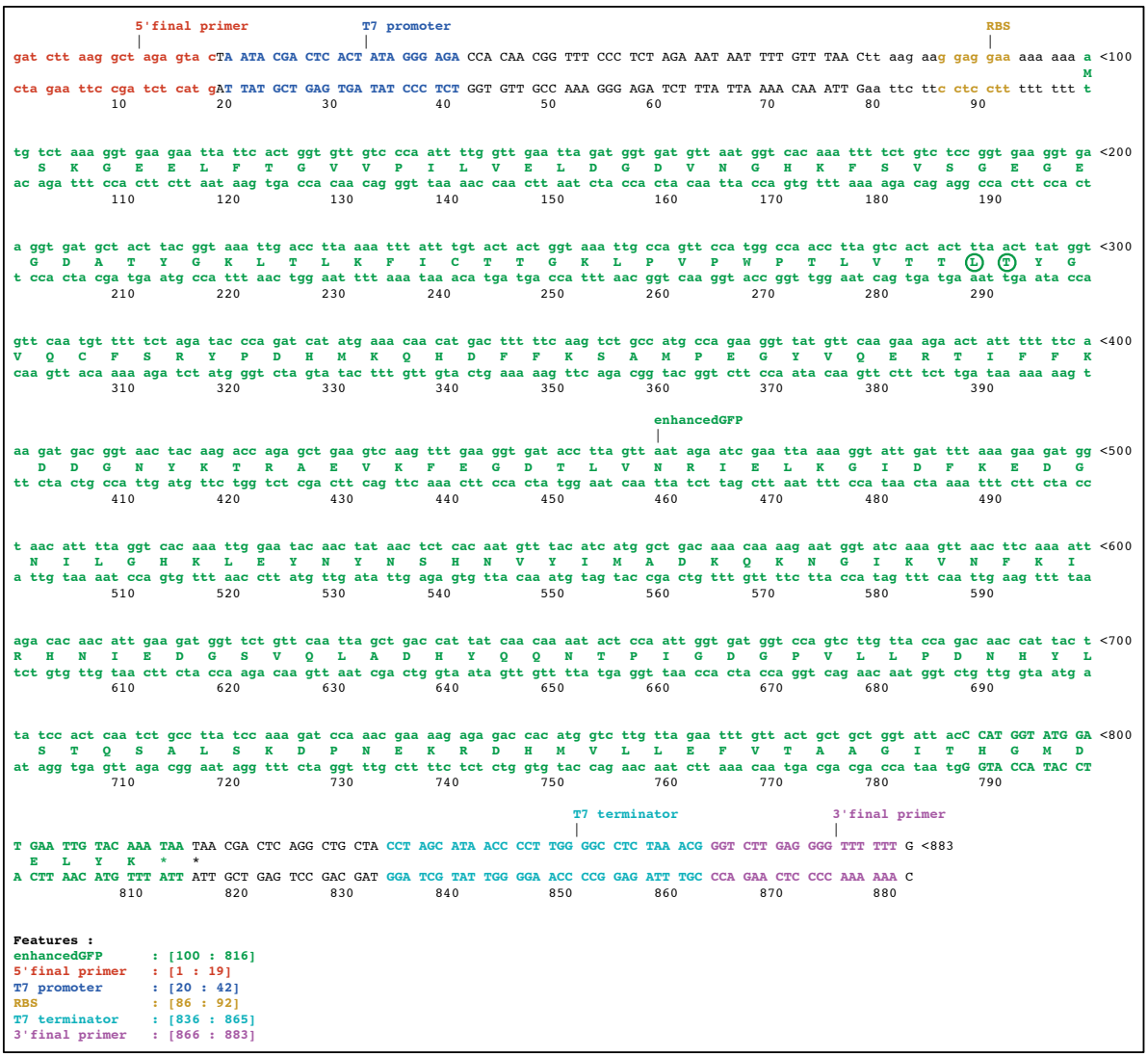


Figure 2.5.1: Complete sequence of expression-ready linear template for yEGFP, which was used to design oligomers for development of asymmetric polymerase extension (APE) assembly.

Table 2.5.1: Oligomer sequences for assembly of yEGFP linear expression construct

Oligomer name (length)	Sequence (5' to 3')
biotin-3'final (initiator oligomer; 18 nt)	biotin-CAAAAAACCCCTCAAGAC
EGFP-ExtOligo1 (90 nt)	**GTATGGATGAATTGTACAAATAATA**ACGACTCAGGCTGCTACCTAGCATAA… CCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTG
EGFP-ExtOligo2 (90 nt)	**tccaaacgaaaagagagaccacatg**gtcttgttagaatttgttactgctgct… ggtattacCCATG**GTATGGATGAATTGTACAAATAATA**
EGFP-ExtOligo3 (90 nt)	**ggtgatggtccagtcttgttaccag**acaaccattacttatccactcaatctg… ccttatccaaaga**tccaaacgaaaagagagaccacatg**
EGFP-ExtOligo4 (90 nt)	**ttagacacaacattgaagatggttc**tgttcaattagctgaccattatcaacaa… aatactccaatt**ggtgatggtccagtcttgttaccag**
EGFP-ExtOligo5 (90 nt)	**taactctcacaatgtttacatcatg**gctgacaaacaaaagaatggtatcaaagt… taacttcaaaa**ttagacacaacattgaagatggttc**
EGFP-ExtOligo6 (90 nt)	**gaattaaaaggtattgattttaaag**aagatggtaacattttaggtcacaaattg… gaatacaacta**taactctcacaatgtttacatcatg**
EGFP-ExtOligo7 (90 nt)	**atgacggtaactacaagaccagagc**tgaagtcaagtttgaaggtgataccttag… ttaatagaatc**gaattaaaaggtattgattttaaag**
EGFP-ExtOligo8 (90 nt)	**acatgacttttttcaagtctgccatg**ccagaaggttatgttcaagaaagaactatt… tttttcaaag**atgacggtaactacaagaccagagc**
EGFP-ExtOligo9 (90 nt)	**accttagtcactactttaacttatg**gtgttcaatgtttttctagatacccagatc… atatgaaaca**acatgacttttttcaagtctgccatg**
EGFP-ExtOligo10 (90 nt)	**ctacttacggtaaattgaccttaaa**atttatttgtactactggtaaattgccagt… tccatggcca**accttagtcactactttaacttatg**
EGFP-ExtOligo11 (90 nt)	**ggttgaattagatggtgatgttaat**ggtcacaaattttctgtctccggtgaaggt… gaaggtgatg**ctacttacggtaaattgaccttaaa**
EGFP-ExtOligo12 (90 nt)	**cttaagaaggaggaaaaaaaaatg**tctaaaggtgaagaattattcactggtgttg… tcccaatttt**ggttgaattagatggtgatgttaat**
EGFP-ExtOligo13 (102 nt)	gatcttaaggctagagtacTAATACGACTCACTATAGGGAGACCACAACGGTTTC… CCTCTAGAAATAATTTTGTTTAA**Cttaagaaggaggaaaaaaaaatg**
3'final check (18 nt)	CAAAAAACCCCTCAAGAC
Step1 check (20 nt)	GTATGGATGAATTGTACAAA
Step2 check (18 nt)	tccaaacgaaaagagaga
Step3 check (21 nt)	ggtgatggtccagtcttgtta
Step4 check (20 nt)	ttagacacaacattgaagat
Step5 check (25 nt)	taactctcacaatgtttacatcatg
Step6 check (23 nt)	gaattaaaaggtattgattttaa
Step7 check (22 nt)	atgacggtaactacaagaccag
Step8 check (20 nt)	acatgacttttttcaagtctg
Step9 check (25 nt)	accttagtcactactttaacttatg
Step10 check (19 nt)	ctacttacggtaaattgac
Step11 check (25 nt)	ggttgaattagatggtgatgttaat
Step12 check (20 nt)	cttaagaaggaggaaaaaaa
Step13 check/5'final (19 nt)	gatcttaaggctagagtac

Table 2.5.2: APE ZFA assembly oligomer sequences

Oligomer Name	Sequence (5' to 3')
Oligo1Finger3 (O1F3) 74 nt	ttgcgacatctgcggtcgtaaattcgct**XXXXXXXXXXXXXXXXXXXX**cacaccaaaatccacctgcgtcaga
Oligo2Linker3-2 (O2link3-2), 60 nt	cacatccgtacccacaccggtgaaaaaccgttcgcttgcgacatctgcggtcgtaaattc
Oligo3Finger2 (O3F2) 77 nt	CCAatgtagaatttgtatgagaaatttctct**XXXXXXXXXXXXXXXXXXXX**cacatccgtacccacaccggtgaaa
Oligo4Linker2-1 (O4link2-1), 60 nt	catattagaattcatactggacaaaAACCATTCCAatgtagaatttgtatgagaaatttc
Oligo5Finger1 (O5F1) 71 nt	tgaatcttgcgaccgtcgtttctct**XXXXXXXXXXXXXXXXXXXX**catattagaattcatactggacaaa
Biotinylated initiator oligo, 25 nt	biotin-tctgacgcaggtggattttggtgtg
Assembly check3-f (5'chk3-finger1)	tgaatcttgcgaccgtcgtttctct
Assembly check2-f (5'chk2-finger2)	catattagaattcatactggacaaaaaccattcc
Assembly check1-f (5'chk1-finger3)	cacatccgtacccacaccggtg
Assembly check-r (zif268-3'chk)	tctgacgcaggtggattttggtg

Where xxxxx indicates location of 21 nt coding for 7 amino acids in positions -1 to 6 of recognition helix. Oligomers 1-5 are used for modular construction of ZFAs. The Assembly check primers were used in initial tests to verify the success of oligomer extension after various rounds of APE assembly. For PCR amplification of full length product after the 5-step assembly, only Assembly check3-f and check-r are needed.

Table 2.5.3 : APE assembly oligomers for Zif268, 37-12, 92-1, 158-2

Oligomer Name	Sequence (5' to 3')
Zif268_Oligo1fin3	**ttgcgacatctgcggtcgtaaattc**gct**cgttctgacgaacgtaaacgtcacaccaaaatccacctgcgtcaga**
Zif268_Oligo3fin2	**CCAatgtagaatttgtatgagaaatttc**tct**cgttctgaccacctgaccacccacatccgtacccacaccggtgaaa**
Zif268_Oligo5fin1	**tgaatcttgcgaccgtcgtttctctcgttctgacgaactgacccgtcatattagaattcatactggacaaa**
37-12Oligo1fin3	ttgcgacatctgcggtcgtaaattcgct**CGTCACGACCAGCTGACCCGT**cacaccaaaatccacctgcgtcaga
37-12Oligo3fin2	CCAatgtagaatttgtatgagaaatttctct**GACCGTGCTAACCTGCGTCGT**cacatccgtacccacaccggtgaaa
37-12Oligo5fin1	tgaatcttgcgaccgtcgtttctct**CGTAACTTCATCCTGCAGCGT**catattagaattcatactggacaaa
92-1Oligo1fin3	ttgcgacatctgcggtcgtaaattcgct**GAACGTGGTAACCTGACCCGT**cacaccaaaatccacctgcgtcaga
92-1Oligo3fin2	CCAatgtagaatttgtatgagaaatttctct**CAGCGTTCTTCTCTGGTTCGT**cacatccgtacccacaccggtgaaa
92-1Oligo5fin1	tgaatcttgcgaccgtcgtttctct**GACTCTCCGACCCTGCGTCGT**catattagaattcatactggacaaa
158-2Oligo1fin3	ttgcgacatctgcggtcgtaaattcgct**CAGTCTACCTCTCTGCAGCGT**cacaccaaaatccacctgcgtcaga
158-2Oligo3fin2	CCAatgtagaatttgtatgagaaatttctct**GTTCGTCACAACCTGACCCGT**cacatccgtacccacaccggtgaaa
158-2Oligo5fin1	tgaatcttgcgaccgtcgtttctct**GACAAAACCAAACTGCGTGTT**catattagaattcatactggacaaa

Table 2.5.4 : Gene synthesis technique error rate comparison (for Figure 2.2.8)

year	Previously reported error rates (publication reference)	Assembly technique	Before
2004	PA Carr et al, Protein mediated error correction for de novo DNA synth, NAR 2004	PCR assembly	1.8 error/kb
2004	Xiong A-S et al, A simple, rapid, high fidelity and cost-effective PCR-based 2step DNA synthesis method for long gene sequences, NAR 2004	PCR assembly PTDS	0.84-6.72 error/kb
2005	Binkowski BF et al, Correcting errors in synthetic DNA through consensus shuffling, NAR 2005	PCR assembly	0.98-1.3 errors/kb
2005	Fuhrmann M et al, Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage, NAR 2005	ligation	5.8-7.2 error/kb
2006	Xiong A-S et al, PCR-based accurate synthesis of long DNA sequences, NatureProtocols 2006	PCR assembly	<1 error/kb
2007	Kong DS et al, Parallel gene synthesis in a microfluidic device, NAR 2007	PCR	1.8-2.2 error/kb
2007	Mamedov TG et al, Rational denovo gene synth by rapid PCA and expression of endothelial protein-c and thrombin receptor genes, J Biotechnol 2007	PCR, PCA, fastPCA	0.53-1.1 error/kb
2008	Marsic D et al, PCR-based gene synthesis to produce recombinant proteins for crystallization, BMC Biotech 2008	SeqTBIO	1-3 error/kb
2009	Ye H et al, Experimental analysis of gene assembly with TopDown one-step real-time gene synth, NAR 2009	PCA assembly	no consideration of error rate
2010	Matzas M et al, High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using HT pyrosequencing, Nature Biotech 2010	PCR assembly	25 error/kb (starting oligomers)
2010	Kosuri S et al, Scalable gene synth by selective amplification of DNA pools from highfidelity microchips, Nat Biotech 2010	PCR assembly	0.67-0.88 errors/kb
2010	Borovkov AY et al, High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligos, NAR 2010	Ligation, Overlapping PCR	2.7-2.9 errors/kb
2011	Quan J et al, Parallel onchip gene synth and application to optimization of protein expression, Nat Biotech 2011	PCA	1.9 error/kb
2012	Saaem I et al, Error correction of microchip synthesized genes using Surveyor nuclease, NAR 2012	nSDA-PCA, PCR	1.9 error/kb
2012	Schwartz JJ et al, Accurate gene snthesis with tag-directed retrieval of sequence-verified DNA molecules, Nat Methods, 2012	Dialout PCR	
2012	Ma S et al, Error Correction in gene synthesis technology, Trends Biotechnol, 2012	various	
2013	Dormitzer PR et al, Synthetic generation of influenza vaccine viruses for rapid response to pandemics, Science translation med, 2013	Ligation+PCA	0.64-0.75errors/kb
2014	Wan W et al, Error removal in microchip-synthesized DNA using immobilized MutS, NAR 2014	PCR	11.44-14.25/kb
2014	Currin A, et al, SpeedyGenes: an imporoved gene syntehsis method for efficient production of error-corrected, synthetic protein libraries for directed evolution, Protein Engineering, Design & Selection 2014	PCR	functional colony output, no seq data
2015	APE solid phase gene synthesis	APE	0.78error/kb (with Phusion polymerase)

```
gat ctt aag gct aga gta cTA ATA CGA CTC ACT ATA GGG AGA CCA CAA CGG TTT CCC TCT AGA AAT AAT T  <70
            10          20          30          40          50          60

TT GTT TAA Ctt aag aag gag gaa aaa aaa atg gaa cgt ccg tac gct tgc ccg gtt gaa tct tgc gac cg  <140
                                    M   E   R   P   Y   A   C   P   V   E   S   C   D   R
            80          90          100         110         120         130

              finger1
t cgt ttc tct cgt tct gac gaa ctg acc cgt cat att aga att cat act gga caa aAA CCA TTC CAa tgt  <210
  R   F   S   R   S   D   E   L   T   R   H   I   R   I   H   T   G   Q   K   P   F   Q   C
            150         160         170         180         190         200

                                finger2
aga att tgt atg aga aat ttc tct cgt tct gac cac ctg acc acc cac atc cgt acc cac acc ggt gaa a  <280
R   I   C   M   R   N   F   S   R   S   D   H   L   T   T   H   I   R   T   H   T   G   E   K
            220         230         240         250         260         270

                                            finger3
aa ccg ttc gct tgc gac atc tgc ggt cgt aaa ttc gct cgt tct gac gaa cgt aaa cgt cac acc aaa at  <350
   P   F   A   C   D   I   C   G   R   K   F   A   R   S   D   E   R   K   R   H   T   K   I
            290         300         310         320         330         340

c cac ctg cgt cag aaa gac cca gcg cca gcg cca tct aaa ggt gaa gaa tta ttc act ggt gtt gtc cca  <420
  H   L   R   Q   K   D   P   A   P   A   P   S   K   G   E   E   L   F   T   G   V   V   P
            360         370         380         390         400         410

att ttg gtt gaa tta gat ggt gat gtt aat ggt cac aaa ttt tct gtc tcc ggt gaa ggt gaa ggt gat g  <490
I   L   V   E   L   D   G   D   V   N   G   H   K   F   S   V   S   G   E   G   E   G   D   A
            430         440         450         460         470         480

ct act tac ggt aaa ttg acc tta aaa ttt att tgt act act ggt aaa ttg cca gtt cca tgg cca acc tt  <560
   T   Y   G   K   L   T   L   K   F   I   C   T   T   G   K   L   P   V   P   W   P   T   L
            500         510         520         530         540         550

a gtc act act tta act tat ggt gtt caa tgt ttt tct aga tac cca gat cat atg aaa caa cat gac ttt  <630
  V   T   T   L   T   Y   G   V   Q   C   F   S   R   Y   P   D   H   M   K   Q   H   D   F
            570         580         590         600         610         620

ttc aag tct gcc atg cca gaa ggt tat gtt caa gaa aga act att ttt ttc aaa gat gac ggt aac tac a  <700
F   K   S   A   M   P   E   G   Y   V   Q   E   R   T   I   F   F   K   D   D   G   N   Y   K
            640         650         660         670         680         690

ag acc aga gct gaa gtc aag ttt gaa ggt gat acc tta gtt aat aga atc gaa tta aaa ggt att gat tt  <770
   T   R   A   E   V   K   F   E   G   D   T   L   V   N   R   I   E   L   K   G   I   D   F
            710         720         730         740         750         760

t aaa gaa gat ggt aac att tta ggt cac aaa ttg gaa tac aac tat aac tct cac aat gtt tac atc atg  <840
  K   E   D   G   N   I   L   G   H   K   L   E   Y   N   Y   N   S   H   N   V   Y   I   M
            780         790         800         810         820         830

gct gac aaa caa aag aat ggt atc aaa gtt aac ttc aaa att aga cac aac att gaa gat ggt tct gtt c  <910
A   D   K   Q   K   N   G   I   K   V   N   F   K   I   R   H   N   I   E   D   G   S   V   Q
            850         860         870         880         890         900

aa tta gct gac cat tat caa caa aat act cca att ggt gat ggt cca gtc ttg tta cca gac aac cat ta  <980
   L   A   D   H   Y   Q   Q   N   T   P   I   G   D   G   P   V   L   L   P   D   N   H   Y
            920         930         940         950         960         970

c tta tcc act caa tct gcc tta tcc aaa gat cca aac gaa aag aga gac cac atg gtc ttg tta gaa ttt  <1050
  L   S   T   Q   S   A   L   S   K   D   P   N   E   K   R   D   H   M   V   L   L   E   F
            990         1000        1010        1020        1030        1040

gtt act gct gct ggt att acC CAT GGT ATG GAT GAA TTG TAC AAA ggg ggt tct cat cat cat cat cat c  <1120
V   T   A   A   G   I   T   H   G   M   D   E   L   Y   K   G   G   S   H   H   H   H   H   H
            1060        1070        1080        1090        1100        1110

at TAA TAA CGA CTC AGG CTG CTA CCT AGC ATA ACC CCT TGG GGC CTC TAA ACG GGT CTT GAG GGG TTT TTT GA
   *   *    1130        1140        1150        1160        1170        1180
```

Features :

T7 promoter: [19 : 42], RBS: [86 : 92]

finger1: [151 : 171], finger2: [235 : 255], finger3: [319 : 339]

proline linker: [365 : 384], yEGFP: [385 : 1095], 6x Histidine tag: [1105 : 1122], T7 terminator: [1145 : 1192]

Figure 2.5.2 : Complete annotated DNA sequence of expression-ready linear template for WT Zif268. To create ZF variants, substitute in the coding sequence (21 nt, in blue) for residues -1 to 6 of recognition helix into the positions of finger 1, 2 and 3.

Table 2.5.5 : Oligomers for generation of linear expression template with GFP fusion

Oligomer Name	Sequence (5' to 3')
yEmCitrine-R1-f	CCTCTAGAAATAATTTTGTTTAACTTAAGAAGGAGGAAAAAAAAatgtctaaaggtgaagaattattcac
yEmCitrine-r	GTAGCAGCCTGAGTCGTTATTATTTGTACAATTCATCCATACCATGG
Proline-linker-EGFP-f	cacaccaaaatccacctgcgtcagaaagacccagcgccagcgccatctaaaggtgaagaattattcac
EGFP-His6-r	GTAGCAGCCTGAGTCGTTATTAatgatgatgatgatgatgagaaccccTTTGTACAATTCATCCATACCATGG
Genespecific-f	AGAAATAATTTTGTTTAACttaagaaggaggaaaaaaaa**atg**gaacgtccgtacgcttgcccggttgaatcttgcgaccgtcgtttctct
Extension-f	gatcttaaggctagagtacTAATACGACTCACTATAGGGAGACCACAACGGTTTCCCTCTAGAAATAATTTTGTTTAACttaagaagga
Extension-r	CAAAAAACCCCTCAAGACCCGTTTAGAGGCCCCAAGGGGTTATGCTAGGTAGCAGCCTGAGTCG
5'final_highTm	Cy3-gatcttaaggctagagtacTAATACGACTCACTATAGGG
3'final_highTm	CAAAAAACCCCTCAAGACCCGTTTAGAG

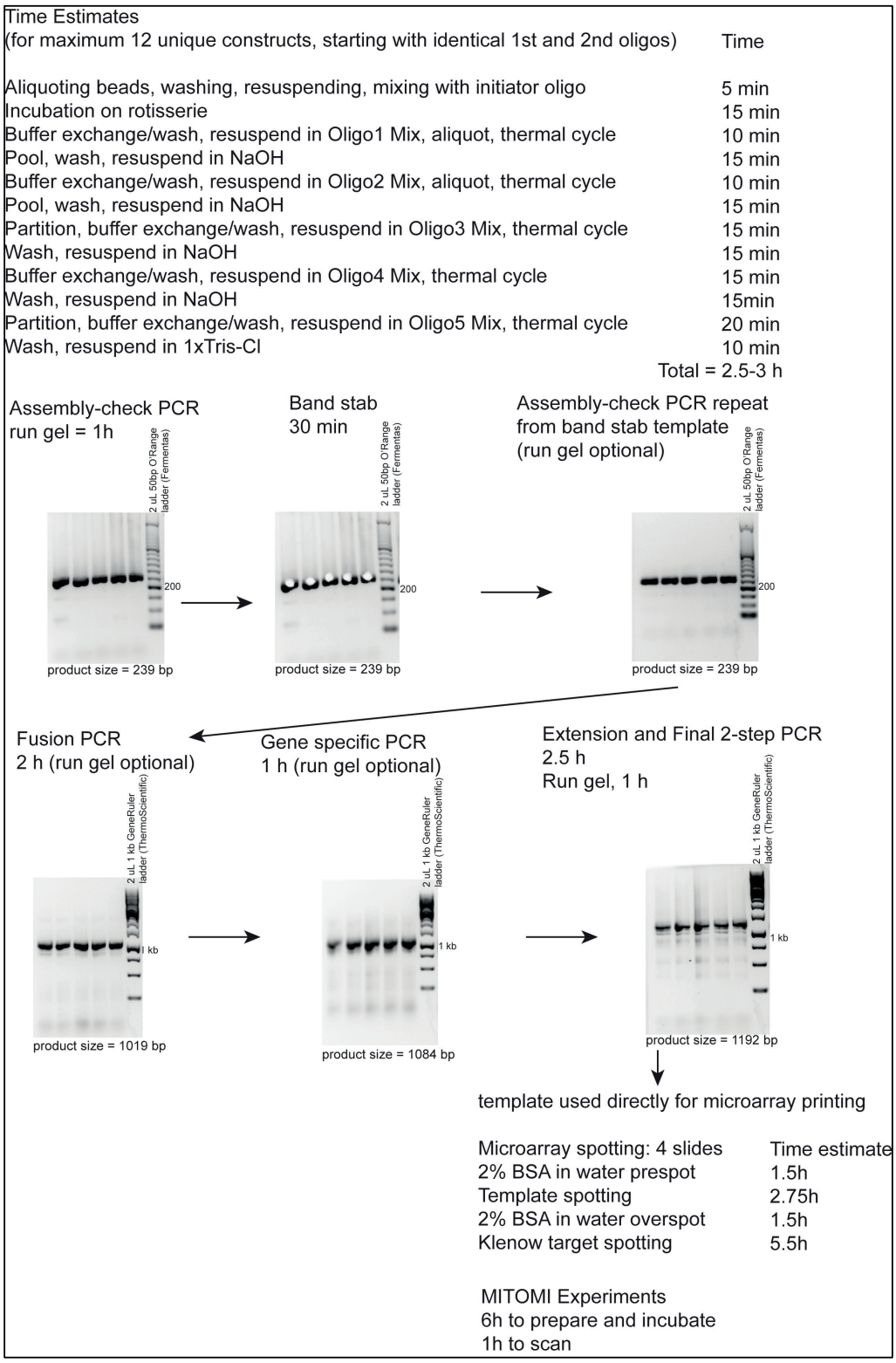


Figure 2.5.3 : Time estimates for performing up to 12 unique (no reaction pooling and splitting) APE assembly reactions by hand. Following the assembly reaction, several PCRs are performed to confirm full-length assembly, and to add on sequence necessary for on-chip expression and detection (5' and 3'UTRs, EGFP tag). Representative gels for the assembly of 5 templates during each step are shown. Finally, another list of time estimates for the microarraying process and running MITOMI experiments is given.

Chapter 3:
C_2H_2 ZF Modular Combinatorics

3.1 Introduction

Following the development of APE assembly for ZFA variants, we were interested in validating the versatility and speed of the platform by creating a library of ZFAs from the oligomer parts we already had available. Optimization of the APE assembly technique was performed with 4 ZFA variants: WT Zif268 and three synthetic proteins 37-12, 92-1 and 158-2 from (156). The amino acid residues of the recognition helices and the DNA target triplets for each ZF domain are given below in Figure 3.1.1.

Zif268 (wt)	**finger 1**	**finger 2**	**finger 3**	
	RSDELTR	RSDHLTT	RSDERKR	
	6 5 4 3 2 1 -1	6 5 4 3 2 1 -1	6 5 4 3 2 1 -1	
C	RKREDSR	TTLHDSR	RTLEDSR	N
5'-	**a GCG**	**TGG**	**GCG t**	

37-12	**finger 1**	**finger 2**	**finger 3**	
	RNFILQR	DRANLRR	RHDQLTR	
	6 5 4 3 2 1 -1	6 5 4 3 2 1 -1	6 5 4 3 2 1 -1	
C	RTLQDHR	RRLNARD	RQLIFNR	N
5'-	**t GAG**	**GAC**	**GTG t**	

92-1	**finger 1**	**finger 2**	**finger 3**	
	DSPTLRR	QRSSLVR	ERGNLTR	
	6 5 4 3 2 1 -1	6 5 4 3 2 1 -1	6 5 4 3 2 1 -1	
C	RTLNGRE	RVLSSRQ	RRLTPSD	N
5'-	**a GAT**	**GTA**	**GCC t**	

158-2	**finger 1**	**finger 2**	**finger 3**	
	DKTKLRV	VRHNLTR	QSTSLQR	
	6 5 4 3 2 1 -1	6 5 4 3 2 1 -1	6 5 4 3 2 1 -1	
C	RQLSTSQ	RTLNHRV	VRLKTKD	N
5'-	**t GTA**	**GAT**	**GGA g**	

Figure 3.1.1 : The four ZFAs and their amino acid residues for determining target specificity are listed. Each finger has a specific DNA target sequence. The residues are reversed in the lower box to indicate the location of the residues when they bind to DNA.

In early MITOMI experiments, we demonstrated that each of these ZFAs were functional (capable of binding their consensus target), but we had not characterized their specificity beyond the original set of four DNA targets. Each ZFA assembly required 3 unique 'finger' oligomers in addition to the 2 universal 'linking' oligomers, as explained in Chapter 2. The finger oligomers were designed such that only the central part of each oligomer uniquely encoded the recognition helix variants (residues -1 to 6). The flanking sequence at the termini of each finger oligomer remained constant, meaning all oligomers designed for a particular finger position could be interchanged. This meant that we could generate up to 64 ZFAs simply by shuffling the 12 finger oligomers that encoded the original set of 4 ZFAs.

APE assembly was applied towards the generation of the 64 ZFAs, and each assembly product was PCR amplified and extended to prepare EGFP-fusion, expression-ready linear templates (see Methods 2.4 for APE assembly and PCR steps). To provide a simple pattern for naming the 64 variants, each of the original ZFAs were labeled A through D (Zif268 = A, 37-12 = B, 92-1 = C, 158-2 = D). Variants made by combining ZF domains from these different ZFAs were named from C to N-terminus (in the direction of target binding). For example, a variant with Finger1 from 37-12, Finger2 from 92-1, and Finger3 from Zif268 would be named ACB (from C- to N-terminus). To characterize the binding affinity and specificity of each of these variants, a panel of target sequences were synthesized, where the expected target consensus for each variant was simply the concatenation of each finger's cognate DNA triplet. Targets were labeled in a similar fashion, from 5' to 3', such that the target had the same name as the protein expected to bind it (ie, protein ACB has the expected consensus target ACB). Each ZFA variant was tested in combination with the full panel of DNA targets to evaluate their specificity using a 1024 unit MITOMI microfluidic device. Details of device preparation and operation, in addition to fluorescent image analysis and affinity calculation are given in the Methods section 3.4.

3.2 Results

We synthesized 64 different ZFAs and quantified binding of each against a library of the corresponding 64 predicted consensus DNA targets. APE-MITOMI successfully

expressed all 64 ZFs (EGFP signal and levels of target DNA binding were observed for all variants). The filled line plot in Figure 3.2.1 was made by rank-ordering all of the ZFA variants by order of their highest observed binding affinity. A solid black line indicates the arbitrary cut-off for ZFAs considered 'non-functional'. These six ZFAs bound to targets very weakly relative to the others, and they were deemed non-functional because their expected consensus target was not found within the top 10 highest bound targets. As displayed in pie charts of Figure 3.2.1, 90.6% (58/64) of the ZFA variants bound DNA. Of those ZFs that bound DNA, 89.7% (52/58) bound the expected consensus target within the top 4 highest-affinity targets and 58.6% (34/58) bound the expected target with highest affinity.

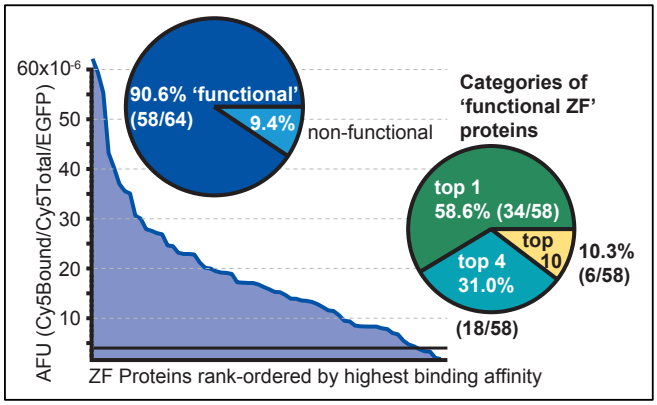


Figure 3.2.1 : Overview of experimental results obtained from combinatoric assembly of ZFAs demonstrating protein expression and functional DNA binding success rates.

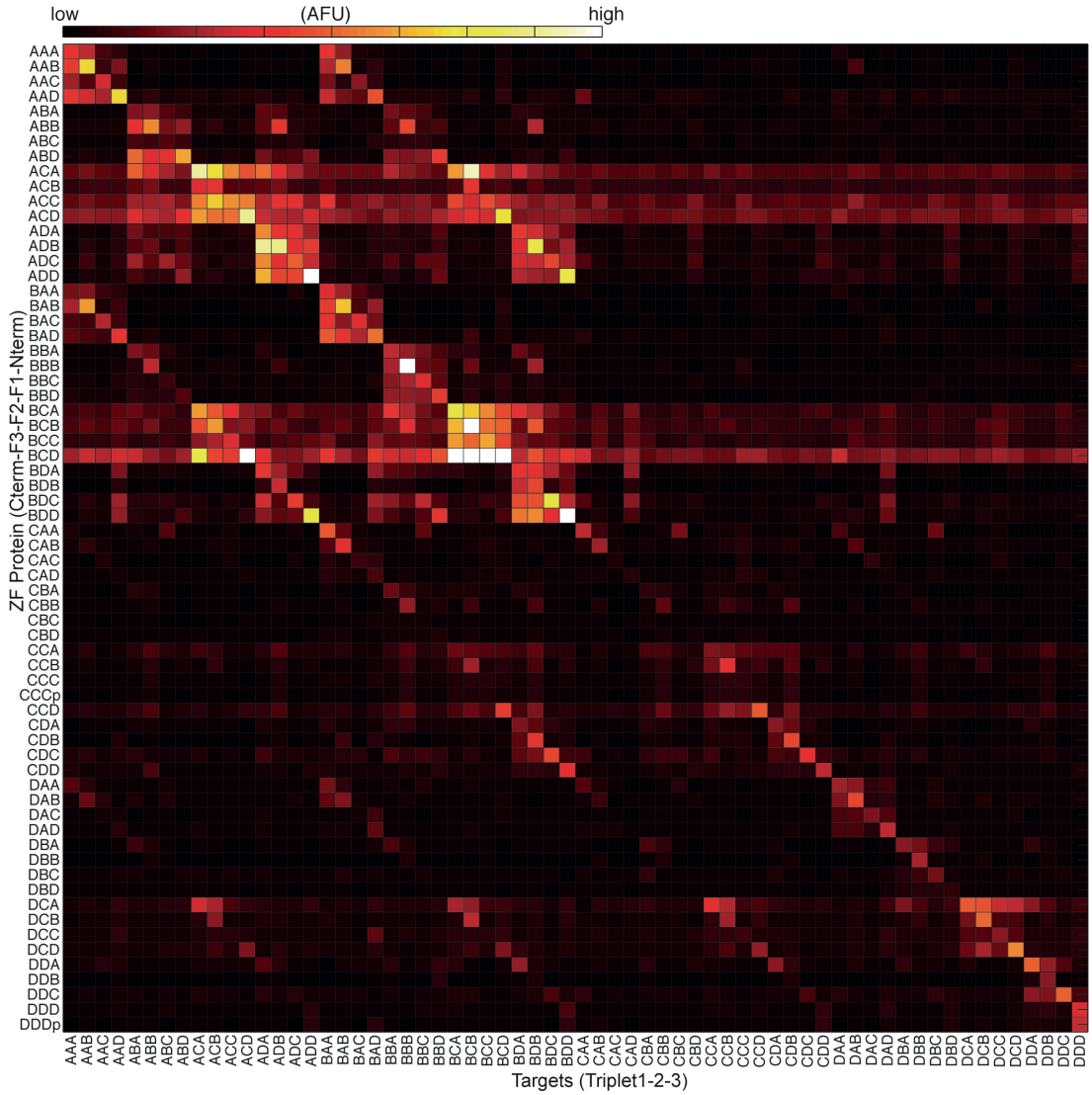


Figure 3.2.2 : Heatmap of dimensionless, relative binding affinities for each assembled ZFA variant (y-axis) to 64 predicted consensus DNA targets (x-axis). The protein naming convention indicates ZF domains from C-to-N (F3 to F1), where AAA ($A_{f3}A_{f2}A_{f1}$)=Zif268, BBB=37-12, CCC=92-1, DDD=158-2 (156); for example, protein ABC = F3 from zif268, F2 from 37-12, F1 from 92-1; target ABC = Zif268 F3 binding consensus triplet GCG, 37-12 F2 binding consensus GAC, 92-1 F1 binding consensus triplet GCC (5'-GCG GAC GCC). Oligomer assembly and target sequences are given in Supplementary Tables 3.5.1 and 2.

To verify that failure to bind and that sequence specificity was unaffected by our approach, we cloned and sequence verified the CCC and DDD ZFA variants generating the plasmid-based versions CCCp and DDDp (see Figure 3.2.2). Both plasmid-based ZFAs gave identical results when compared to non-sequenced versions. Due to the low APE assembly error-rate as well as the on-chip purification of only properly folded ZFA-EGFP fusions, all future experiments were performed without sequence verification of constructs. The heat map shows the quantitative binding specificity of each of the 64+2 ZFA variants. Except for 6 ZFA-DNA pairs out of 4096 potential combinations, all of the affinity data in

Figure 3.2.2 are the average of a least 2 datapoints from different experiments. On average, each interaction was evaluated 5 times in different MITOMI experiments. The heatmap displays the average affinity values obtained from over 24,000 data points.

3.3 Discussion

This approach of rapidly generating ZFAs and evaluating their target specificities can be used to identify orthogonal ZFAs for use in the design and implementation of synthetic genetic networks (156) and to eliminate ZFAs with sub-optimal binding characteristics such as low affinity or extensive non-specific binding. As seen in Figure 3.2.2, we observed that ZFAs containing a 'C' finger in position F2 exhibited high levels of non-specific binding, especially when combined with finger variants A and B in position F3 (see Figure 3.2.2).

Due to the cross-strand interaction of residue 2 in the recognition helix, and variations in the types of assays used to evaluate DNA specificity, the modularity of ZFAs has been repeatedly questioned (117, 271). Studies evaluating ZF nuclease cleavage frequently reported high failure rates for their modularly derived variants. ZF nuclease cleavage relies on the dimerization of two FokI nuclease domains brought into proximity by fusing them to site-specific ZFAs that target regions on either side of the desired cut site (144). DNA cleavage failure is a difficult situation to troubleshoot due to the number of variables that can lead to failure: non-specific or low-affinity binding of the ZFAs, inadequate linker length for fusion to the FokI domain, insufficient amounts of the proteins to cause cleavage, and complications with the reporter used to quantify cleaving efficiency. By assaying ZFAs purely on their ability to bind DNA, it is commonly seen that ZFAs constructed by modular combinations of characterized domains were functional (117, 272, 273). The modular combination results we present here corroborate with results observed in other *in vivo* and *in vitro* studies.

3.4 Methods

3.4.1 dsDNA Target Synthesis

Double stranded DNA targets (dsDNA) for zinc-finger array binding were prepared via isothermal Klenow extension as previously described (210). Oligomers were designed such that the DNA contains the 9 nt target sequence with single nucleotide flanks (11 nt total). At the 3'end of the DNA is the complementary sequence for a Cy5-labeled primer (5'CompCy5; Table 3.4.1). The reaction consists of 2 steps, one annealing step, and one extension step. Each 20 µL annealing reaction contains 1x NEB Buffer 2, 10 µM 5'CompCy5, and 15 µM Target oligomer. This mixture was placed on a thermocycler, heated to 94°C for 5 minutes, then cooled to 37°C (10% ramp) for 5 minutes, then held at 20°C. Following the annealing program, 10 µL of extension mix (1x NEB Buffer 2, 3mM dNTPs, and 2.5 units Klenow exo-) are spiked into the annealing mix, and the reaction is thermocycled according to the following routine: 37°C, 90min; 75°C heat kill, 20min; 10% ramp down to 30°C, 30s; hold at 4°C.

Table 3.4.1 : Klenow extension, Cy5-labelled target oligomer design

Oligomer Name	Sequence (5' to 3')
5'CompCy5	Cy5-GTCATACCGCCGGA
Target design	GGCCAATT X XXX XXX XXX X TTTCCGGCGGTATGAC

Where xxxx indicates location of target sequence (11 nt target = 9 nt binding site with flanking nt on each end)

3.4.2 Microarray printing

Preparation of epoxy-silane glass slides

For microarray printing, epoxy-silane functionalized glass slides were prepared, following an adapted protocol (274). A MilliQ water and ammonia solution (NH_4OH, 25%) mixture was prepared in a 5:1 ratio, respectively, and heated to 80°C. Then, 150 mL of hydrogen peroxide (H_2O_2, 30%) was added to the mixture, and glass microscope slides were placed in the cleaning bath for 30 minutes. The glass slides were then removed and rinsed in fresh MilliQ water, dried with N_2, and placed in a second bath containing 1% 3-glycidoxypropyltrimethoxymethylsilane (97% purity) in toluene, and incubated for at least 20 minutes at ambient temperature. Then, the glass slides were removed, rinsed in fresh

toluene, dried with N_2, and baked at 80°C for 30 minutes. The glass slides were removed from the oven, allowed to cool, and stored under vacuum at room temperature in opaque storage boxes until used. Immediately prior to microarray printing, both sides of each glass slide is briefly rinsed with fresh isopropanol, dried with N_2, rinsed with fresh toluene, and dried with N_2 again.

Sample microarraying

All samples to be printed onto a microarray were prepared in a 384-well microtiter plate. Each zinc-finger array assembly (1192 bp PCR product) was co-spotted with a target DNA (Klenow extended products) in duplicate onto epoxy-silane coated glass slides using a microarray robot (QArray2) with a 946MP4 microspotting pin (Arrayit). Up to four slides were prepared in a single printing session. Each spot on the array was generated by four consecutive printing programs. Immediately following completion of printing from a given sample plate, the sample wells were covered with adhesive PCR foil seals (ThermoScientific) and stored at -20°C until needed for future printing runs. In general, Klenow target plates were re-used for up to 5 printing runs before the volume of each well became too low to be used. After the final printing program, the microspotting pin is cleaned by sonication for 15min in a 15 mL Falcon tube containing one drop of dish detergent and 10 mL of deionized water, then sonicated for 30min in a 15 mL Falcon tube containing 10 mL of 70% ethanol in water. Before use, the pin tip is rinsed under deionized water, dried using a high pressure (100 psi) compressed air gun, then inspected under a microscope to verify the tip is clean and free of debris. Finally, the shaft of the pin is briefly polished with a dry Kimwipe to prevent sticking when placed into the robotic printer head.

Microarray Printing Routine

The first BSA printing routine is to block the surface of the glass, to limit the amount of template and target DNA that is irreversibly attached to the epoxy-silane surface. The second BSA printing routine is implemented to deposit a thin blocking layer on top of the printed template spots, to reduce the risk of cross-contamination due to template carry-over during target printing (since target printing does not involve same-sample washing

steps). The full printing routine and operational details can be found in Table 3.4.2 and Figure 2.5.3.

Table 3.4.2 : Microarray robotic printing routine

Step	Printing reagent	Stamps /spot	Stamps /ink	Humidity	Wash frequency
1 (prespot)	0.5% BSA in water	1	4	45-55%	No same-sample wash, 1 source well per linear template
2	Linear template printing (6 µL final PCR product + 64 µL 2% BSA in water)	2	4	45-55%	Wash after 32 inks
3 (overspot)	0.5% BSA in water	1	4	45-55%	No same-sample wash, 1 source well per linear template
4	Cy5-DNA target printing (10 µL Klenow product + 60 µL 2% BSA in water)	3	3	45-55%	Wash after 32 inks

3.4.3 MITOMI chip fabrication and operation

Mold Fabrication

The MITOMI microfluidic device (204, 275) (Figure 1.4.1-3) consists of two superimposed layers, the flow layer and the control layer. Each layer is fabricated in polydimethylsiloxane (PDMS) using 4" silicon wafer molds fabricated using standard lithography techniques(206). Each wafer (control and flow) contained three pattern replicates for a 1024-chamber (16 rows by 64 columns) MITOMI device.

Mask fabrication was carried out using a Heidelberg DWL200 laser lithography system with 10mm writing head and solid state wavelength stabilized laser diode (max. 110 mW at 405 nm). Each layer of the MITOMI device was reproduced as a chrome mask. After laser writing, the chrome mask is cycled twice for 15s in developer mixture (1:5 MP351 and deionized water, respectively), 45s of agitation, then rinsed and dried. The developed mask is then chrome etched for 110s, rinsed, cleaned twice for 15min in 1165-remover bath, rinsed and air dried.

The flow layer mold is first cleaned for 7 min in a Tepla300 plasma stripper with 400 mL/min O2 at 500W and 2.45 GHz. The wafer is then treated with hexamethyldisilazane (HMDS) using an ATMsse hotplate at 125°C for 12min. Positive photoresist AZ9260 is spin-coated on the cleaned wafer for 10s at 800 rpm, then 40s at 1800 rpm (ramp 1000 rpm/s) to produce a substrate height of 14 µm. The wafer is then baked on a 115°C

hotplate for 6min. The soft-baked positive resist is then allowed to rehydrate for 1h. The wafer is then exposed during 3 intervals of 18s with a 10s pauses between each exposure on a MA6 mask aligner (power 360 mJ/cm^2, intensity 10 mW/cm^2, broad-spectrum lamp, hard contact exposure mode). After a 1h relaxation time, the wafer is developed in a DV10 chamber via multiple, automated cycles of rinsing/agitation with development mixture (1:4 ratio of AZ400K and deionized water, respectively) until the features are visible. Finally, the wafer is heated to 160°C for 20 minutes to anneal and round the features of the flow wafer to create a profile that allows complete valve closure.

The control layer mold is first cleaned following the same plasma treatment protocol as the flow layer mold. Negative photoresist SU-8 GM1060 (Gestertec) is spin-coated on the cleaned wafer for 10s at 500 rpm (ramp 100 rpm/s), 10s at 1500 rpm (ramp 100 rpm/s), 1s at 2500 rpm, and finally 6s at 1500 rpm to produce a substrate height of 14 µm. The wafer is baked on a hotplate for 30min at 130°C, then 25min at 30°C. The wafer is then exposed on a MA6 mask aligner for 13.2s (power 360 mJ/cm^2, intensity 10mW/cm^2, broad-spectrum lamp, hard contact exposure mode). The exposed wafer is developed manually by bathing in PGMEA twice for 1.5min, then rinsed in isopropanol and dried with an air gun.

Device Fabrication

Prior to PDMS casting, both the flow and control layer wafers are subjected to vapor deposition of trimethylchlorosilane (TMCS, EMD Millipore Corp.) for at least 30min by placing them within a sealed plastic container with a small dish containing 0.25 mL liquid TMCS. TMCS treatment is repeated for at least 15min before all subsequent PDMS casting rounds. The control layer wafer is placed into an aluminum foil-lined glass Petri dish, and 60g of Sylgard elastomer (5:1 mix of elastomer base and curing agent, respectively) is mixed for 1min at 2000rpm (400×g) and degassed for 2min at 2200 rpm (440×g) in a centrifugal mixer. The elastomer mixture is poured on top of the control layer in the Petri dish, and degassed in a vacuum dessicator for 20min at ambient temperature.

For the flow layer, 21 g of PDMS mixture is prepared at the ratio of 20:1 (base:curing agent), then mixed and degassed in a centrifugal mixer according to the same speeds and times as the control layer. The flow wafer is carefully centered on top of a spin-coater platform using wafer tweezers, and the flow layer PDMS mixture is poured in

the center, taking care not to create any bubbles. The mixture is spin-coated onto the wafer with a 15s ramp and 35s spin at 2800 rpm. The degassed PDMS on the control layer wafer is removed from the vacuum chamber. Residual bubbles are removed with a scalpel and any pieces of dust are carefully removed from the control channel grid using the tip of the scalpel blade.

Both the control and flow layers are then placed into an oven at 80°C for 28-30min. After baking, both Petri dishes are removed from the oven and briefly allowed to cool. The control layer is then cut with a scalpel in a rectangle around each pattern replicate, and each rectangle of cured PDMS is carefully peeled away from the silicon wafer. Holes are punched through each of the control line input channels on the patterned side of the PDMS block. The patterned side of the control layer is cleaned twice with Scotch Magic Tape to remove dust and debris then quickly placed on top of the flow layer replicates. A stereomicroscope is used to precisely align the features of the control layer so that they overlap with the chambers visible on the flow layer. Once aligned, the assembled device is bonded at 80°C for 90-180min. The bonded devices are removed from the oven and briefly allowed to cool. A scalpel is guided around the outer edge of the control layer PDMS block to cut the thin flow layer. Then each individual device is gently peeled from the flow layer wafer, and holes are punched through the patterned side inlets and outlet of the flow layer. Each device is then cleaned with Magic Scotch tape and trimmed to fit within the boundary defined by the glass slide-holding cartridge of the microarray scanner. The assembled device is aligned with a printed microarray on an epoxy-silane glass slide using a stereomicroscope and bonded overnight at 80°C before use.

The flow layer mold is cleaned of residual polymerized PDMS by pouring on another layer of mixed PDMS (this can be leftover control- or flow-layer PDMS mixtures prepared earlier; to stall cross-linking for several hours, store the PDMS mixture at 4°C), and baked at 80°C for at least 1 hour. The resulting thicker layer of PDMS can be easily peeled away from the flow layer, resulting in a clean surface to repeat the process. Both the cleaned flow wafer and control wafer are cleaned with high pressure (100 psi) compressed air gun to dislodge pieces of dust or PDMS before being treated with TMCS.

Device Setup

Assembled MITOMI chips bonded to microarray printed glass slides were stored at 40°C following an overnight bonding at 80°C until used. To begin an experiment, control line tubing is filled with PBS using a syringe, and pins are placed into the appropriate locations to feed into the control valves of the microfluidic device. The control lines are actuated at low pressure (10 psi) to begin filling the control lines of the microfluidic device. Once all of the control lines are filled, the sandwich valves and button valves are deactivated, and the pressure is increased to 20-22 psi to ensure complete closure of all other valves.

Surface Derivatization, Protein Synthesis, Binding Assay, and Device Readout

Biotin-BSA (2 mg/mL) is flowed through the device for 15min at 3.5 psi. The chip is then washed with 0.01% Tween20 in PBS for 5min to wash away unbound biotin-BSA. Next, neutravidin (1 mg/mL) is flowed for 15min followed by 0.01% Tween20 in PBS for 5min. The button valves are then activated and biotin-BSA is again flowed across the chip for 10min, blocking all of the neutravidin binding sites except those protected under the area of the button valve. The chip is again washed with 0.01% Tween20 in PBS for 5min. Then a solution containing 0.5 µL biotinylated antibody to GFP (1 mg/mL stock, Abcam ab6658) in 100 µL 1% BSA in PBS is flowed across the chip for 5min, the button valve is deactivated, and the antibody solution is flown for 15min, allowing the antibody to bind to the available neutravidin under the button valve. Then, the chip is flushed with 0.01% Tween20 in PBS for 5min, and with PBS for 5min. The button valve is then activated, and ITT mixture (NEB PURExpress, 10 µL SolnA, 7.5 µL SolnB, 0.5 µL RNAse Inhibitor (Roche), 7 µL PCR grade water) is flowed for 10min. The exit valve is activated for 2min while the ITT is being flowed on-chip to build up pressure. The neck valve is deactivated, and the ITT mixture is allowed to fill the DNA chambers for 1-2min.

Once the DNA chambers are filled, the neck valves are activated, the exit is opened, and fresh ITT is allowed to flow across the chip for 10min. Then the sandwich valve is activated while flowing ITT mix during the last minute of ITT washing. Once the sandwich valves are partially closed, the button valve is deactivated, the neck valve is deactivated, the exit is closed, and the flow of ITT is stopped. The inlet tree valve

controlling entry to the chamber array is closed, and the inlet tree is briefly flushed with 0.01% Tween20 in PBS. Then the entire chip is placed on a flatbed thermal cycler set to 37°C, and incubated for 3-5h. During this time, the DNA array spots are rehydrated in the ITT mix, transcription and translation occur, synthesized zinc-finger/EGFP fusion protein diffuses and is bound by the anti-GFP moiety located under the button valve, and target DNA diffuses and interacts with the various zinc finger DNA binding domains. After incubation, the chip is placed into an ArrayWoRx microarray scanner, and an image is taken in three fluorescent channels (A488/GFP, Cy3, Cy5) to determine relative amounts of solution phase target DNA in the MITOMI chamber. The button valves are then activated, the sandwich valves are deactivated, and the neck valve is activated again. The flow space is washed with 0.01% Tween20 in PBS for 5min to remove unbound target DNA, then the chip is scanned again in the three fluorescent channels, giving the total protein signal (EGFP/A488 signal) and the relative amount of target bound (Cy5 signal). Each zinc finger fusion template was tagged with Cy3 during the final PCR step, and though signal from this channel was captured in each scan, it was not factored into downstream binding-specificity analyses, primarily because little Cy3 signal was detected as being bound by the ZF proteins and normalization for protein amount was performed with the EGFP fluorescence.

3.4.4 Image Analysis and Affinity Value Calculations

Images acquired from the experiment were processed using a 1024 unit detection array in GenePix v6.0. Raw tif files from the ArrayWoRx scanner were loaded into the GenePix software, and using the grid detection tool, an array of 1024 circular areas was snapped onto the EGFP spots detected in the A488 channel after washing. Small adjustments to the grid were performed by hand for poorly detected locations. Mean and median fluorescence measurements were taken in each fluorescent channel before and after washing. In addition, local background measurements were taken by dragging the detection array off the button valve locations into the space just outside the reaction chamber. For each scan, each circular area in the array was background corrected using its own local background measurement. Datapoints were filtered to ensure that EGFP levels were at least 500 AFU (arbitrary fluorescence units) or higher, and Cy5 target levels

at least 1000 AFU or higher. These filtered, and background-corrected data points were used to calculate 'relative affinity' values (reported in AFU), using Equation 3.4.1 below:

$$\textit{relative affinity} = \frac{\dfrac{\text{Cy5}_{\text{bound}}\ \text{(after wash)}}{\text{Cy5}_{\text{total}}\ \text{(before wash)}}}{\text{A488 (EGFP after wash)}} \qquad \text{(Equation 3.4.1)}$$

These relative affinity values were used to compare ZFA binding affinity to various targets across different experiments. In general, at least 2 data points were averaged together to arrive at a single affinity value (as in the heat map generated in Figure 3.2.2).

3.5 Supplementary Tables

Table 3.5.1 : DNA target sequences for ZF combinatorics (Figure)

Oligomer Name	Sequence (5' to 3')
target_BAA	GGC CAA TTT GAG TGG GCG TTT TCC GGC GGT ATG AC
target_CAA	GGC CAA TTA GAT TGG GCG TTT TCC GGC GGT ATG AC
target_DAA	GGC CAA TTT GTA TGG GCG TTT TCC GGC GGT ATG AC
target_ABA	GGC CAA TTA GCG GAC GCG TTT TCC GGC GGT ATG AC
target_ACA	GGC CAA TTA GCG GTA GCG TTT TCC GGC GGT ATG AC
target_ADA	GGC CAA TTA GCG GAT GCG TTT TCC GGC GGT ATG AC
target_AAB	GGC CAA TTA GCG TGG GTG TTT TCC GGC GGT ATG AC
target_AAC	GGC CAA TTA GCG TGG GCC TTT TCC GGC GGT ATG AC
target_AAD	GGC CAA TTA GCG TGG GGA GTT TCC GGC GGT ATG AC
target_ABB	GGC CAA TTA GCG GAC GTG TTT TCC GGC GGT ATG AC
target_CBB	GGC CAA TTA GAT GAC GTG TTT TCC GGC GGT ATG AC
target_DBB	GGC CAA TTT GTA GAC GTG TTT TCC GGC GGT ATG AC
target_BAB	GGC CAA TTT GAG TGG GTG TTT TCC GGC GGT ATG AC
target_BCB	GGC CAA TTT GAG GTA GTG TTT TCC GGC GGT ATG AC
target_BDB	GGC CAA TTT GAG GAT GTG TTT TCC GGC GGT ATG AC
target_BBA	GGC CAA TTT GAG GAC GCG TTT TCC GGC GGT ATG AC
target_BBC	GGC CAA TTT GAG GAC GCC TTT TCC GGC GGT ATG AC
target_BBD	GGC CAA TTT GAG GAC GGA GTT TCC GGC GGT ATG AC
target_ACC	GGC CAA TTA GCG GTA GCC TTT TCC GGC GGT ATG AC
target_BCC	GGC CAA TTT GAG GTA GCC TTT TCC GGC GGT ATG AC
target_DCC	GGC CAA TTT GTA GTA GCC TTT TCC GGC GGT ATG AC
target_CAC	GGC CAA TTA GAT TGG GCC TTT TCC GGC GGT ATG AC
target_CBC	GGC CAA TTA GAT GAC GCC TTT TCC GGC GGT ATG AC
target_CDC	GGC CAA TTA GAT GAT GCC TTT TCC GGC GGT ATG AC
target_CCA	GGC CAA TTA GAT GTA GCG TTT TCC GGC GGT ATG AC
target_CCB	GGC CAA TTA GAT GTA GTG TTT TCC GGC GGT ATG AC
target_CCD	GGC CAA TTA GAT GTA GGA GTT TCC GGC GGT ATG AC
target_ADD	GGC CAA TTA GCG GAT GGA GTT TCC GGC GGT ATG AC
target_BDD	GGC CAA TTT GAG GAT GGA GTT TCC GGC GGT ATG AC
target_CDD	GGC CAA TTA GAT GAT GGA GTT TCC GGC GGT ATG AC
target_DAD	GGC CAA TTT GTA TGG GGA GTT TCC GGC GGT ATG AC
target_DBD	GGC CAA TTT GTA GAC GGA GTT TCC GGC GGT ATG AC
target_DCD	GGC CAA TTT GTA GTA GGA GTT TCC GGC GGT ATG AC
target_DDA	GGC CAA TTT GTA GAT GCG TTT TCC GGC GGT ATG AC
target_DDB	GGC CAA TTT GTA GAT GTG TTT TCC GGC GGT ATG AC
target_DDC	GGC CAA TTT GTA GAT GCC TTT TCC GGC GGT ATG AC
target_BCA	GGC CAA TTT GAG GTA GCG TTT TCC GGC GGT ATG AC
target_BDA	GGC CAA TTT GAG GAT GCG TTT TCC GGC GGT ATG AC
target_CBA	GGC CAA TTA GAT GAC GCG TTT TCC GGC GGT ATG AC
target_CDA	GGC CAA TTA GAT GAT GCG TTT TCC GGC GGT ATG AC
target_DBA	GGC CAA TTT GTA GAC GCG TTT TCC GGC GGT ATG AC
target_DCA	GGC CAA TTT GTA GTA GCG TTT TCC GGC GGT ATG AC

target_ACB	GGC CAA TTA GCG GTA GTG TTT TCC GGC GGT ATG AC
target_ADB	GGC CAA TTA GCG GAT GTG TTT TCC GGC GGT ATG AC
target_CAB	GGC CAA TTA GAT TGG GTG TTT TCC GGC GGT ATG AC
target_CDB	GGC CAA TTA GAT GAT GTG TTT TCC GGC GGT ATG AC
target_DAB	GGC CAA TTT GTA TGG GTG TTT TCC GGC GGT ATG AC
target_DCB	GGC CAA TTT GTA GTA GTG TTT TCC GGC GGT ATG AC
target_ABC	GGC CAA TTA GCG GAC GCC TTT TCC GGC GGT ATG AC
target_ADC	GGC CAA TTA GCG GAT GCC TTT TCC GGC GGT ATG AC
target_BAC	GGC CAA TTT GAG TGG GCC TTT TCC GGC GGT ATG AC
target_BDC	GGC CAA TTT GAG GAT GCC TTT TCC GGC GGT ATG AC
target_DAC	GGC CAA TTT GTA TGG GCC TTT TCC GGC GGT ATG AC
target_DBC	GGC CAA TTT GTA GAC GCC TTT TCC GGC GGT ATG AC
target_ABD	GGC CAA TTA GCG GAC GGA GTT TCC GGC GGT ATG AC
target_ACD	GGC CAA TTA GCG GTA GGA GTT TCC GGC GGT ATG AC
target_BAD	GGC CAA TTT GAG TGG GGA GTT TCC GGC GGT ATG AC
target_BCD	GGC CAA TTT GAG GTA GGA GTT TCC GGC GGT ATG AC
target_CAD	GGC CAA TTA GAT TGG GGA GTT TCC GGC GGT ATG AC
target_CBD	GGC CAA TTA GAT GAC GGA GTT TCC GGC GGT ATG AC
target_268wt	GGC CAA TTA GCG TGG GCG TTT TCC GGC GGT ATG AC
target_37-12	GGC CAA TTT GAG GAC GTG TTT TCC GGC GGT ATG AC
target_92-1	GGC CAA TTA GAT GTA GCC TTT TCC GGC GGT ATG AC
target_158-2	GGC CAA TTT GTA GAT GGA GTT TCC GGC GGT ATG AC

Table 3.5.2 : APE assembly oligomers for ZF combinatorics

Oligomer Name	**Sequence (5' to 3')**
Zif268_Oligo1fin3	**ttgcgacatctgcggtcgtaaattc**gct**cgttctgacgaacgtaaacgtcacaccaaaatccacctgcgtcaga**
Zif268_Oligo3fin2	**CCAatgtagaatttgtatgagaaatttc**tct**cgttctgaccacctgaccacccacatccgtacccacaccggtgaaa**
Zif268_Oligo5fin1	**tgaatcttgcgaccgtcgtttctctcgttctgacgaactgacccgtcatattagaattcatactggacaaa**
37-12Oligo1fin3	ttgcgacatctgcggtcgtaaattcgct**CGTCACGACCAGCTGACCCGT**cacaccaaaatccacctgcgtcaga
37-12Oligo3fin2	CCAatgtagaatttgtatgagaaatttctct**GACCGTGCTAACCTGCGTCGT**cacatccgtacccacaccggtgaaa
37-12Oligo5fin1	tgaatcttgcgaccgtcgtttctct**CGTAACTTCATCCTGCAGCGT**catattagaattcatactggacaaa
92-1Oligo1fin3	ttgcgacatctgcggtcgtaaattcgct**GAACGTGGTAACCTGACCCGT**cacaccaaaatccacctgcgtcaga
92-1Oligo3fin2	CCAatgtagaatttgtatgagaaatttctct**CAGCGTTCTTCTCTGGTTCGT**cacatccgtacccacaccggtgaaa
92-1Oligo5fin1	tgaatcttgcgaccgtcgtttctct**GACTCTCCGACCCTGCGTCGT**catattagaattcatactggacaaa
158-2Oligo1fin3	ttgcgacatctgcggtcgtaaattcgct**CAGTCTACCTCTCTGCAGCGT**cacaccaaaatccacctgcgtcaga
158-2Oligo3fin2	CCAatgtagaatttgtatgagaaatttctct**GTTCGTCACAACCTGACCCGT**cacatccgtacccacaccggtgaaa
158-2Oligo5fin1	tgaatcttgcgaccgtcgtttctct**GACAAAACCAAACTGCGTGTT**catattagaattcatactggacaaa

Chapter 4: C_2H_2 ZF DNA Specificity Engineering

4.1 Introduction

The ability to recombine ZF domains with distinct DNA specificities into linked ZFAs that target longer, unique sequences is a powerful tool for biotechnology and synthetic biology. Following the successful generation and characterization of the 64 ZFA variants in Chapter 3, we next applied APE-MITOMI to engineering ZF specificity. We chose two different target sites, and starting from a ZFA with a different consensus target, sequentially introduced finger variants at different positions and evaluated their binding affinity towards a panel of closely related targets.

4.2 Results

4.2.1 ZFA Specificity Engineering for 'GTA GAT GGC'

We first decided to engineer a ZFA with a consensus sequence of 'GTA GAT GGC', taking advantage of the relatively well-populated selection of GNN-binding recognition helices (RHs) in the Zinc Finger Consortium database (ZF DB). Starting with the WT Zif268 protein, we first modified F2 of Zif268 by replacing it with 16 RHs listed as 'GAT' binders (Figure 4.2.1). Characterizing these variants showed that there was considerable variability in the specificity and binding affinity of the 16 RHs tested. Due to observations that the aspartic acid residue at position 2 of each α-helix of Zif268 is important for determining sequence specificity (139), causing specificity constraints, the same set of 16 RHs were also used in a different F1/F3 context (37-12 framework, Figure 4.2.1). Additionally, by combining residues from RHs with high affinity and specificity towards

'GAT', a set of designed variants not found in the ZF DB were made, and these variants were placed into still another F1/F3 context (158-2 framework, Figure 4.2.1).

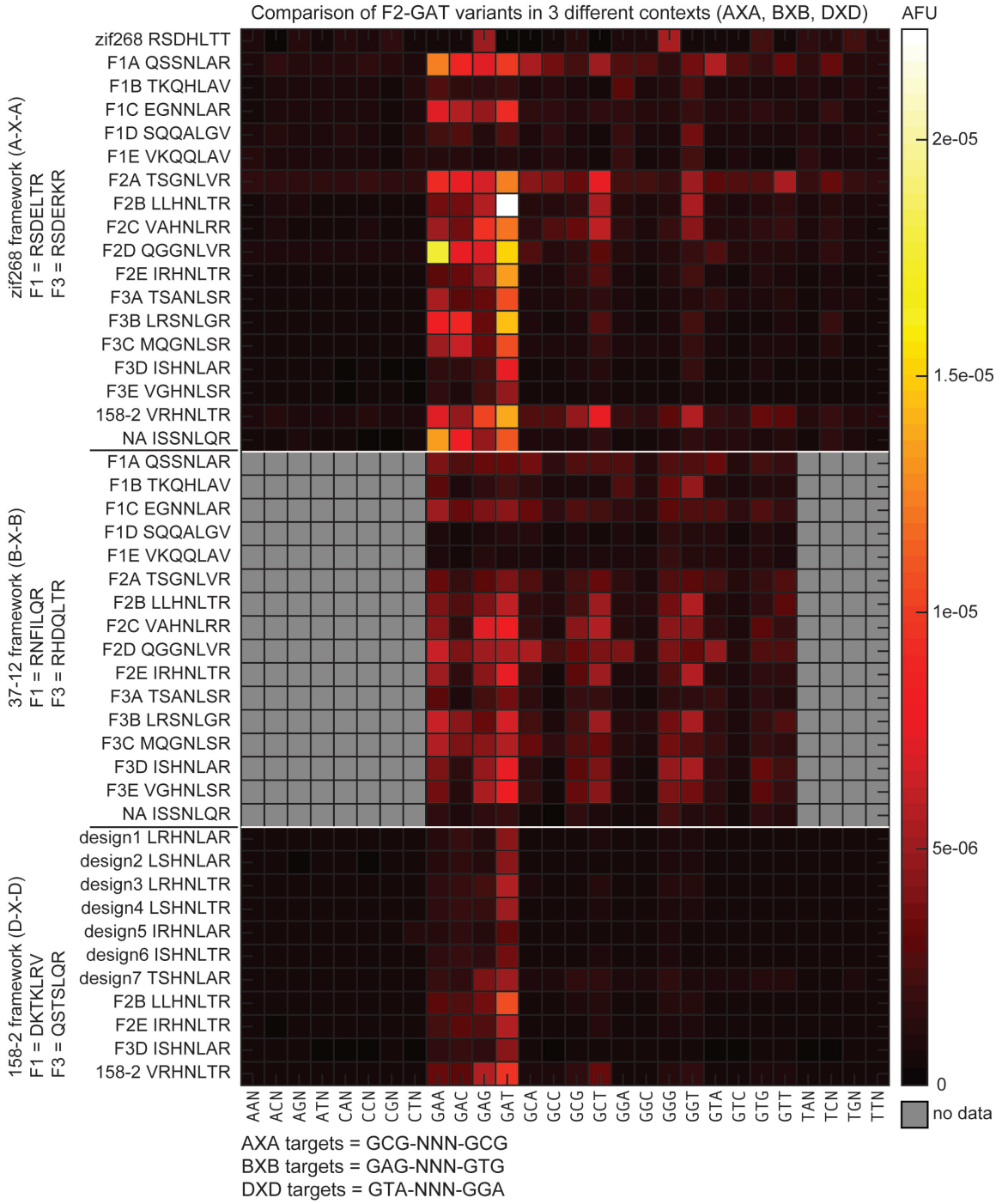


Figure 4.2.1 : Heat map of DNA target affinity data from all ZFAs containing F2 variants that were selected to bind the triplet GAT in different contexts (different F1/F3 combinations). The topmost 18 RH's placed in the zif268 F1/F3 context exhibited the highest affinities for the GAT target, whereas the same set placed within the 37-12 F1/F3 context exhibited weakened affinities with no clear affinity for GAT. As a final screen, the highest affinity variants from the zif268 screen were placed into the 158-2 F1/F3 context, in addition to seven 'designed' RHs based on residue combinations from the highest affinity/lowest non-specific variants (F2B, F2E and F3D). As observed in the zif268 context, the RH F2B (LLHNLTR) had the highest affinity, and was used as the F2 variant in subsequent screens for the other finger positions.

One RH variant, LLHNLTR, consistently exhibited the highest affinity and decent specificity for ‘GAT’ in different F1/F3 frameworks. This RH also displayed low levels of affinity towards other ‘GAN’ triplets, but considering its elevated affinity for our desired target, we deemed its specificity satisfactory. We thus used this helix for the next selection step, which involved substituting RH variants into the F1 position to target ‘GGC’. Initially, we tested the F1 variants using F3 from Zif268, and the new F2 variant from the ‘GAT’ selection round. These protein variants exhibited high levels of degeneracy to the extent that it was not clear which, if any, of the F1 RH variants were specific to ‘GGC’ (Figure 4.2.2). In an effort to improve the target specificity, we performed a second screen using the same set of F1 variants for ‘GGC’, but with F3 from ZFA 158-2, along with the selected F2 variant from the ‘GAT’ selection (Figure 4.2.3).

The variants in Figure 4.2.3 displayed a similar specificity profile to those observed in Figure 4.2.2, but with reduced the non-specific background. Nearly all of the RH variants, which were annotated in the ZF DB as binding ‘GGC’, appeared to have higher affinity for the closely related target ‘GTC’. To advance the engineering process, we selected ESSKLKR as the best ‘GGC’ binder in position F1, considering its level of specificity for ‘GGC’ relative to ‘GTC’. We then performed a third and final screen by substituting RH variants which were reported to bind ‘GTA’ into F3 position (Figure 4.2.4). The RH that displayed the highest ‘GTA’ binding affinity and specificity, QSSALTR, did not exist in the ZF DB. It was generated by selecting the highest frequency residue in each position from all of the ‘GTA’ RH variants listed in the database.

In order to characterize the actual specificity of our synthetic ZFA (F1 ESSKLKR, F2 LLHNLTR, F3 QSSALTR), a 1-off target library was prepared for the consensus sequences: ‘GTA GAT GGC’ and ‘GTA GAT GTC’ (Figure 4.2.5). Given the relatively poor performance observed during the F1 (‘GGC’ variants) selection step, the engineered ZFA exhibited a surprisingly high affinity and specificity for the intended target consensus sequence (‘GTA GAT GGC’). In an attempt to improve the specificity of our engineered variant for the intended target, we performed an additional screen using 14 designed RH variants which were anticipated to bind ‘GGC’ based on online DNA binding site predictors (75, 76, 145, 276) (Figure 4.5.1), but nearly all these variants displayed a higher affinity for GTC and thus failed to further improve the specificity of the engineered ZFA (Figure 4.2.5).

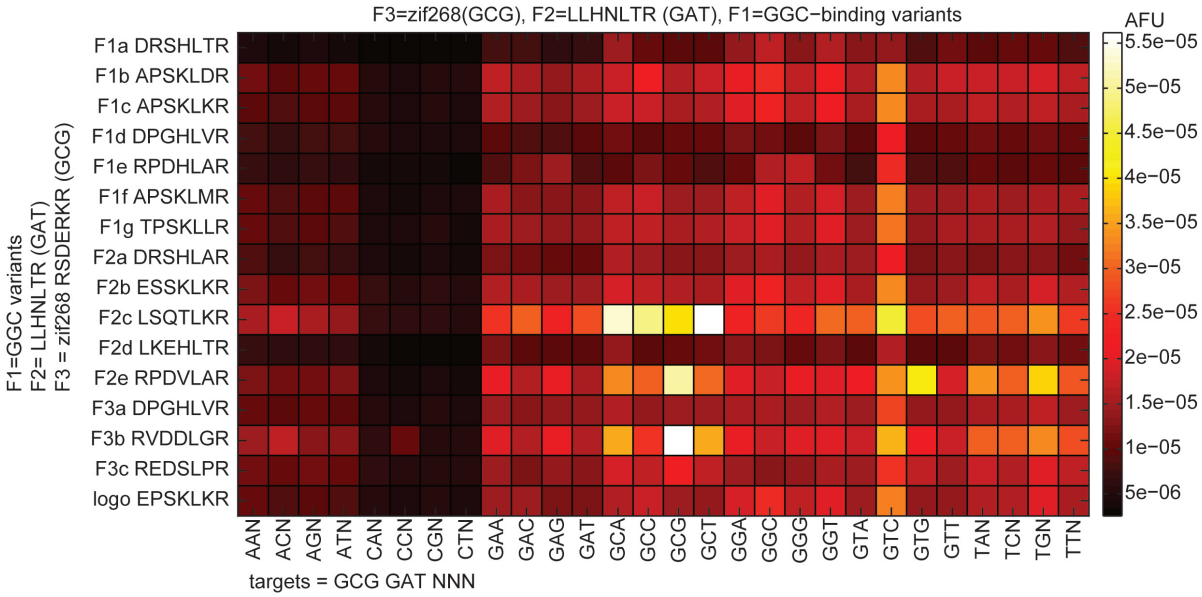


Figure 4.2.2 : Heat map of affinity data from F1 RH variants selected to bind GGC with F3 from zif268 and F2-LLHNLTR from the GAT selection screen. Due to nonspecific binding for nearly all GNN targets, there is no clear RH with high specificity for GGC, and so a second screen was performed with a different F3 (Figure 4.2.3).

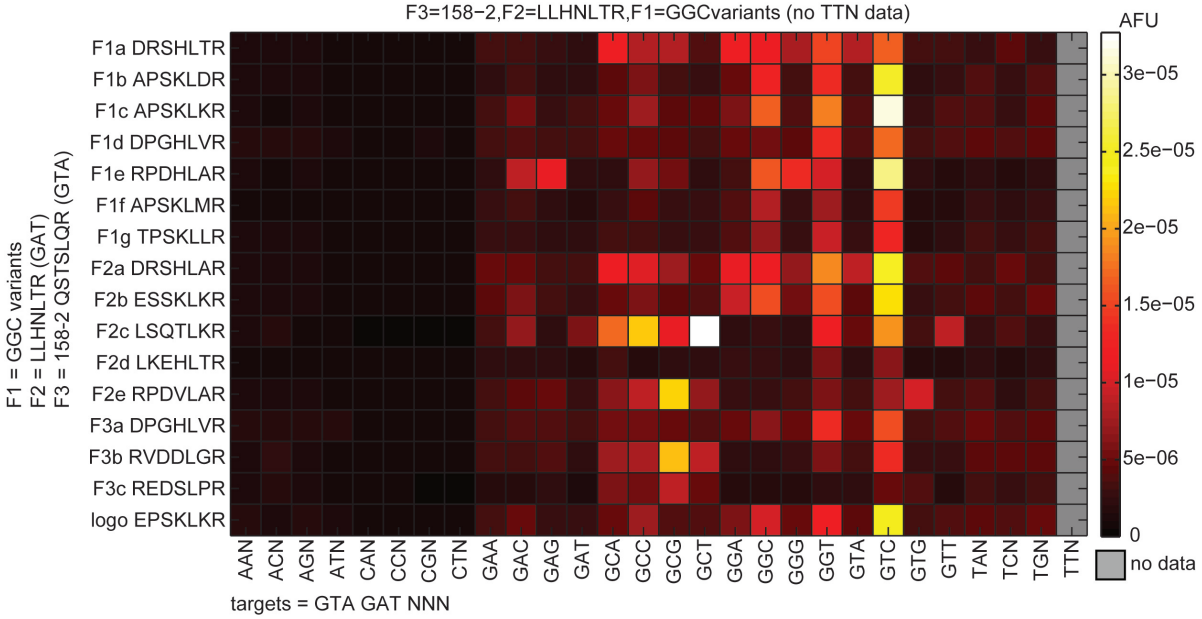


Figure 4.2.3 : Heat map of affinity data from F1 RH variants selected to bind GGC with F3 from 158-2, which resulted in a weakened affinity across the entire target range, but also reduced the non-specific binding ‘noise’ seen in Figure 4.2.2. In this screen, while there is no high-specificity variant for GGC, by comparing the relative specificities for GGC and GTC, F2B ESSKLKR was selected. The ‘logo’ RH at the bottom of the heat map was generated by taking the highest frequency residue in each RH position from all available GGC variants listed in the Zinc Finger Database.

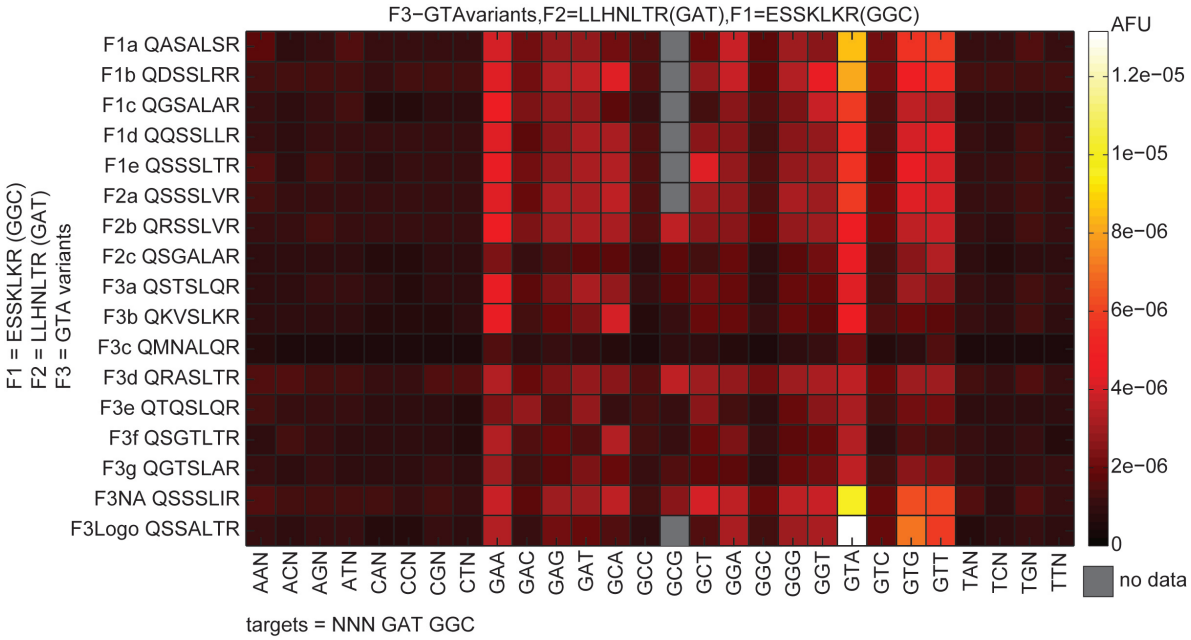


Figure 4.2.4 : Heat map of affinity data from F3 RH variants selected to bind GTA, with the selected RHs from the F2 and F1 screening rounds. In this screen, the highest affinity variant was the 'logo' design, which was generated by taking all of the available GTA variants listed in the Zinc Finger Database and selecting the highest frequency residue at each position (QSSALTR). This RH was chosen to complete the 3 selection rounds towards developing a ZFA that recognizes the sequence GTA GAT GGC.

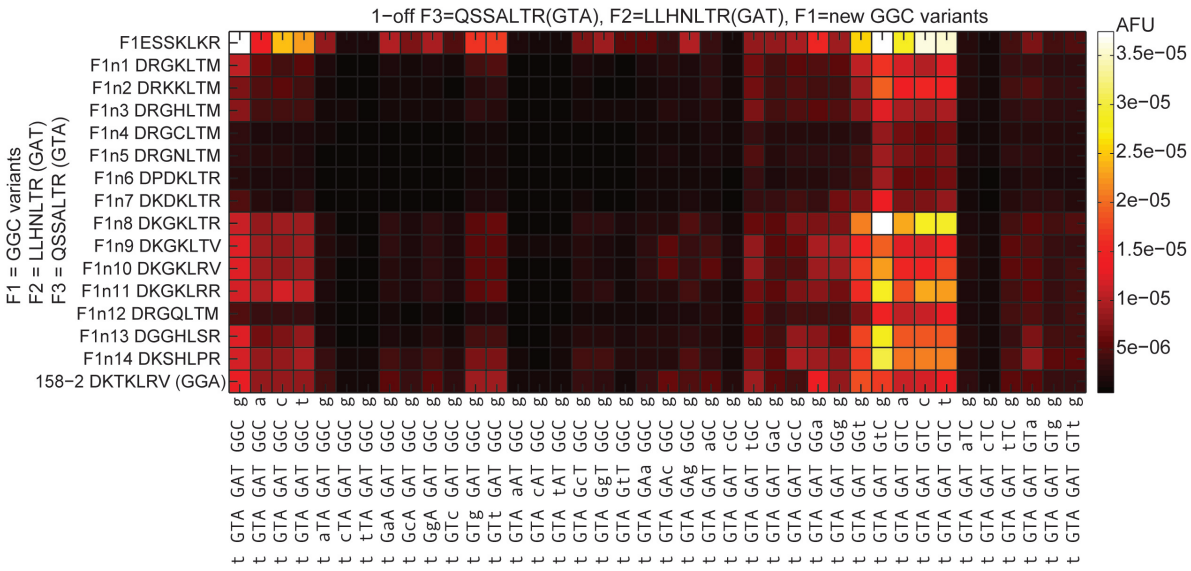


Figure 4.2.5 : Heat map of affinity data from the final engineered variant (top row) selected to bind GTA GAT GGC, in addition to 14 other F1 variants which were predicted (Figure 2.5.1) to bind more specifically to GGC than those tested in the earlier F1 selection round (Figure 4.2.3). Here all of the variants are tested against a 1-off target library to generate a detailed summary of specificity towards the target of interest. Despite the DNA specificity predictions given for the designed F1 variants, most of them also have a preference for GTC rather than the desired target GGC.

4.2.2 ZFA Specificity Engineering for ‘GCC CAC GTG’

We repeated the same sequential selection process to engineer another ZFA recognizing the sequence ‘GCC CAC GTG’, which is a more challenging target sequence given the ‘CAC’ in the F2 position. The most populated ZF designs in the ZF DB exists for fingers targeting ‘GNN’ motifs. The selection of designs for ‘CNN’ binding variants is considerably reduced, and to generate enough designs that bind ‘CAC’, we selected RHs from recent publications (see caption for Figure 4.2.6) but also conceived of new designs based on structures from those publications.

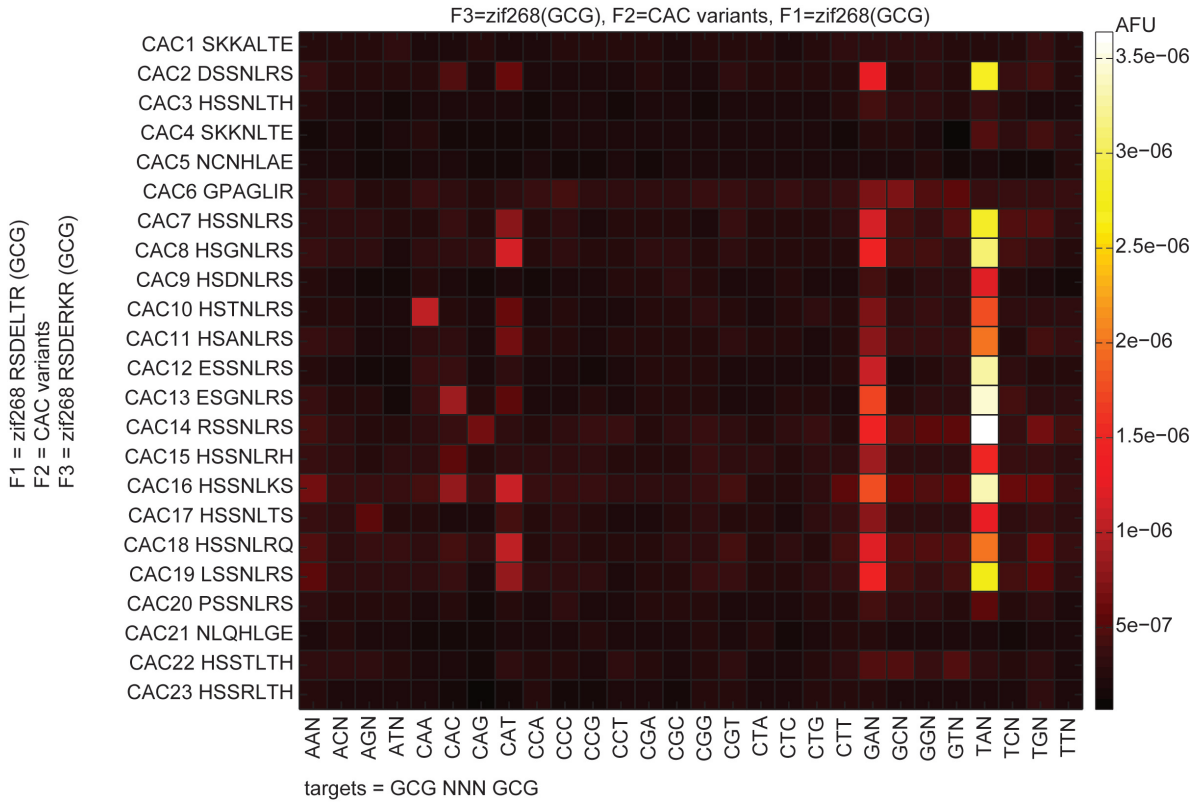


Figure 4.2.6 : Complete heat map data from all ZF TFs containing F2 variants that were selected or designed to bind the triplet CAC with F1/ F3 from Zif268. Due to a low list of options from the Zinc Finger Consortium Database, RHs were taken from recent publications, which reported CAC-binding ZF domains. CAC1, 5 and 6 came from (140), CAC2 and 3 came from (75), and CAC21 was taken from a patent application (2004, EP1421177A2). The remaining RHs were designed around the amino acid residue logos presented for CAC (75) or from half-site designs reported in (152). In this initial screen, using F1/F3 from zif268, it appeared that none of the RHs were functional for binding CAC, and instead we observed strong affinity for GAN or TAN targets.

As with the specificity engineering process in 4.2.1, we started with the Zif268 F1/F3 domains, and inserted a set of 23 F2 variants that bind 'CAC' (Figure 4.2.6). This first screen hinted at a few RH variants that may bind 'CAC', but the majority of the designs displayed high specificity and affinity towards 'GAN' and 'TAN' targets. We believe this result can be explained by the strong cross-site interaction of the Zif268 F3 aspartic acid in position 2 of the RH, which had been previously observed (139). The natural F2 target in Zif268 is TGG/GGG, and so the cross-strand interaction would prefer an A or C in the first base of the F2 target complement, which was only available in the GNN and TNN targets. Since the F2 variants were selected for binding 'CAC', the second base that is recognized is an 'A'. Because this screen with F1/F3 from Zif268 did not produce any high affinity CAC binding variants, we performed a second screen (Figure 4.2.7) using F1/F3 from a recent publication testing 'CNN' binding variants (75).

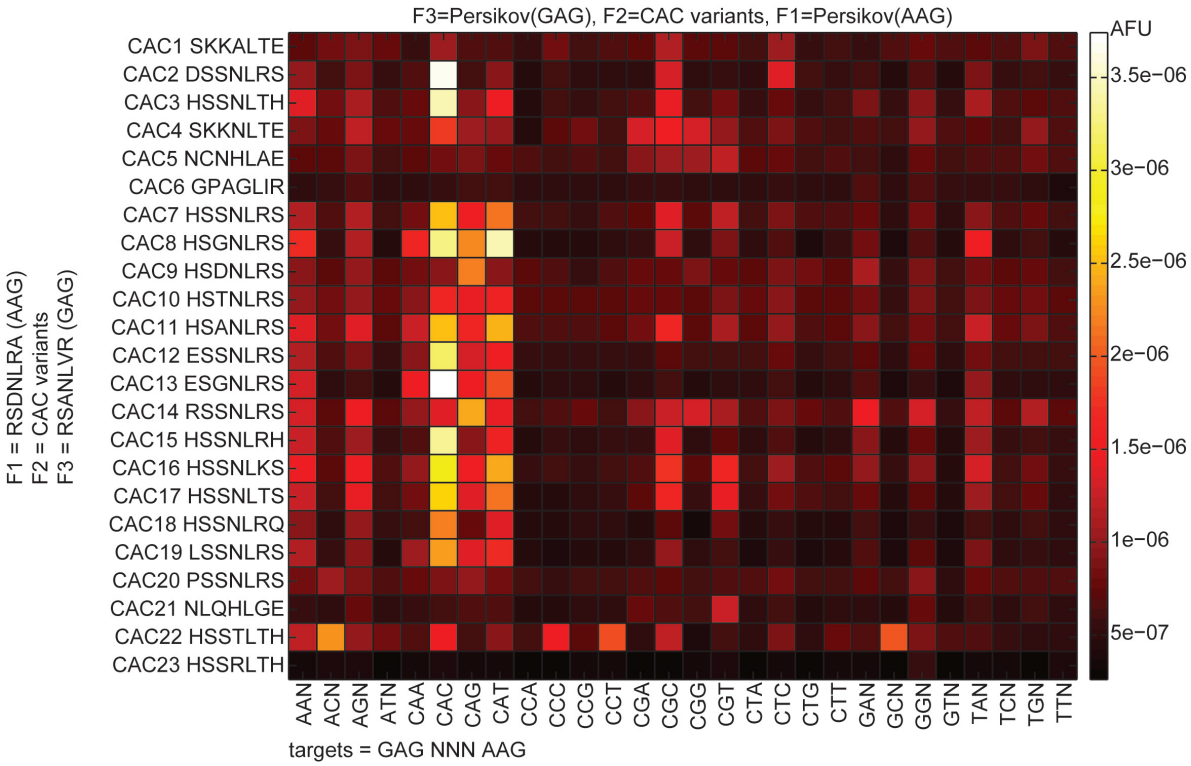


Figure 4.2.7 : Heat map of affinity data from all ZFAs containing F2 variants that were selected or designed to bind the triplet CAC with F1/ F3 from (75). This F1/F3 set is the context within which CAC2 and 3 were originally tested, so we knew at least these variants should be specific for CAC. As reported in ref 22, CAC1 displayed only weak binding towards CAC. While most of the variant designs did not come from published examples, many of them were capable of binding CAC, but with reduced specificity. In the end, variant CAC13 was chosen for subsequent screens since it was a novel design and exhibited relatively high specificity and affinity towards CAC.

Some of the 'CAC' designs we tested within F1/F3 from Zif268 were originally developed and tested within the F1/F3 context (referred to as the 'Persikov' framework for the first author) from this publication. Since these designs had been previously validated in another study, we were confident they would function as a good control for screening new 'CAC' variants. As expected, within the proper F1/F3 context, many of the designs in our screen showed high affinity and specificity for the desired 'CAC' target. We chose the F2 variant ESGNLRS since it displayed high affinity and specificity towards CAC and was an entirely novel ZF design. The next step involved substituting variants into the F1 position that were expected to bind 'GTG', and here, the F1 variant was very easy to select since the majority of the designs from the ZF DB did not display high affinity towards the target panel (Figure 4.2.8). The RH variant RKDVLTR was chosen since it displayed exceptional binding affinity and specificity to the 'GTG' target.

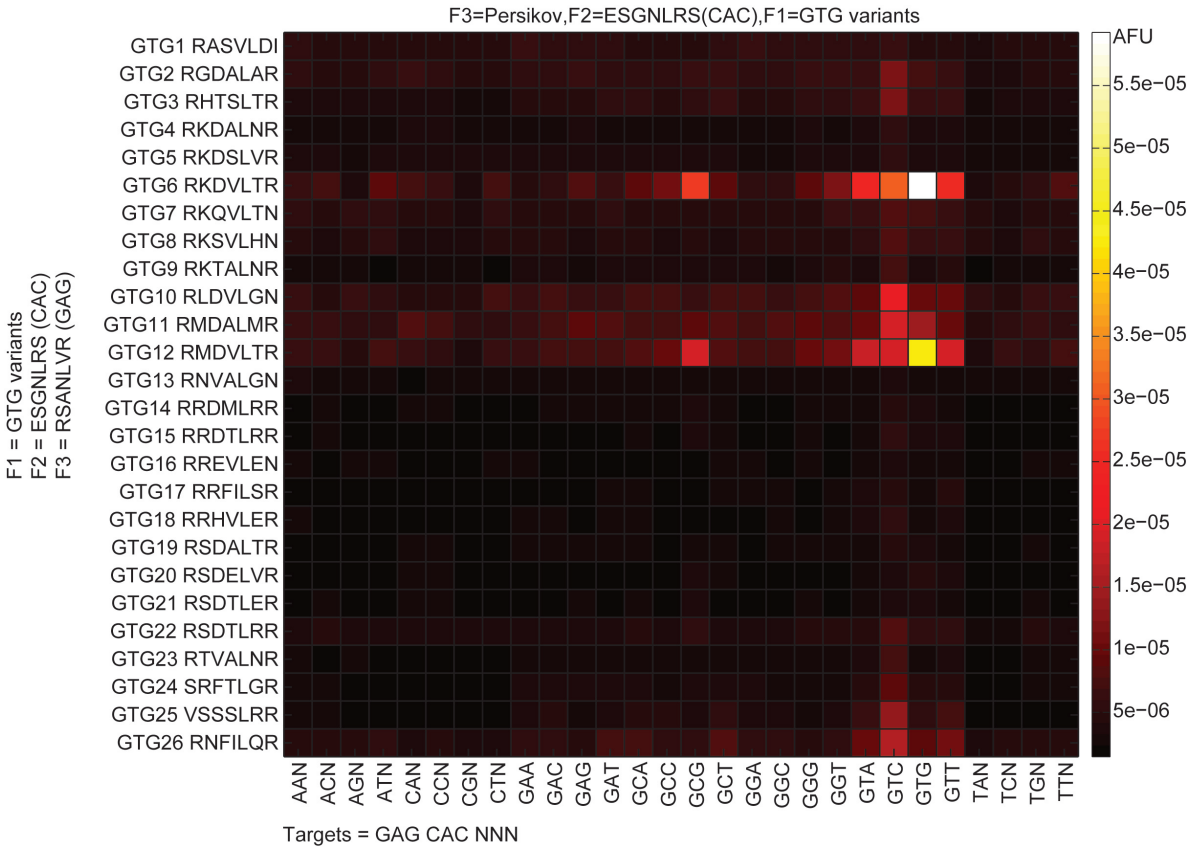


Figure 4.2.8 : Heat map of affinity data from F1 RH variants selected to bind GTG with F3 from (75) and F2 (ESGNLRS) from the CAC selection in Figure 4.2.7. A large majority of the RHs tested were not functional, or bound with relatively low affinity, and many of the variants had a binding preference for GTC in addition to the desired target GTG. RH variant RKDVLTR was selected for the subsequent screens in spite of its secondary binding preference for GTC.

The third and final selection round involved inserting RH variants for F3 that were expected to bind the target ‘GCC’. Finding a high affinity, yet specific RH in position F3 for GCC was more challenging and we settled on using a RH that displayed high affinity towards both ‘GCC’ and ‘GTC’ (Figure 4.2.9). The RH variant EGGTLRR was selected for use in the final engineered ZFA design.

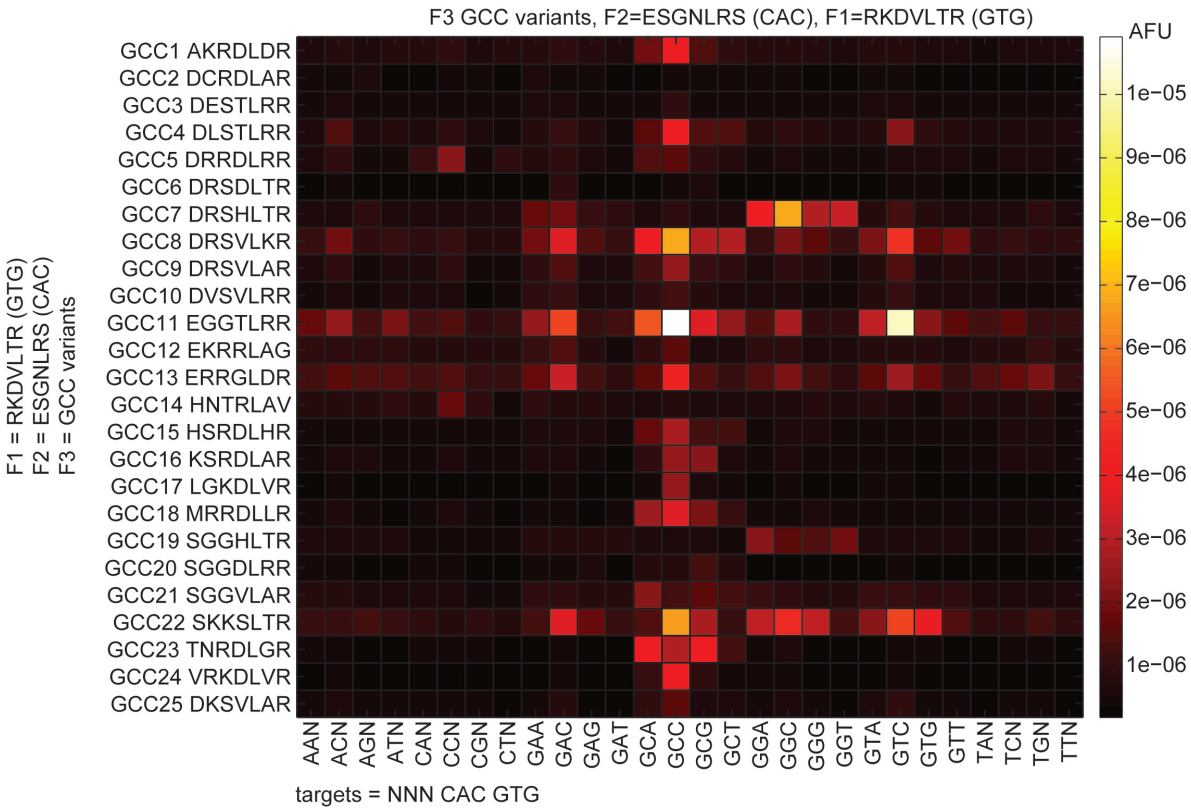


Figure 4.2.9 : Heat map of affinity data from F3 RH variants selected to bind GCC with F1 from the GTG selection in Figure 4.2.8 and F2 from the CAC selection in Figure 4.2.7. In this selection, the majority of the variants were functional and displayed high specificity for GCC, but the highest affinity variants also displayed non-specific binding to other triplets. A selection of these variants was also characterized against a 1-off target library (Figure 4.2.10).

Although each individual RH appeared to bind the intended target sequence with high specificity, when testing the engineered ZFA against a one-off library, the highest affinity target was ‘GTC CAT GTG’, not the desired target ‘GCC CAC GTG’, which was bound with a lower affinity (Figure 4.2.10). Other RHs from the F3 ‘GCC’ screen (Figure 4.2.9), which displayed lower affinity but higher specificity for ‘GCC’ were also tested against a one-off library, but all of these exhibited higher affinity for other targets than ‘GCC CAC GTG’.

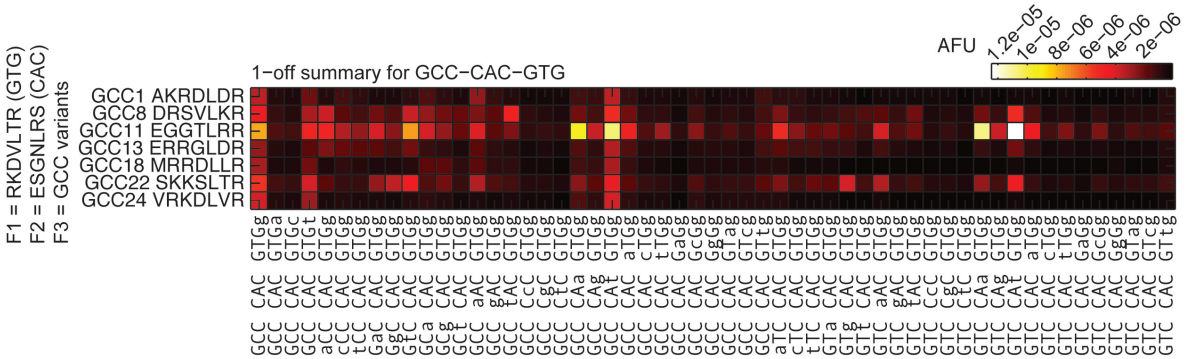


Figure 4.2.10 : Heat map of affinity data from the final engineered ZFA selected to bind GCC CAC GTG, in addition to several other F3 (GCC) variants, which displayed lower affinity but higher specificity to GCC. Here all of the variants are measured against a 1-off target library based on the consensus sequences 'GCC CAC GTG' and 'GTC CAC GTG' to generate a detailed evaluation of their specificity towards the target of interest. All of the variants have some affinity for CAT rather than CAC, and it becomes clear that a few of the F3 variants have a preference for GTC rather than GCC. In most cases, the target of interest (GCC CAC GTG) is bound, but with equivalent or lower affinity than other targets.

These results indicate that engineering ZFAs to bind a particular sequence is relatively easy to achieve using a stepwise selection process, but engineering the precise specificity landscape is more challenging. The individual specificity profiles of RHs can vary considerably, and in some instances do not reflect their annotated or predicted sequence specificity given in the ZF Consortium Database or online binding site predictors. One significant advantage of APE-MITOMI based assembly and characterization is the fact that the method returns information on the precise specificity landscape and affinity of the synthesized transcription factors during all stages of assembly. APE-MITOMI allows the generation of ZFAs with similar consensus sequences, but different specificity and/or affinity profiles and enables the examination of these characteristics for the function of native and synthetic transcriptional regulatory networks.

4.3 Discussion

As demonstrated in the two engineering studies detailed in this chapter, by using previously published RHs from a variety of sources, in addition to binding site predictor softwares, the rational design of ZFAs is possible. The use of the APE-MITOMI enables the rapid production of linear template libraries of ZFA gene variants and high-throughput,

quantitative mapping of their binding specificities. Although the Zinc Finger Consortium Database was a useful resource for selecting ZF domain parts, future experiments should incorporate a computational screening step prior to selecting variants for APE assembly. This would improve the rate of selecting functional designs and help to limit the number of variants that need to be screened before finding designs that meet the engineering objectives.

Although both of our engineered variants exhibited off-target binding, these designs were only the first implementation of our rapid protein engineering platform. We did not explore the possibilities of making larger ZFAs that contain more than 3 linked ZF domains. Longer ZFAs bind to longer target sites, so that degeneracy in a single finger position may be overcome by including additional domains with high specificity. In addition to improving the specificity of an engineered ZFA with additional ZF domains, the affinity can also be augmented (277). Another potential option to explore for improving the specificity of engineered variants would be to perform the rounds of selection in a different order. Beerli and Barbas recommend three different strategies for producing ZFAs with engineered binding specificity: parallel, sequential and bipartite selection (154).

As studies related to ZF domains are continuously being performed, new results are offering added insight into how amino acid sequence, linker structure, and neighboring ZF domains influence target specificity (278, 279). Data from these studies are leading to improved models of ZF-DNA interactions and aiding motif prediction algorithms, which in turn will help make a technique like APE-MITOMI more powerful for generating and testing rationally designed variants.

4.4 Methods

The methods used in this Chapter are identical to those detailed in Chapter 2 (for APE assembly and linear template preparation) and 3 (for ZFA assembly, MITOMI characterization, and data analysis).

4.5 Supplementary Figures

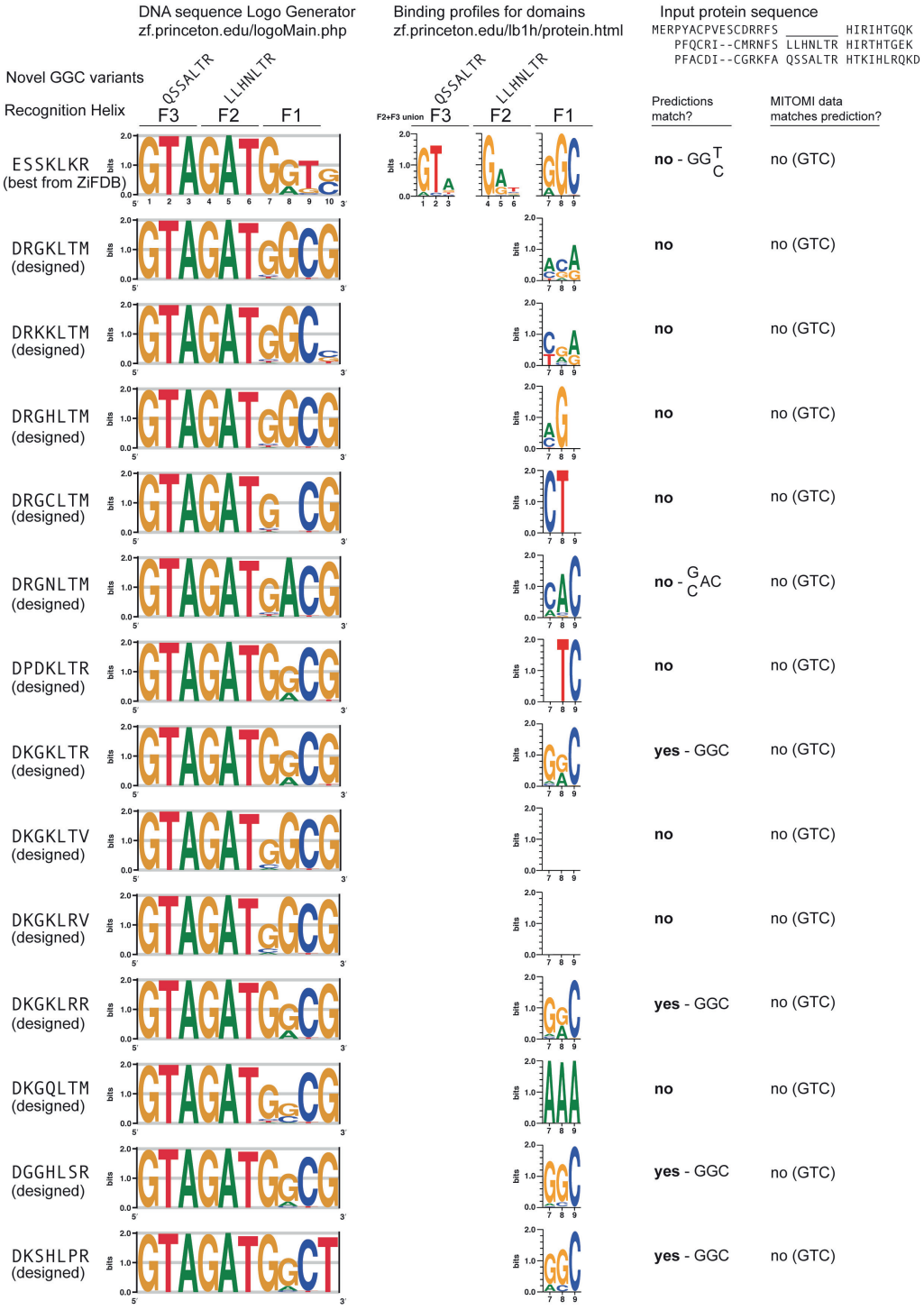


Figure 4.5.1 : Two tables of predicted DNA binding specificities for GGC-binding variants taken from the ZF Consortium Database, and novel variants created by substituting in new residues, using online prediction programs (75, 76, 145, 276). The amino acid sequence of the complete ZFA sequence (containing all three ZF domains) is given as input to the programs, which detect the RHs (residues -1 to 6) then predict the DNA binding site of each ZF domain. In the figure above, the leftmost column provides the amino acid sequence of the RH variant of interest. The output of each program is given in either the second or third column, respectively, as sequence logos. The fourth column indicates whether the two predictions agree with each other, and the final column compares the prediction with the observed MITOMI binding preference (target bound with the highest affinity).

Chapter 5: C_2H_2 ZF Affinity Variant Engineering

5.1 Introduction

Focus in ZF engineering has been on rationally designing ZF specificity, but it is unclear to what extent affinity can be tuned and whether affinity can be tuned independently of sequence specificity. The ability to tune ZF affinity is of interest in creating synthetic transcriptional regulatory networks (156) and it would drastically simplify the task of engineering ZFAs if affinity could be tuned independently of specificity. Based on the Zif268 crystal structure we selected 28 residues that could be involved in determining the protein's affinity to DNA, and changed these residues to alanine (Figure 5.1.1). These variants were assembled and characterized using APE-MITOMI to evaluate how these modifications altered binding specificity and affinity. Each experiment was performed with WT Zif268 present in each MITOMI device, and all variants were tested against a dilution set of the Zif268 consensus target, as well as single concentrations of a 1-off target library based on the Zif268 consensus sequence ('GCG TGG GCG'). Changes in binding affinity towards the consensus target were determined using data from the consensus target dilution set. Changes in binding specificity were determined using data from the 1-off target library.

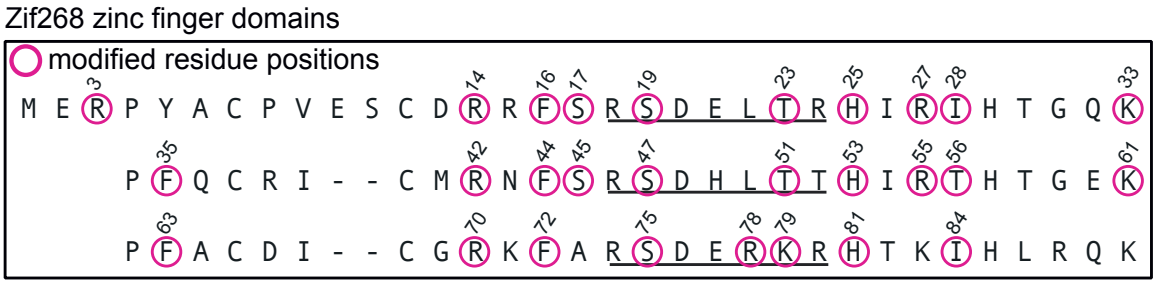


Figure 5.1.1 : Alignment of amino acid residues from the ZF domains of Zif268, where magenta circles indicate positions that make non-specific contacts with the DNA target backbone, based on the crystal structure of Zif268 with its consensus target. These positions were mutated to alanine to evaluate their effect on DNA binding affinity. Underlined residues indicate the location of the recognition helix (-1 to 6) of each ZF domain.

5.2 Results

In all previous analyses of ZFA binding, a 'relative affinity' value in arbitrary fluorescence units was calculated from observed levels of bound target, total target, and total protein to generate a single value (see Equation 3.4.1). For analyzing the affinity variants, the fold change in binding affinity towards the WT Zif268 consensus target was determined using the target dilution series data set for each variant. By taking the ratio of surface bound target DNA to surface bound protein ($Cy5_{bound}$/GFP) and plotting the ratio as a function of total target DNA ($Cy5_{total}$, also in AFU), a linear fit was used to derive a slope value which was taken as the 'relative affinity' for the consensus target (Figure 5.2.1).

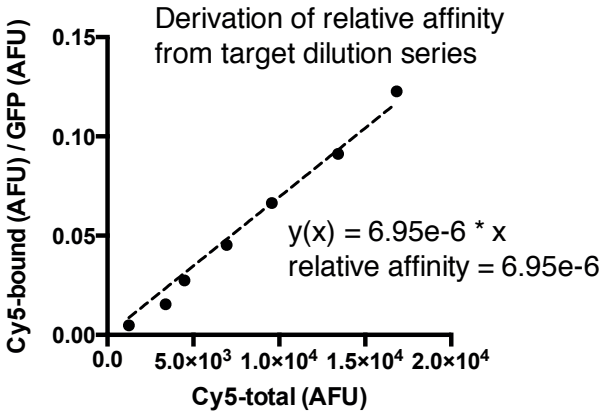


Figure 5.2.1 : Example of plotting MITOMI data from a target dilution series to derive a 'relative affinity'.

Relative affinity values were calculated in this manner for each variant. Fold change in binding affinity was determined by dividing the average relative affinity for each ZFA

variant by the average relative affinity for Zif268. As expected, most alanine substitutions resulted in modest decreases in affinity as compared to wildtype (WT) Zif268 (up to 2-fold; Figure 5.2.2). To our surprise, a few modifications led to increased binding affinity (R14A, I28A, R70A). In addition to testing the alanine substitution variants, residues from Sp1 (a TF with a 3-finger ZFA domain, commonly used as an alternative to the Zif268 framework (117, 280)), were placed into the few positions which exhibited increased binding affinity in the alanine scan. Four of these non-alanine residue changes also led to increases in binding affinity, and may partially explain differences in binding affinity observed between artificial TFs designed using the Zif268 and Sp1 frameworks.

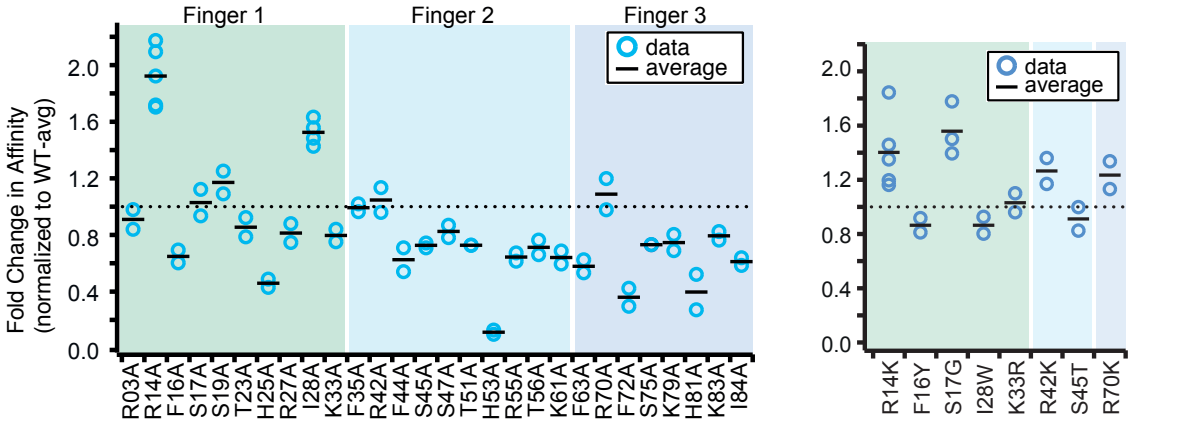


Figure 5.2.2 : (Left) Results of single residue alanine scan at the residue positions listed and the resulting fold-change in binding affinity relative the WT zif268 consensus target. (Right) Results of single residue substitutions using non-alanine amino acids, taken from the Sp1 framework.

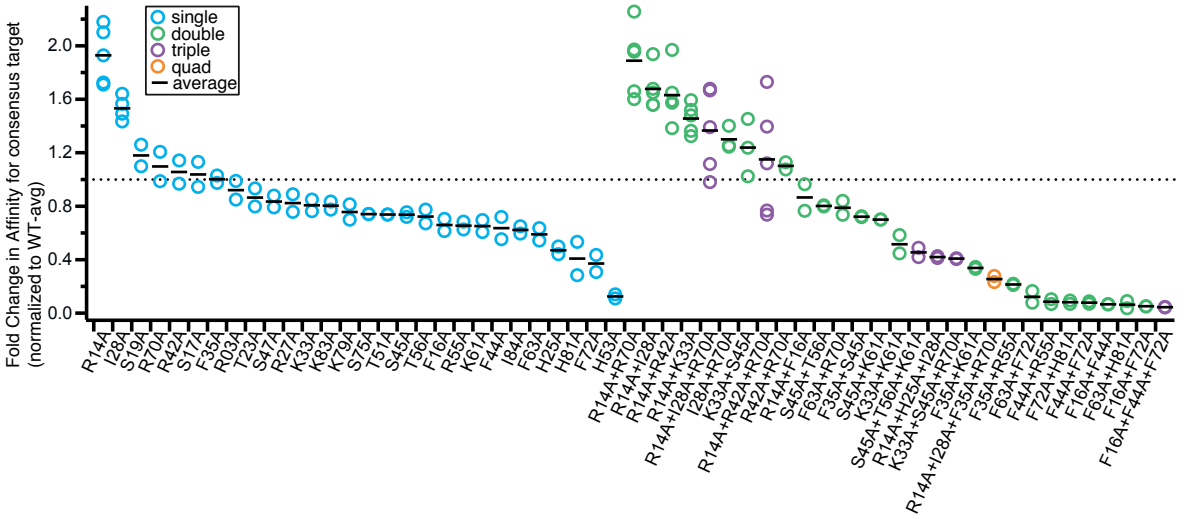
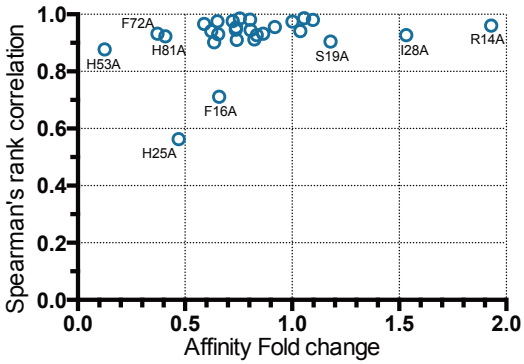


Figure 5.2.3 : Rank ordered fold-change in binding affinity relative to WT consensus target for single and multiple mutant variants.

Given that the single substitutions primarily resulted in modest affinity decreases, we proceeded to assemble and test 22 double substitutions, 6 triple substitutions and 1 quadruple substitution (Figure 5.2.3). These novel mutants allowed us to extend the dynamic range to 1/5 of WT and allowed us to smoothly tune affinity over the entire accessible range between 2x increased and 5x reduced affinity.

We could also show that most of the alanine substitution mutants retain their specificity landscape, indicating that affinity can be tuned independently of specificity. The Spearman's rank correlation coefficient was determined by comparing the relative affinity values of each variant with those of Zif268, when tested with a 1-off library based on the Zif268 consensus target (Figure 5.2.4). Except for two alanine mutants (F16A and H25A), most of the other affinity variants exhibit nearly identical binding specificity compare to WT Zif268, but display at wide range of affinities.

name	Fold change	Spearmans r
R03A	0.91975	0.95553
R14A	1.92895	0.96043
F16A	0.65925	0.71152
S17A	1.03805	0.94123
S19A	1.17970	0.90451
T23A	0.86478	0.93176
H25A	0.47010	0.56274
R27A	0.82351	0.91201
I28A	1.53270	0.92675
K33A	0.80639	0.94420
F35A	1.00195	0.97462
R42A	1.05607	0.98630
F44A	0.63581	0.90239
S45A	0.73692	0.94611
S47A	0.83495	0.92732
T51A	0.73796	0.95412
H53A	0.12509	0.87681
R55A	0.65476	0.93045
T56A	0.72269	0.97588
K61A	0.65126	0.97577
F63A	0.58931	0.96605
R70A	1.09723	0.98021
F72A	0.37127	0.93171
S75A	0.74140	0.91020
K79A	0.75666	0.98539
H81A	0.40791	0.92325
K83A	0.80416	0.98101
I84A	0.62245	0.94153



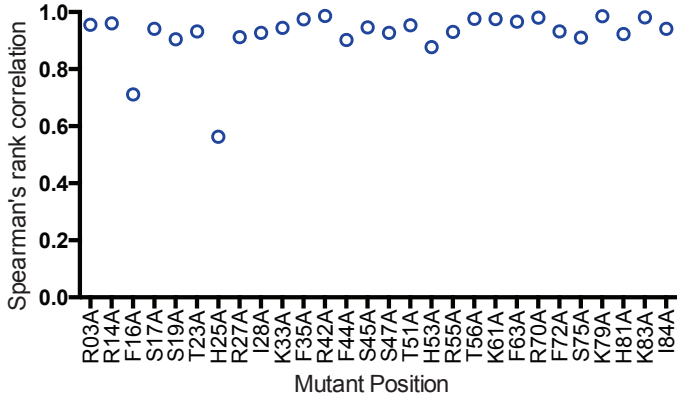


Figure 5.2.4 : Analysis of data obtained from examining Zif268 alanine substitution affinity variants (Figure 5.2.2) tested against a 1-off target library for the Zif268 consensus target (GCG TGG GCG). Spearman's rank correlation coefficient was determined by comparing relative affinity values with Zif268 for the 1-off target library. Nearly all of the variants correlate highly, meaning none have lost their target specificity in spite of their affinity being modulated (fold change in affinity values derived from experiments using a dilution series of the Zif268 consensus target).

The majority of variants have a Spearman's rank correlation greater than 0.9, except for 2 variants, F16A and H25A, which displayed severe departures from the specificity of the WT protein as a result of the mutation. These variants have a higher than expected affinity for certain targets, and lower than expected affinity for others as seen in the individual scatter plots displayed in Figure 5.5.1.

5.3 Discussion

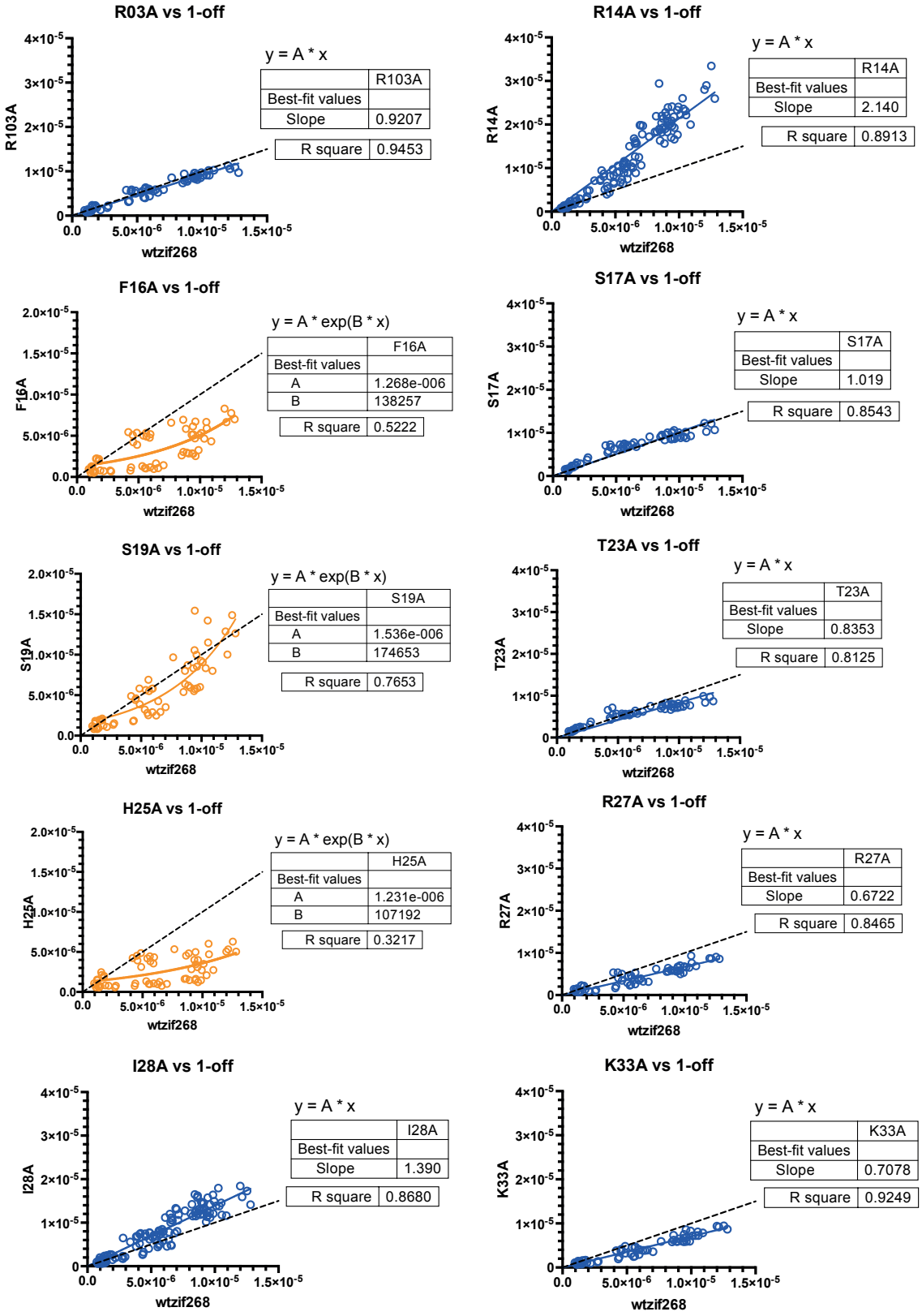
In this chapter, we demonstrate the utility of APE-MITOMI to rapidly construct ZFA variants for modifying target binding affinity. This work was initially inspired by results given in a recent publication detailing the development of synthetic TFs (sTF) for eukaryotic gene networks(156). It is interesting to note that in this publication, they report the creation of sTFs with tunable transcriptional output by replacing multiple arginines with alanines in the ZF-backbone, but to see a decrease in output, they needed to make three or more mutations. These *in vivo* results correspond well with what was observed *in vitro* using our affinity variants, since many of the single and double R→A mutations lead to an increased binding affinity. The ability to tune affinity independently from specificity is an incredibly useful tool for improving weak, but highly specific interactions. Since these modifications are made to the framework sequence of ZFAs, it should be applicable to any of the protein designs that have been made in Chapter 3 or 4. Another option for tuning affinity, which was not explored in this thesis, would be to explore how the linking sequence between ZF domains affects DNA binding specificity or affinity. Previous studies reported the development of extremely high affinities simply by changing the linker design (154, 277, 281) .

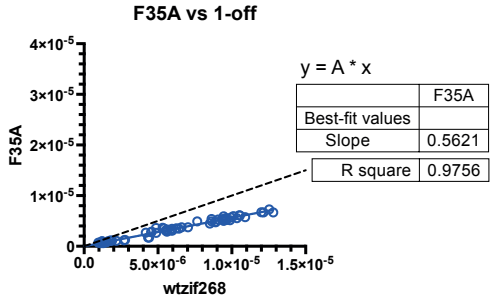
5.4 Methods

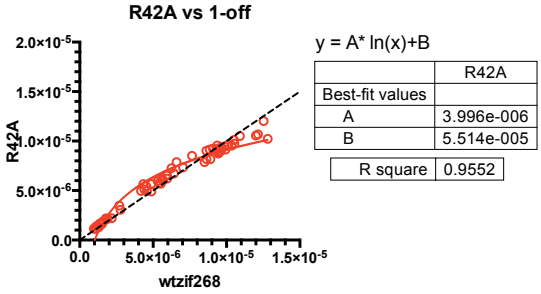
The methods used in this Chapter are identical to those detailed in Chapter 2 (for APE assembly and linear template preparation) and 3 (for ZFA assembly, MITOMI characterization, and data analysis).

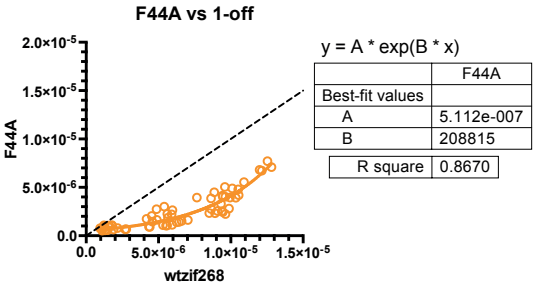
5.5 Supplementary Figures

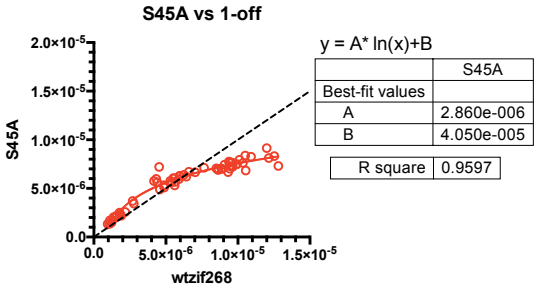
Figure 5.5.1 : For each variant, we fit a curve to the 1-off target data set, using a linear (blue), exponential (yellow) or log (red) equation to capture the behavior of the data. The black dashed line signifies the behavior of WT Zif268 plotted against itself, to compare against the behavior of the mutant variants. Variants with increased binding affinity towards 1-off targets have curve fits above the dashed black line. Those with decreased binding affinity have curve fits below the dashed black line. The curve fit constants and R^2 value is reported beside each plot.

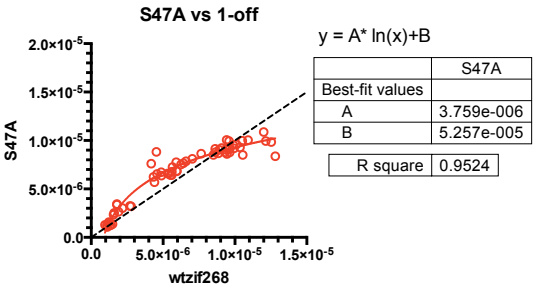


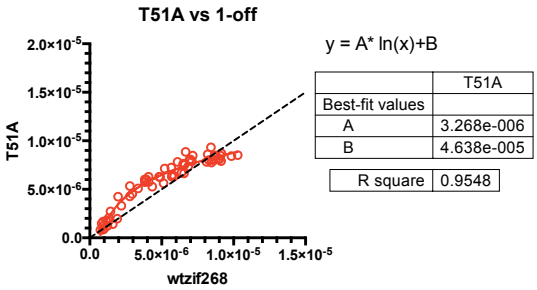

F35A vs 1-off
y = A * x
F35A
Best-fit values
Slope 0.5621
R square 0.9756
F35A
wtzif268

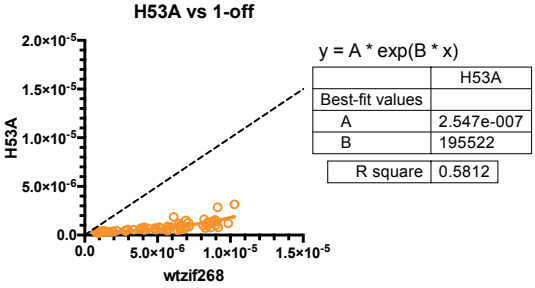

R42A vs 1-off
y = A* ln(x)+B
R42A
Best-fit values
A 3.996e-006
B 5.514e-005
R square 0.9552
R42A
wtzif268

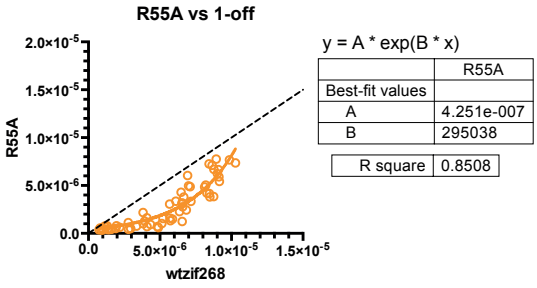

F44A vs 1-off
y = A * exp(B * x)
F44A
Best-fit values
A 5.112e-007
B 208815
R square 0.8670
F44A
wtzif268

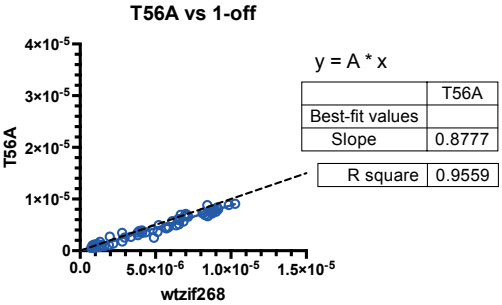

S45A vs 1-off
y = A* ln(x)+B
S45A
Best-fit values
A 2.860e-006
B 4.050e-005
R square 0.9597
S45A
wtzif268

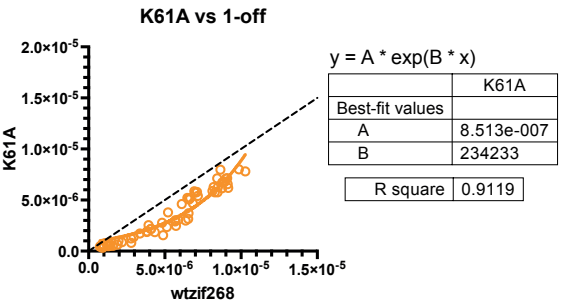

S47A vs 1-off
y = A* ln(x)+B
S47A
Best-fit values
A 3.759e-006
B 5.257e-005
R square 0.9524
S47A
wtzif268

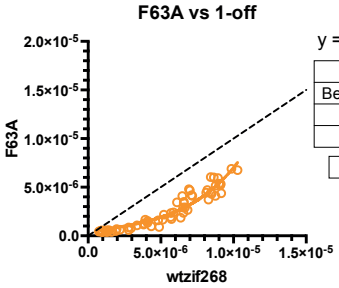

T51A vs 1-off
y = A* ln(x)+B
T51A
Best-fit values
A 3.268e-006
B 4.638e-005
R square 0.9548
T51A
wtzif268


H53A vs 1-off
y = A * exp(B * x)
H53A
Best-fit values
A 2.547e-007
B 195522
R square 0.5812
H53A
wtzif268


R55A vs 1-off
y = A * exp(B * x)
R55A
Best-fit values
A 4.251e-007
B 295038
R square 0.8508
R55A
wtzif268

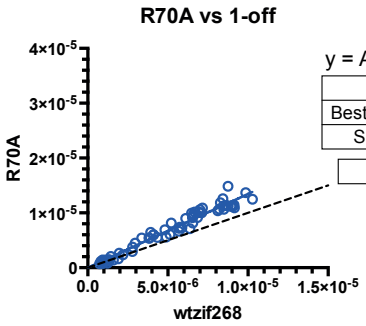

T56A vs 1-off
y = A * x
T56A
Best-fit values
Slope 0.8777
R square 0.9559
T56A
wtzif268


K61A vs 1-off
y = A * exp(B * x)
K61A
Best-fit values
A 8.513e-007
B 234233
R square 0.9119
K61A
wtzif268



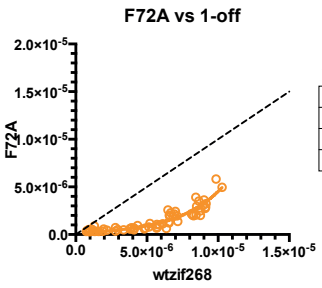
y = A * exp(B * x)

	F63A
Best-fit values	
A	5.198e-007
B	260156
R square	0.9078



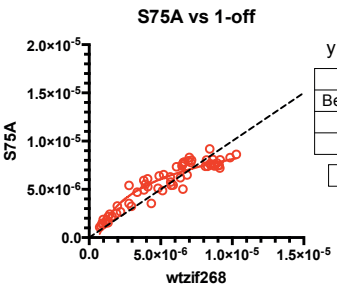
y = A * x

	R70A
Best-fit values	
Slope	1.345
R square	0.9565



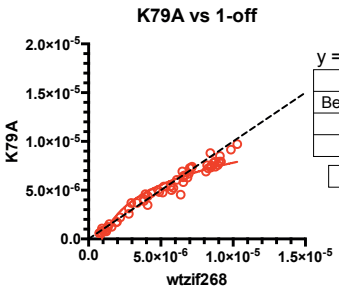
y = A * exp(B * x)

	F72A
Best-fit values	
A	1.872e-007
B	318330
R square	0.8612



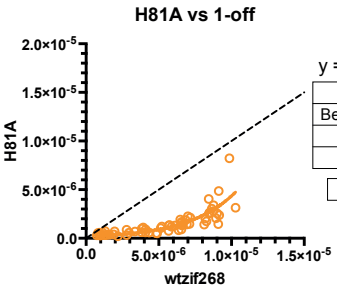
y = A* ln(x)+B

	S75A
Best-fit values	
A	2.981e-006
B	4.245e-005
R square	0.9087



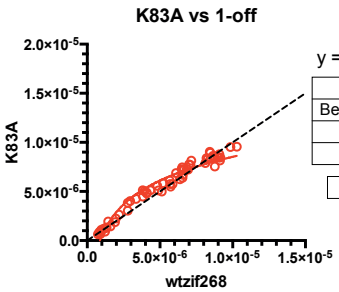
y = A* ln(x)+B

	K79A
Best-fit values	
A	3.278e-006
B	4.555e-005
R square	0.9275



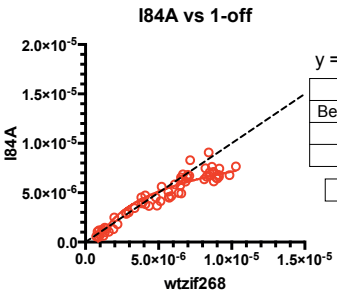
y = A * exp(B * x)

	H81A
Best-fit values	
A	1.922e-007
B	311842
R square	0.6771



y = A* ln(x)+B

	K83A
Best-fit values	
A	3.500e-006
B	4.883e-005
R square	0.9505



y = A* ln(x)+B

	I84A
Best-fit values	
A	2.904e-006
B	4.057e-005
R square	0.9010

Chapter 6: Conclusions and Outlook

The work presented in this thesis outlines the development of a novel approach towards accelerating protein engineering. By coupling a gene assembly strategy with high-throughput on-chip protein expression, purification, and quantitative characterization, we demonstrate that it is possible to completely circumvent molecular cloning and cell-based processing steps generally needed in protein biochemical analysis. Integration of these methods reduced the time required between protein design and protein characterization from weeks to days. We validated the utility and versatility of this process by synthesizing over 400 zinc finger array (ZFA) variants, and characterized the binding specificity of each for a total of over 98,000 protein-DNA interaction measurements.

In Chapter 2, we described the process for developing a solid-phase, ligase and restriction enzyme-free gene assembly technique that utilizes commodity-scale DNA oligomers without the need for purification steps. The error rate of this technique was found to be competitive with other methods, and the modularity of the assembly process allows for the construction of both diverse and repetitive gene elements. Since the assembly technique utilizes magnetic beads, it is amenable to scale-up via robotic liquid handling systems.

In Chapter 3, we explored the modularity of ZFAs, and using a limited set of DNA oligomers, assembled a library of protein variants and mapped their binding specificity profiles using MITOMI. The results of this study validated the speed and sensitivity of APE-MITOMI, and proved that our platform could be extended towards other ZFA engineering applications.

In Chapters 4 and 5, we applied the APE-MITOMI process towards rationally designing specificity and affinity ZFA variants. We took advantage of existing ZF protein structure databases and models to engineer two ZFA variants that bind specific target sequences. The binding specificity landscapes for both variants were mapped throughout each cycle of design, assembly and characterization, providing a detailed overview of the engineering process. Additionally, by selectively introducing alanine substitutions into the ZFA framework outside of regions that control specificity, we were able to create ZFA variants for tuning binding affinity without affecting specificity.

The application of APE-MITOMI towards the engineering of C_2H_2 ZF variants was only a proof-of-concept for rapid prototyping of synthetic biological designs. In general, the APE gene assembly technique could be used for potentially any kind of protein engineering application, particularly for instances where protein variants are very similar in sequence, and where crystal structures are readily available to guide the selection of novel designs. Some instances of projects that could already be initiated include examining the protein or RNA-binding activity of ZFAs, or engineering other types of protein domains. Repeat or solenoid proteins would be interesting to explore, since they are modular, well characterized, can mediate a range of protein-protein interactions, and have been applied to engineering applications (282). Still other protein structures that could be studied are scFvs and nanobodies (283), designed ankyrin repeat proteins (DARPins) (284), PDZ domains, leucine zipper domains, SH2/SH3 domains, and also engineering new functions from existing protein structures (130). One interesting option outside of protein applications would be to study oligonucleotide aptamers (70).

It has been shown previously that over 400 full-length drosophila transcription factors of varying size and families could be expressed on a single MITOMI device (210), indicating that there is no major bottleneck to on-chip protein expression. Although all of the genes

studied in this thesis were prepared on bench-top by hand, the APE gene synthesis method can be readily automated on liquid handling robots, eliminating the last limitation in the process. It should thus be possible to characterize thousands of synthetic protein variants per week, which in turn will enable exploration of the protein sequence-function relationship in unprecedented detail, aid in the development of accurate computational predictions of protein function, and allow us to rapidly engineer novel proteins.

Bibliography

1. Church,G.M., Elowitz,M.B., Smolke,C.D., Voigt,C.A. and Weiss,R. (2014) Realizing the potential of synthetic biology. *Nat Rev Mol Cell Biol*, **15**, 289–294.

2. Roy,S. and Caruthers,M. (2013) Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. *Molecules*, **18**, 14268–14284.

3. Nirenberg,M. (2004) Historical review: Deciphering the genetic code--a personal account. Trends in Biochem Sci, **29**(1), 46-54.

4. Yanofsky,C. (2007) Establishing the Triplet Nature of the Genetic Code. *Cell*, **128**, 815–818.

5. Elkin,L.O. (2003) Rosalind Franklin and the Double Helix. *Physics Today*, **56**, 42–48.

6. Letsinger,R.L. and Mahadevan,V. (1965) Oligonucleotide Synthesis on a Polymer Support. *J. Am. Chem. Soc.*, **87**, 3526–3527.

7. Letsinger,R.L., Finnan,J.L. and Heavner,G.A. (1975) Nucleotide chemistry (part XX)- Phosphite coupling procedure for generating internucleotide links. *Journal of the Am Chem Soc* **97**, 3278–3279.

8. Beaucage,S.L. and Caruthers,M.H. (1981) Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Letters, **22**(20), 1859-1862*.

9. Caruthers,M.H. (2013) The Chemical Synthesis of DNA/RNA: Our Gift to Science. *J. Biol. Chem.*, **288**, 1420–1427.

10. Tang,N., Ma,S. and Tian,J. (2013) Chapter1: New Tools for Cost-Effective DNA Synthesis. Synthetic Biology, First Edition. Elsevier Inc, 3-21.

11. Kosuri,S. and Church,G.M. (2014) Large-scale de novo DNA synthesis: technologies and applications. *Nat Meth*, **11**, 499–507.

12. Allen,S.D., Luebke,T.M. and Rose,S.D. (2007) Ultramers- the longest oligonucleotides available with mass spectrometry QC. *Integrated DNA Technologies Technical Report*.

13. LeProust,E.M., Peck,B.J., Spirin,K., McCuen,H.B., Moore,B., Namsaraev,E. and Caruthers,M.H. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Research*, **38**, 2522–2540.

14. Carlson,R. (2009) The changing economics of DNA synthesis. *Nature Biotechnology*, **27**, 1091–1094.

15. Pease,A.C., Solas,D., Sullivan,E.J., Cronin,M.T., Holmes,C.P. and Fodor,S.P. (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 5022–5026.

16. Tian,J., Ma,K. and Saaem,I. (2009) Advancing high-throughput gene synthesis technology. *Mol. BioSyst.*, **5**, 714.

17. Kosuri,S., Eroshenko,N., Leproust,E.M., Super,M., Way,J., Li,J.B. and Church,G.M. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nature Biotechnology*, **28**, 1295–1299.

18. Halweg-Edwards,A.L., Grau,W.C., Winkler,J.D., Garst,A.D. and Gill,R.T. (2015) The emergence of commodity-scale genetic manipulation. *Current Opinion in Chemical Biology*, **28**, 150–155.

19. Khorana,H.G. (2012) Total synthesis of a gene. *Resonance*, **17**, 1174–1197.

20. Hughes,R.A., Miklos,A.E. and Ellington,A.D. (2011) Gene synthesis: methods and applications. *Meth. Enzymol.*, **498**, 277–309.

21. Ellis,T., Adie,T. and Baldwin,G.S. (2011) DNA assembly for synthetic biology: from parts to pathways and beyond. *Integrative Biology*, **3**, 109–118.

22. Casini,A., Storch,M., Baldwin,G.S. and Ellis,T. (2015) Bricks and blueprints: methods and standards for DNA assembly. *Nat Rev Mol Cell Biol*, 10.1038/nrm4014.

23. Haimovich,A.D., Muir,P. and Isaacs,F.J. (2015) Genomes by design. *Nat. Rev. Genet.*, **16**, 501–516.

24. Au,L.C., Yang,F.Y., Yang,W.J., Lo,S.H. and Kao,C.F. (1998) Gene synthesis by a LCR-based approach: High-level production of leptin-L54 using synthetic gene in Escherichia coli. *Biochemical and Biophysical Research Communications*, **248**, 200–203.

25. Bang,D. and Church,G.M. (2008) Gene synthesis by circular assembly amplification. *Nat Meth*, **5**, 37–39.

26. Marchand,J.A. and Peccoud,J. (2012) Building block synthesis using the polymerase chain assembly method. *Methods Mol. Biol.*, **852**, 3–10.

27. Prodromou,C. and Pearl,L.H. (1992) Recursive PCR: a novel technique for total gene synthesis. *Protein Engineering Design and Selection*, **5**, 827–829.

28. Ye,H., Huang,M.C., Li,M.-H. and Ying,J.Y. (2009) Experimental analysis of gene assembly with TopDown one-step real-time gene synthesis. *Nucleic Acids Research*, **37**, e51.

29. Huang,M.C., Cheong,W.C., Ye,H. and Li,M.-H. (2012) TopDown Real-Time Gene Synthesis. In Peccoud,J. (ed), *Gene Synthesis*, Methods in Molecular Biology. Humana Press, Totowa, NJ, Vol. 852, pp. 23–34.

30. Cheong,W.C., Lim,L.S., Huang,M.C., Bode,M. and Li,M.-H. (2010) New insights into the de novo gene synthesis using the automatic kinetics switch approach. *Analytical Biochemistry*, **406**, 51–60.

31. Zhou,X., Cai,S., Hong,A., You,Q., Yu,P., Sheng,N., Srivannavit,O., Muranjan,S., Rouillard,J.M., Xia,Y., *et al.* (2004) Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Research*, **32**, 5409–5417.

32. Tian,J., Gong,H., Sheng,N., Zhou,X., Gulari,E., Gao,X. and Church,G. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432**, 1050–1054.

33. Quan,J., Saaem,I., Tang,N., Ma,S., Negre,N., Gong,H., White,K.P. and Tian,J. (2011) Parallel on-chip gene synthesis and application to optimization of protein expression. *Nature Biotechnology*, **29**, 449–452.

34. Kim,H., Jeong,J. and Bang,D. (2011) Hierarchical gene synthesis using DNA microchip oligonucleotides. *Journal of Biotechnology*, **151**, 319–324.

35. Kim,H., Han,H., Ahn,J., Lee,J., Cho,N., Jang,H., Kwon,S. and Bang,D. (2012) 'Shotgun DNA synthesis' for the high-throughput construction of large DNA molecules. *Nucleic Acids Research*, 10.1093/nar/gks546.

36. Borovkov,A.Y., Loskutov,A.V., Robida,M.D., Day,K.M., Cano,J.A., Le Olson,T., Patel,H., Brown,K., Hunter,P.D. and Sykes,K.F. (2010) High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. *Nucleic Acids Research*, **38**, e180.

37. Ma,S., Saaem,I. and Tian,J. (2012) Error correction in gene synthesis technology. *Trends in Biotechnology*, **30**, 147–154.

38. Briggs,A.W., Rios,X., Chari,R., Yang,L., Zhang,F., Mali,P. and Church,G.M. (2012) Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic Acids Research*, 10.1093/nar/gks624.

39. Reyon,D., Tsai,S.Q., Khayter,C., Foden,J.A., Sander,J.D. and Joung,J.K. (2012) FLASH assembly of TALENs for high-throughput genome editing. *Nature Biotechnology*, **30**, 460–465.

40. Kim,H., Han,H., Shin,D. and Bang,D. (2010) A fluorescence selection method for accurate large-gene synthesis. *ChemBioChem*, **11**, 2448–2452.

41. Quan,J. and Tian,J. (2009) Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS ONE*, **4**, e6441.

42. Quan,J. and Tian,J. (2014) Circular polymerase extension cloning. *Methods Mol. Biol.*, **1116**, 103–117.

43. Engler,C., Gruetzner,R., Kandzia,R. and Marillonnet,S. (2009) Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS ONE*, **4**, e5553.

44. Weber,E., Engler,C., Gruetzner,R., Werner,S. and Marillonnet,S. (2011) A modular cloning system for standardized assembly of multigene constructs. *PLoS ONE*, **6**, e16765.

45. Li,M.Z. and Elledge,S.J. (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat Meth*, **4**, 251–256.

46. de Kok,S., Stanton,L.H., Slaby,T., Durot,M., Holmes,V.F., Patel,K.G., Platt,D., Shapland,E.B., Serber,Z., Dean,J., *et al.* (2014) Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth Biol*, **3**, 97–106.

47. Gibson,D.G., Benders,G.A., Axelrod,K.C., Zaveri,J., Algire,M.A., Moodie,M., Montague,M.G., Venter,J.C., Smith,H.O. and Hutchison,C.A. (2008) One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 20404–20409.

48. Muller,H., Annaluru,N., Schwerzmann,J.W., Richardson,S.M., Dymond,J.S., Cooper,E.M., Bader,J.S., Boeke,J.D. and Chandrasegaran,S. (2012) Assembling large DNA segments in yeast. *Methods Mol. Biol.*, **852**, 133–150.

49. Gibson,D.G. (2012) Oligonucleotide assembly in yeast to produce synthetic DNA fragments. *Methods Mol. Biol.*, **852**, 11–21.

50. Mitchell,L.A., Chuang,J., Agmon,N., Khunsriraksakul,C., Phillips,N.A., Cai,Y., Truong,D.M., Veerakumar,A., Wang,Y., Mayorga,M., *et al.* (2015) Versatile genetic assembly system (VEGAS) to assembly pathways for expression in *S. cercisiae*. *Nucleic Acids Research*, **43**, e88–e88.

51. Schmid-Burgk,J.L., Xie,Z., Frank,S., Virreira Winter,S., Mitschka,S., Kolanus,W., Murray,A. and Benenson,Y. (2012) Rapid hierarchical assembly of medium-size DNA cassettes. *Nucleic Acids Research*, **40**, e92–e92.

52. Fu,J., Bian,X., Hu,S., Wang,H., Huang,F., Seibert,P.M., Plaza,A., Xia,L., Stewart,A.F., ller,R.M.U., *et al.* (2012) Full-length RecE enhances linear-linear homologous recombination and facilitates direct cloning for bioprospecting. *Nature Biotechnology*, **30**, 440–446.

53. Zhang,Y., Werling,U. and Edelmann,W. (2012) SLiCE: a novel bacterial cell extract-based DNA cloning method. *Nucleic Acids Research*, **40**, e55–e55.

54. Gibson,D.G., Young,L., Chuang,R.-Y., Venter,J.C., Hutchison,C.A. and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Meth*, **6**, 343–345.

55. Lundqvist,M., Edfors,F., Sivertsson,Å., Hallström,B.M., Hudson,E.P., Tegel,H., Holmberg,A., Uhlén,M. and Rockberg,J. (2015) Solid-phase cloning for high-throughput assembly of single and multiple DNA parts. *Nucleic Acids Research*, **43**, e49.

56. Rajkumar,A.S., Dénervaud,N. and Maerkl,S.J. (2013) Mapping the fine structure of a eukaryotic promoter input-output function. *Nat. Genet.*, **45**, 1207–1215.

57. Schlabach,M.R., Hu,J.K., Li,M. and Elledge,S.J. (2010) Synthetic design of strong promoters. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 2538–2543.

58. Mogno,I., Kwasnieski,J.C. and Cohen,B.A. (2013) Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Research*, **23**, 1908–1915.

59. Melnikov,A., Murugan,A., Zhang,X., Tesileanu,T., Wang,L., Rogov,P., Feizi,S., Gnirke,A., Callan,C.G.J., Kinney,J.B., *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, **30**, 271–277.

60. Ivarsson,Y., Arnold,R., McLaughlin,M., Nim,S., Joshi,R., Ray,D., Liu,B., Teyra,J., Pawson,T., Moffat,J., *et al.* (2014) Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2542–2547.

61. Araya,C.L. and Fowler,D.M. (2011) Deep mutational scanning: assessing protein function on a massive scale. *Trends in Biotechnology*, **29**, 435–442.

62. Fowler,D.M., Stephany,J.J. and Fields,S. (2014) Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat Protoc*, **9**, 2267–2284.

63. Fowler,D.M. and Fields,S. (2014) Deep mutational scanning: a new style of protein science. *Nat Meth*, **11**, 801–807.

64. Stanton,B.C., Nielsen,A.A.K., Tamsir,A., Clancy,K., Peterson,T. and Voigt,C.A. (2014) Genomic mining of prokaryotic repressors for orthogonal logic gates. *Nat. Chem. Biol.*, **10**, 99–105.

65. DiMarco,R.L. and Heilshorn,S.C. (2012) Multifunctional Materials through Modular Protein Engineering. *Adv. Mater.*, **24**, 3923–3940.

66. Montiel,D., Kang,H.-S., Chang,F.-Y., Charlop-Powers,Z. and Brady,S.F. (2015) Yeast homologous recombination-based promoter engineering for the activation of silent natural product biosynthetic gene clusters. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 8953–8958.

67. Lynch,S.A. and Gill,R.T. (2012) Synthetic biology: New strategies for directing design. *Metabolic Engineering*, **14**, 205–211.

68. Gibson,D.G. (2014) Programming biological operating systems: genome design, assembly and activation. *Nat Meth*, **11**, 521–526.

69. Sun,H., Zhu,X., Lu,P.Y., Rosato,R.R., Tan,W. and Zu,Y. (2014) Oligonucleotide aptamers: new tools for targeted cancer therapy. *Mol Ther Nucleic Acids*, **3**, e182.

70. Kimoto,M., Yamashige,R., Matsunaga,K.-I., Yokoyama,S. and Hirao,I. (2013) Generation of high-affinity DNA aptamers using an expanded genetic alphabet. *Nature Biotechnology*, **31**, 453–457.

71. Rothemund,P. (2006) Folding DNA to create nanoscale shapes and patterns. *Nature*, **440**, 297–302.

72. Shih,W.M. and Lin,C. (2010) Knitting complex weaves with DNA origami. *Curr. Opin. Struct. Biol.*, **20**, 276–282.

73. Fisher,M.A. and Tullman-Ercek,D. (2013) Change, exchange, and rearrange: protein engineering for the biotechnological production of fuels, pharmaceuticals, and other chemicals. *Curr. Opin. Biotechnol.*, **24**, 1010–1016.

74. Ernst,A., Gfeller,D., Kan,Z., Seshagiri,S., Kim,P.M., Bader,G.D. and Sidhu,S.S. (2010) Coevolution of PDZ domain–ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. BioSyst.*, **6**, 1782–1790.

75. Persikov,A.V., Rowland,E.F., Oakes,B.L., Singh,M. and Noyes,M.B. (2014) Deep sequencing of large library selections allows computational discovery of diverse sets of zinc fingers that bind common targets. *Nucleic Acids Research*, **42**, 1497–1508.

76. Persikov,A.V., Wetzel,J.L., Rowland,E.F., Oakes,B.L., Xu,D.J., Singh,M. and Noyes,M.B. (2015) A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. *Nucleic Acids Research*, **43**, 1965–1984.

77. Packer,M.S. and Liu,D.R. (2015) Methods for the directed evolution of proteins. *Nat. Rev. Genet.*, **16**, 379–394.

78. Reetz,M.T. and Carballeira,J.D. (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat Protoc*, **2**, 891–903.

79. Renata,H., Wang,Z.J. and Arnold,F.H. (2015) Expanding the Enzyme Universe: Accessing Non-Natural Reactions by Mechanism-Guided Directed Evolution. *Angew. Chem. Int. Ed.*, **54**, 3351–3367.

80. Marks,D.S., Hopf,T.A. and Sander,C. (2012) Protein structure prediction from sequence variation. *Nature Biotechnology*, **30**, 1072–1080.

81. Koga,N., Tatsumi-Koga,R., Liu,G., Xiao,R., Acton,T.B., Montelione,G.T. and Baker,D. (2012) Principles for designing ideal protein structures. *Nature*, **491**, 222–227.

82. Correia,B.E., Bates,J.T., Loomis,R.J., Baneyx,G., Carrico,C., Jardine,J.G., Rupert,P., Correnti,C., Kalyuzhniy,O., Vittal,V., *et al.* (2014) Proof of principle for epitope-focused vaccine design. *Nature*, **507**, 201–206.

83. Woolfson,D.N., Bartlett,G.J., Burton,A.J., Heal,J.W., Niitsu,A., Thomson,A.R. and Wood,C.W. (2015) De novo protein design: how do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.*, **33**, 16–26.

84. Dougherty,M.J. and Arnold,F.H. (2009) Directed evolution: new parts and optimized function. *Curr. Opin. Biotechnol.*, **20**, 486–491.

85. Reetz,M.T. (2013) The Importance of Additive and Non-Additive Mutational Effects in Protein Engineering. *Angew. Chem. Int. Ed.*, **52**, 2658–2666.

86. Kitzman,J.O., Starita,L.M., Lo,R.S., Fields,S. and Shendure,J. (2015) Massively parallel single-amino-acid mutagenesis. *Nat Meth*, **12**, 203–.

87. Cohen,J. (2001) How DNA shuffling works. *Science*, **293**, 237–237.

88. Dominy,C.N. and Andrews,D.W. (2003) Site-Directed Mutagenesis by Inverse PCR. In *E coli Plasmid Vectors*. Humana Press, New Jersey, Vol. 235, pp. 209–224.

89. Foo,J.L., Ching,C.B., Chang,M.W. and Leong,S.S.J. (2012) The imminent role of protein engineering in synthetic biology. *Biotechnology Advances*, **30**, 541–549.

90. Yang,H., Li,J., Shin,H.-D., Du,G., Liu,L. and Chen,J. (2014) Molecular engineering of industrial enzymes: recent advances and future prospects. *Appl Microbiol Biotechnol*, **98**, 23–29.

91. Hubbard,B.P., Badran,A.H., Zuris,J.A., Guilinger,J.P., Davis,K.M., Chen,L., Tsai,S.Q., Sander,J.D., Joung,J.K. and Liu,D.R. (2015) Continuous directed evolution of DNA-binding proteins to improve TALEN specificity. *Nat Meth*, 10.1038/nmeth.3515.

92. Esvelt,K.M., Carlson,J.C. and Liu,D.R. (2011) A system for the continuous directed evolution of biomolecules. *Nature*, **472**, 499–503.

93. Kanno,T. and Tozawa,Y. (2010) Protein engineering accelerated by cell-free technology. *Methods Mol. Biol.*, **607**, 85–99.

94. Ellis,T., Wang,X. and Collins,J.J. (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature Biotechnology*, **27**, 465–471.

95. Shekhawat,S.S. and Ghosh,I. (2011) Split-protein systems: beyond binary protein-protein interactions. *Current Opinion in Chemical Biology*, **15**, 789–797.

96. Dodevski,I., Markou,G.C. and Sarkar,C.A. (2015) Conceptual and methodological advances in cell-free directed evolution. *Curr. Opin. Struct. Biol.*, **33**, 1–7.

97. Tawfik,D.S. and Griffiths,A.D. (1998) Man-made cell-like compartments for molecular evolution. *Nature Biotechnology*, **16**, 652–656.

98. Bratkovič,T. (2009) Progress in phage display: evolution of the technique and its applications. *CMLS, Cell. Mol. Life Sci.*, **67**, 749–767.

99. Hanes,J. and Plückthun,A. (1997) In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 4937–4942.

100. Ueda,T., Kanamori,T. and Ohashi,H. (2010) Ribosome display with the PURE technology. *Methods Mol. Biol.*, **607**, 219–225.

101. Packer,M.S. and Liu,D.R. (2015) Methods for the directed evolution of proteins. *Nat. Rev. Genet.*, **16**, 379–394.

102. Kundrotas,P.J., Zhu,Z., Janin,J. and Vakser,I.A. (2012) Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 9438–9441.

103. Biasini,M., Bienert,S., Waterhouse,A., Arnold,K., Studer,G., Schmidt,T., Kiefer,F., Cassarino,T.G., Bertoni,M., Bordoli,L., *et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, **42**, gku340–W258.

104. Meier,A. and Söding,J. (2015) Probabilistic multi-template protein homology modeling. *PLoS Comput Biol*.

105. Hecht,M., Bromberg,Y. and Rost,B. (2013) News from the Protein Mutability Landscape. *Journal of Molecular Biology*, **425**, 3937–3948.

106. Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.

107. Levo,M. and Segal,E. (2014) In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.*, **15**, 453–468.

108. Geertz,M. and Maerkl,S.J. (2010) Experimental strategies for studying transcription factor-DNA binding specificities. *Brief Funct Genomics*, **9**, 362–373.

109. Verdin,E. and Ott,M. (2015) 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond.

110. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Research*, **38**, 7364–7377.

111. Hudson,W.H. and Ortlund,E.A. (2014) The structure, function and evolution of proteins that bind DNA and RNA. *Nat Rev Mol Cell Biol*, **15**, 749–760.

112. Wilson,D., Charoensawan,V., Kummerfeld,S.K. and Teichmann,S.A. (2007) DBD--taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Research*, **36**, D88–D92.

113. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.

114. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., *et al.* (2013) DNA-Binding Specificities of Human Transcription Factors. *Cell*, **152**, 327–339.

115. Weirauch,M.T. and Hughes,T.R. (2011) A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. *Subcell. Biochem.*, **52**, 25–73.

116. Laity,J.H., Lee,B.M. and Wright,P.E. (2001) Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.*, **11**, 39–46.

117. Lam,K.N., van Bakel,H., Cote,A.G., van der Ven,A. and Hughes,T.R. (2011) Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Research*, **39**, 4680–4690.

118. Najafabadi,H.S., Mnaimneh,S., Schmitges,F.W., Garton,M., Lam,K.N., Yang,A., Albu,M., Weirauch,M.T., Radovani,E., Kim,P.M., *et al.* (2015) C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnology*, **33**, 555–562.

119. Wolfe,S.A., Nekludova,L. and Pabo,C.O. (2000) DNA Recognition by Cys 2His 2Zinc Finger Proteins. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 183–212.

120. Stubbs,L., Sun,Y. and Caetano-Anolles,D. (2011) Chapter4: Function and Evolution of C2H2 Zinc Finger Arrays. In Hughes,T.R. (ed), *A Handbook of Transcription Factors*, Subcellular Biochemistry. Springer Netherlands, Dordrecht, Vol. 52, pp. 75–94.

121. Segal,D.J. and Barbas,C.F.,III (2001) Custom DNA-binding proteins come of age: polydactyl zinc-finger proteins. *Curr. Opin. Biotechnol.*, **12**, 632–637.

122. Carlson,C.D., Warren,C.L., Hauschild,K.E., Ozers,M.S., Qadir,N., Bhimsaria,D., Lee,Y., Cerrina,F. and Ansari,A.Z. (2010) Specificity landscapes of DNA binding molecules elucidate biological function. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 4544–4549.

123. Emerson,R.O. and Thomas,J.H. (2009) Adaptive Evolution in Zinc Finger Transcription Factors. *PLoS Genet.*, **5**.

124. Iuchi,S. (2001) Three classes of C2H2 zinc finger proteins. *CMLS, Cell. Mol. Life Sci.*, **58**, 625–635.

125. Shimizu,Y., Soellue,C., Meckler,J.F., Adriaenssens,A., Zykovich,A., Cathomen,T. and Segal,D.J. (2011) Adding Fingers to an Engineered Zinc Finger Nuclease Can Reduce Activity. *Biochemistry*, **50**, 5033–5041.

126. Nomura,W., Masuda,A., Ohba,K., Urabe,A., Ito,N., Ryo,A., Yamamoto,N. and Tamamura,H. (2012) Effects of DNA Binding of the Zinc Finger and Linkers for Domain Fusion on the Catalytic Activity of Sequence-Specific Chimeric Recombinases Determined by a Facile Fluorescent System. *Biochemistry*, **51**, 1510–1517.

127. Theunissen,O., Rudt,F., Guddat,U., Mentzel,H. and Pieler,T. (1992) Rna and Dna-Binding Zinc Fingers in Xenopus Tfiiia. *Cell*, **71**, 679–690.

128. Hall,T.M.T. (2005) Multiple modes of RNA recognition by zinc finger proteins. *Curr. Opin. Struct. Biol.*, **15**, 367–373.

129. Bianchi,E., Folgori,A., Wallace,A., Nicotra,M., Acali,S., Phalipon,A., Barbato,G., Bazzo,R., Cortese,R., Felici,F., *et al.* (1995) A Conformationally Homogeneous Combinatorial Peptide Library. *Journal of Molecular Biology*, **247**, 154–160.

130. Binz,H.K., Amstutz,P. and Plückthun,A. (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nature Biotechnology*, **23**, 1257–1268.

131. Pavletich,N.P. and Pabo,C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, **252**, 809–817.

132. ElrodErickson,M., Rould,M.A., Nekludova,L. and Pabo,C.O. (1996) Zif268 protein-DNA complex refined at 1.6 angstrom: A model system for understanding zinc finger-DNA interactions. *Structure/Folding and Design*, **4**, 1171–1180.

133. Wolfe,S.A., Greisman,H.A., Ramm,E.I. and Pabo,C.O. (1999) Analysis of zinc fingers optimized via phage display: Evaluating the utility of a recognition code. *Journal of Molecular Biology*, **285**, 1917–1934.

134. Pabo,C.O. and Nekludova,L. (2000) Geometric analysis and comparison of protein-DNA interfaces: Why is there no simple code for recognition? *Journal of Molecular Biology*, **301**, 597–624.

135. Elrod-Erickson,M. and Pabo,C.O. (1999) Binding studies with mutants of Zif268. Contribution of individual side chains to binding affinity and specificity in the Zif268 zinc finger-DNA complex. *J. Biol. Chem.*, **274**, 19281–19285.

136. Miller,J.C. and Pabo,C.O. (2001) Rearrangement of side-chains in a zif268 mutant highlights the complexities of zinc finger-DNA recognition. *Journal of Molecular Biology*, **313**, 309–315.

137. Segal,D.J., Dreier,B., Beerli,R.R. and Barbas,C.F. (1999) Toward controlling gene expression at will: Selection and design of zinc finger domains recognizing each of the 5 '-GNN-3 ' DNA target sequences. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 2758–2763.

138. Dreier,B., Segal,D.J. and Barbas,C.F.,III (2000) Insights into the molecular recognition of the 5′-GNN-3′ family of DNA sequences by zinc finger domains. *Journal of Molecular Biology*, **303**, 489–502.

139. Dreier,B., Beerli,R.R., Segal,D.J., Flippin,J.D. and Barbas,C.F. (2001) Development of zinc finger domains for recognition of the 5 '-ANN-3 ' family of DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.*, **276**, 29466–29478.

140. Dreier,B., Fuller,R.P., Segal,D.J., Lund,C.V., Blancafort,P., Huber,A., Koksch,B. and Barbas,C.F. (2005) Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.*, **280**, 35588–35597.

141. Bulyk,M.L., Huang,X.H., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 7158–7163.

142. Liu,J. and Stormo,G.D. (2005) Quantitative analysis of EGR proteins binding to DNA: assessing additivity in both the binding site and the protein. *BMC Bioinformatics*, **6**, 176.

143. Desjarlais,J.R. and Berg,J.M. (1993) Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 2256–2260.

144. Carroll,D., Morton,J.J., Beumer,K.J. and Segal,D.J. (2006) Design, construction and in vitro testing of zinc finger nucleases. *Nat Protoc*, **1**, 1329–1341.

145. Persikov,A.V. and Singh,M. (2013) De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Research*, **42**, 97–108.

146. Fu,F. and Voytas,D.F. (2013) Zinc Finger Database (ZiFDB) v2.0: a comprehensive database of C_2H_2 zinc fingers and engineered zinc finger arrays. *Nucleic Acids Research*, **41**, D452–5.

147. Maeder,M.L., Thibodeau-Beganny,S., Osiak,A., Wright,D.A., Anthony,R.M., Eichtinger,M., Jiang,T., Foley,J.E., Winfrey,R.J., Townsend,J.A., *et al.* (2008) Rapid 'open-source' engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell*, **31**, 294–301.

148. Sander,J.D., Dahlborg,E.J., Goodwin,M.J., Cade,L., Zhang,F., Cifuentes,D., Curtin,S.J., Blackburn,J.S., Thibodeau-Beganny,S., Qi,Y., *et al.* (2011) Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nat Meth*, **8**, 67–69.

149. Wilson,K.A., McEwen,A.E., Pruett-Miller,S.M., Zhang,J., Kildebeck,E.J. and Porteus,M.H. (2013) Expanding the Repertoire of Target Sites for Zinc Finger Nuclease-mediated Genome Modification. *Mol Ther Nucleic Acids*, **2**.

150. Bhakta,M.S. and Segal,D.J. (2010) The generation of zinc finger proteins by modular assembly. *Methods Mol. Biol.*, **649**, 3–30.

151. Gonzalez,B., Schwimmer,L.J., Fuller,R.P., Ye,Y., Asawapornmongkol,L. and Barbas,C.F. (2010) Modular system for the construction of zinc-finger libraries and proteins. *Nat Protoc*, **5**, 791–810.

152. Gupta,A., Christensen,R.G., Rayla,A.L., Lakshmanan,A., Stormo,G.D. and Wolfe,S.A. (2012) An optimized two-finger archive for ZFN-mediated gene targeting. *Nat Meth*, **9**, 588–590.

153. Mandell,J.G. and Barbas,C.F. (2006) Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. *Nucleic Acids Research*, **34**, W516–23.

154. Beerli,R.R. and Barbas,C.F. (2002) Engineering polydactyl zinc-finger transcription factors. *Nature Biotechnology*, **20**, 135–141.

155. Sera,T. (2009) Zinc-finger-based artificial transcription factors and their applications. *Adv. Drug Deliv. Rev.*, **61**, 513–526.

156. Khalil,A.S., Lu,T.K., Bashor,C.J., Ramirez,C.L., Pyenson,N.C., Joung,J.K. and Collins,J.J. (2012) A synthetic biology framework for programming eukaryotic transcription functions. *Cell*, **150**, 647–658.

157. Keung,A.J., Bashor,C.J., Kiriakov,S., Collins,J.J. and Khalil,A.S. (2014) Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. *Cell*, **158**, 110–120.

158. Gaj,T., Guo,J., Kato,Y., Sirk,S.J. and Barbas,C.F.I. (2012) Targeted gene knockout by direct delivery of zinc-finger nuclease proteins. *Nat Meth*, **9**, 805–.

159. Gaj,T., Liu,J., Anderson,K.E., Sirk,S.J. and Barbas,C.F. (2014) Protein delivery using Cys2-His2 zinc-finger domains. *ACS Chem. Biol.*, **9**, 1662–1667.

160. Ghosh,I., Stains,C.I., Ooi,A.T. and Segal,D.J. (2006) Direct detection of double-stranded DNA: molecular methods and applications for DNA diagnostics. *Mol. BioSyst.*, **2**, 551–560.

161. Kim,M.-S., Stybayeva,G., Lee,J.Y., Revzin,A. and Segal,D.J. (2011) A zinc finger protein array for the visual detection of specific DNA sequences for diagnostic applications. *Nucleic Acids Research*, **39**.

162. Aik T Ooi, Cliff I Stains, Indraneel Ghosh,A.David J Segal (2006) Sequence-Enabled Reassembly of β-Lactamase (SEER-LAC): A Sensitive Method for the Detection of Double-Stranded DNA†. *Biochemistry*, **45**, 3620–3625.

163. Wu,J., Kandavelou,K. and Chandrasegaran,S. (2007) Custom-designed zinc finger nucleases: What is next? *CMLS, Cell. Mol. Life Sci.*, **64**, 2933–2944.

164. Cornu,T.I., Thibodeau-Beganny,S., Guhl,E., Alwin,S., Eichtinger,M., Joung,J.K. and Cathomen,T. (2008) DNA-binding Specificity Is a Major Determinant of the Activity and Toxicity of Zinc-finger Nucleases. *Molecular Therapy*, **16**, 352–358.

165. Anand,P., Schug,A. and Wenzel,W. (2013) Structure based design of protein linkers for zinc finger nuclease. *FEBS Lett.*, **587**, 3231–3235.

166. Chen,F., Pruett-Miller,S.M., Huang,Y., Gjoka,M., Duda,K., Taunton,J., Collingwood,T.N., Frodin,M. and Davis,G.D. (2011) High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat Meth*, **8**, 753–U96.

167. Doyon,Y., Vo,T.D., Mendel,M.C., Greenberg,S.G., Wang,J., Xia,D.F., Miller,J.C., Urnov,F.D., Gregory,P.D. and Holmes,M.C. (2011) Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nat Meth*, **8**, 74–U108.

168. Miller,J.C., Holmes,M.C., Wang,J., Guschin,D.Y., Lee,Y.-L., Rupniewski,I., Beausejour,C.M., Waite,A.J., Wang,N.S., Kim,K.A., *et al.* (2007) An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature Biotechnology*, **25**, 778–785.

169. Perez,E.E., Wang,J., Miller,J.C., Jouvenot,Y., Kim,K.A., Liu,O., Wang,N., Lee,G., Bartsevich,V.V., Lee,Y.-L., *et al.* (2008) Establishment of HIV-1 resistance in CD4(+) T cells by genome editing using zinc-finger nucleases. *Nature Biotechnology*, **26**, 808–816.

170. Isalan,M. (2012) Zinc-finger nucleases: how to play two good hands. *Nat Meth*, **9**, 32–34.

171. Guo,J., Gaj,T. and Barbas,C.F.I. (2010) Directed Evolution of an Enhanced and Highly Efficient FokI Cleavage Domain for Zinc Finger Nucleases. *Journal of Molecular Biology*, **400**, 96–107.

172. Townsend,J.A., Wright,D.A., Winfrey,R.J., Fu,F., Maeder,M.L., Joung,J.K. and Voytas,D.F. (2009) High-frequency modification of plant genes using engineered zinc-finger nucleases. *Nature*, **459**, 442–U161.

173. Kim,H.J., Lee,H.J., Kim,H., Cho,S.W. and Kim,J.-S. (2009) Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. *Genome Research*, **19**, 1279–1288.

174. Gaj,T., Gersbach,C.A. and Barbas,C.F. (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, **31**, 397–405.

175. Stormo,G.D. and Zhao,Y. (2007) Putting numbers on the network connections. *Bioessays*, **29**, 717–721.

176. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.

177. Segal,E. and Widom,J. (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev. Genet.*, **10**, 443–456.

178. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.

179. Cai,Y.-H. and Huang,H. (2012) Advances in the study of protein-DNA interaction. *Amino Acids*, **43**, 1141–1146.

180. Smith,A.J.P. and Humphries,S.E. (2009) Characterization of DNA-Binding Proteins Using Multiplexed Competitor EMSA. *Journal of Molecular Biology*, **385**, 714–717.

181. Galas,D.J. and Schmitz,A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*, **5**, 3157–3170.

182. Brenowitz,M., Senear,D.F. and Kingston,R.E. (2001) DNase I footprint analysis of protein-DNA binding. *Curr Protoc Mol Biol*, **Chapter 12**, Unit 12.4.

183. Hampshire,A.J., Rusling,D.A., Broughton-Head,V.J. and Fox,K.R. (2007) Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands. *Methods*, **42**, 128–140.

184. Stoltenburg,R., Reinemann,C. and Strehlitz,B. (2007) SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular Engineering*, **24**, 381–403.

185. Wang,J., Lu,J., Gu,G. and Liu,Y. (2011) In vitro DNA-binding profile of transcription factors: methods and new insights. *J. Endocrinol.*, **210**, 15–27.

186. Oliphant,A.R., Brandl,C.J. and Struhl,K. (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell. Biol.*, **9**, 2944–2949.

187. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J., *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, **20**, 861–873.

188. Zykovich,A., Korf,I. and Segal,D.J. (2009) Bind-n-Seq: high-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing. *Nucleic Acids Research*, **37**, gkp802–e151.

189. Grove,C.A. and Walhout,A.J.M. (2008) Transcription factor functionality and transcription regulatory networks. *Mol. BioSyst.*, **4**, 309–314.

190. Xie,Z., Hu,S., Qian,J., Blackshaw,S. and Zhu,H. (2011) Systematic characterization of protein-DNA interactions. *Cell. Mol. Life Sci.*, **68**, 1657–1668.

191. Kanagawa,T. (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, **96**, 317–323.

192. Kalle,E., Kubista,M. and Rensing,C. (2014) Multi-template polymerase chain reaction. *Biomolecular Detection and Quantification*, **2**, 11–29.

193. Bulyk,M.L., Gentalen,E., Lockhart,D.J. and Church,G.M. (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nature Biotechnology*, **17**, 573–577.

194. Bulyk,M.L. (2006) DNA microarray technologies for measuring protein-DNA interactions. *Curr. Opin. Biotechnol.*, **17**, 422–430.

195. Mukherjee,S., Berger,M.F., Jona,G., Wang,X.S., Muzzey,D., Snyder,M., Young,R.A. and Bulyk,M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.

196. Freckleton,G., Lippman,S.I., Broach,J.R. and Tavazoie,S. (2009) Microarray profiling of phage-display selections for rapid mapping of transcription factor-DNA interactions. *PLoS Genet.*, **5**, e1000449.

197. Hauschild,K.E., Stover,J.S., Boger,D.L. and Ansari,A.Z. (2009) CSI-FID: high throughput label-free detection of DNA binding molecules. *Bioorg. Med. Chem. Lett.*, **19**, 3779–3782.

198. Patwardhan,R.P., Lee,C., Litvin,O., Young,D.L., Pe'er,D. and Shendure,J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, **27**, 1173–1175.

199. Nutiu,R., Friedman,R.C., Luo,S., Khrebtukova,I., Silva,D., Li,R., Zhang,L., Schroth,G.P. and Burge,C.B. (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature Biotechnology*, **29**, 659–U146.

200. Fägerstam,L.G., Frostell-Karlsson,Å., Karlsson,R., Persson,B. and Rönnberg,I. (1992) Biospecific interaction analysis using surface plasmon resonance detection applied to kinetic, binding site and concentration analysis. *Journal of Chromatography A*, **597**, 397–410.

201. Majka,J. and Speck,C. (2007) Analysis of protein–DNA interactions using surface plasmon resonance. *Analytics of Protein–DNA Interactions*, **104**, 13–36.

202. Campbell,C.T. and Kim,G. (2007) SPR microscopy and its applications to high-throughput analyses of biomolecular binding events and their kinetics. *Biomaterials*, **28**, 2380–2392.

203. Rockel,S., Geertz,M. and Maerkl,S.J. (2012) MITOMI: A Microfluidic Platform for In Vitro Characterization of Transcription Factor–DNA Interaction. *Gene Regulatory Networks*, **786**.

204. Maerkl,S.J. and Quake,S.R. (2007) A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science*.

205. Unger,M.A., Chou,H.P., Thorsen,T., Scherer,A. and Quake,S.R. (2000) Monolithic microfabricated valves and pumps by multilayer soft lithography. *Science*, **288**, 113–116.

206. Thorsen,T., Maerkl,S.J. and Quake,S.R. (2002) Microfluidic large-scale integration. *Science*, **298**, 580–584.

207. Duffy,D.C., McDonald,J.C., Schueller,O.J. and Whitesides,G.M. (1998) Rapid Prototyping of Microfluidic Systems in Poly(dimethylsiloxane). *Anal. Chem.*, **70**, 4974–4984.

208. Fordyce,P.M., Gerber,D., Tran,D., Zheng,J., Li,H., DeRisi,J.L. and Quake,S.R. (2010) De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature Biotechnology*, **28**, 970–975.

209. Geertz,M., Shore,D. and Maerkl,S.J. (2012) Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 16540–16545.

210. Rockel,S., Geertz,M., Hens,K., Deplancke,B. and Maerkl,S.J. (2013) iSLIM: a comprehensive approach to mapping and characterizing gene regulatory networks. *Nucleic Acids Research*, **41**, e52.

211. Maerkl,S.J. and Quake,S.R. (2009) Experimental determination of the evolvability of a transcription factor. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 18650–18655.

212. Gerber,D., Maerkl,S.J. and Quake,S.R. (2009) An in vitro microfluidic approach to generating protein-interaction networks. *Nat Meth*, **6**, 71–74.

213. Bates,S.R. and Quake,S.R. (2009) Highly parallel measurements of interaction kinetic constants with a microfabricated optomechanical device. *Appl. Phys. Lett.*, **95**, 073705.

214. Bates,S.R. and Quake,S.R. (2014) Mapping of Protein-Protein Interactions of E. coli RNA Polymerase with Microfluidic Mechanical Trapping. *PLoS ONE*, **9**, e91542.

215. Garcia-Cordero,J.L., Nembrini,C., Stano,A., Hubbell,J.A. and Maerkl,S.J. (2013) A high-throughput nanoimmunoassay chip applied to large-scale vaccine adjuvant screening. *Integrative Biology*, **5**, 650–658.

216. Einav,S., Gerber,D., Bryson,P.D. and Sklan,E.H. (2008) Discovery of a hepatitis C target and its pharmacological inhibitors by microfluidic affinity analysis. *Nature*, **26**, 1019–1027.

217. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S., *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, **31**, 126–134.

218. Xu,Y.Z., Kanagaratham,C., Jancik,S. and Radzioch,D. (2013) Promoter deletion analysis using a dual-luciferase reporter system. *Methods Mol. Biol.*, **977**, 79–93.

219. Shoemaker,D.D., Lashkari,D.A., Morris,D., Mittmann,M. and Davis,R.W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.*, **14**, 450–456.

220. Gertz,J., Siggia,E.D. and Cohen,B.A. (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, **457**, 215–218.

221. Voss,T.C. and Hager,G.L. (2013) Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.*, **15**, 69–81.

222. Jackson,V. (1978) Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. *Cell*, **15**, 945–954.

223. Aparicio,O., Geisberg,J.V., Sekinger,E., Yang,A., Moqtaderi,Z. and Struhl,K. (2005) Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo. *Curr Protoc Mol Biol*, **Chapter 21**, Unit 21.3 Supplement 69.

224. Collas,P. (2010) The Current State of Chromatin Immunoprecipitation. *Mol Biotechnol*, **45**, 87–100.

225. Gilfillan,G.D., Hughes,T., Sheng,Y., Hjorthaug,H.S., Straub,T., Gervin,K., Harris,J.R., Undlien,D.E. and Lyle,R. (2012) Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics*, **13**, 645.

226. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

227. Walhout,A. and Vidal,M. (1999) A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome Research*, **9**, 1128–1134.

228. Xu,D.J. and Noyes,M.B. (2015) Understanding DNA-binding specificity by bacteria hybrid selection. *Brief Funct Genomics*, **14**, 3–16.

229. Rhee,H.S. and Pugh,B.F. (2012) ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol*, **Chapter 21**, Unit 21.24 Supplement 100.

230. Kasinathan,S., Orsi,G.A., Zentner,G.E., Ahmad,K. and Henikoff,S. (2014) High-resolution mapping of transcription factor binding sites on native chromatin. *Nat Meth*, **11**, 203–209.

231. van Steensel,B. and Henikoff,S. (2000) Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nature Biotechnology*, **18**, 424–428.

232. Orian,A. (2006) Chromatin profiling, DamID and the emerging landscape of gene expression. *Curr. Opin. Genet. Dev.*, **16**, 157–164.

233. Meng,X., Brodsky,M.H. and Wolfe,S.A. (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature Biotechnology*, **23**, 988–994.

234. Noyes,M.B. (2012) Analysis of specific protein-DNA interactions by bacterial one-hybrid assay. *Methods Mol. Biol.*, **786**, 79–95.

235. Reece-Hoyes,J.S. and Marian Walhout,A.J. (2012) Yeast one-hybrid assays: A historical and technical perspective. *Methods*, **57**, 441–447.

236. Wilson,T.E., Padgett,K.A., Johnston,M. and Milbrandt,J. (1993) A Genetic Method for Defining Dna-Binding Domains - Application to the Nuclear Receptor Ngfi-B. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 9186–9190.

237. Deplancke,B., Dupuy,D., Vidal,M. and Walhout,A.J.M. (2004) A gateway-compatible yeast one-hybrid system. *Genome Research*, **14**, 2093–2101.

238. Ouwerkerk,P.B.F. and Meijer,A.H. (2001) Yeast One-Hybrid Screening for DNA-Protein Interactions John Wiley & Sons, Inc., Hoboken, NJ, USA.

239. Gupta,A., Christensen,R.G., Bell,H.A., Goodwin,M., Patel,R.Y., Pandey,M., Enuameh,M.S., Rayla,A.L., Zhu,C., Thibodeau-Beganny,S., *et al.* (2014) An improved predictive recognition model for Cys(2)-His(2) zinc finger proteins. *Nucleic Acids Research*, **42**, 4800–4812.

240. Buchner,E. (1997) Alcoholic fermentation without yeast cells. *Cornish-Bowden (1997) pp*.

241. Nirenberg,M.W. and Matthaei,J.H. (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U.S.A.*, **47**, 1588–1602.

242. Hodgman,C.E. and Jewett,M.C. (2012) Cell-free synthetic biology Thinking outside the cell. *Metabolic Engineering*, **14**, 261–269.

243. Carlson,E.D., Gan,R., Hodgman,C.E. and Jewett,M.C. (2012) Cell-free protein synthesis: Applications come of age. *Biotechnology Advances*, **30**, 1185–1194.

244. Kwon,Y.-C. and Jewett,M.C. (2015) High-throughput preparation methods of crude extract for robust cell-free protein synthesis. *Sci. Rep.*, **5**, 8663.

245. Sun,Z.Z., Yeung,E., Hayes,C.A., Noireaux,V. and Murray,R.M. (2013) Linear DNA for Rapid Prototyping of Synthetic Biological Circuits in an Escherichia coli Based TX-TL Cell-Free System. *ACS synthetic biology* **3**, 387–397.

246. Pardee,K., Green,A.A., Ferrante,T., Cameron,D.E., DaleyKeyser,A., Yin,P. and Collins,J.J. (2014) Paper-Based Synthetic Gene Networks. *Cell*, **159**, 940–954.

247. Rosenblum,G. and Cooperman,B.S. (2014) Engine out of the chassis: Cell-free protein synthesis and its uses. *FEBS Lett.*, **588**, 261–268.

248. Shimizu,Y., Kanamori,T. and Ueda,T. (2005) Protein synthesis by pure translation systems. *Methods*, **36**, 299–304.

249. Sun,Z.Z., Hayes,C.A., Shin,J., Caschera,F., Murray,R.M. and Noireaux,V. (2013) Protocols for Implementing an Escherichia coli Based TX-TL Cell-Free Expression System for Synthetic Biology. *JoVE (Journal of Visualized Experiments)*, 10.3791/50762.

250. Kim,D.M. and Swartz,J.R. (1999) Prolonging cell-free protein synthesis with a novel ATP regeneration system. *Biotechnol. Bioeng.*, **66**, 180–188.

251. Caschera,F. and Noireaux,V. (2015) A cost-effective polyphosphate-based metabolism fuels an all E. coli cell-free expression system. *Metabolic Engineering*, **27**, 29–37.

252. Anderson,M.J., Stark,J.C., Hodgman,C.E. and Jewett,M.C. (2015) Energizing eukaryotic cell-free protein synthesis with glucose metabolism. *FEBS Lett.*, **589**, 1723–1727.

253. Stech,M., Quast,R.B., Sachse,R., Schulze,C., Wüstenhagen,D.A. and Kubick,S. (2014) A Continuous-Exchange Cell-Free Protein Synthesis System Based on Extracts from Cultured Insect Cells. *PLOS ONE*, **9**, e96635.

254. Jackson,K., Kanamori,T., Ueda,T. and Fan,Z.H. (2014) Protein synthesis yield increased 72 times in the cell-free PURE system. *Integrative Biology*, **6**, 781–788.

255. Shimizu,Y., Inoue,A., Tomari,Y., Suzuki,T., Yokogawa,T., Nishikawa,K. and Ueda,T. (2001) Cell-free translation reconstituted with purified components. *Nature Biotechnology*, **19**, 751–755.

256. Ohashi,H., Kanamori,T., Shimizu,Y. and Ueda,T. (2010) A Highly Controllable Reconstituted Cell-Free System -a Breakthrough in Protein Synthesis Research. *CPB*, **11**, 267–271.

257. Kazuta,Y., Adachi,J., Matsuura,T., Ono,N., Mori,H. and Yomo,T. (2008) Comprehensive analysis of the effects of Escherichia coli ORFs on protein translation reaction. *Mol. Cell Proteomics*, **7**, 1530–1540.

258. Young,C.L., Britton,Z.T. and Robinson,A.S. (2012) Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. *Biotechnology Journal*, **7**, 620–634.

259. Mureev,S., Kovtun,O., Nguyen,U.T.T. and Alexandrov,K. (2009) Species-independent translational leaders facilitate cell-free expression. *Nature Biotechnology*, **27**, 747–752.

260. Noireaux,V., Bar-Ziv,R. and Libchaber,A. (2003) Principles of cell-free genetic circuit assembly. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 12672–12677.

261. Shin,J. and Noireaux,V. (2010) Efficient cell-free expression with the endogenous E. Coli RNA polymerase and sigma factor 70. *J Biol Eng*, **4**, 8.

262. Shin,J. and Noireaux,V. (2012) An E. coli Cell-Free Expression Toolbox: Application to Synthetic Gene Circuits and Artificial Cells. *ACS Synth Biol*, 10.1021/sb200016s.

263. Young,L. and Dong,Q. (2004) Two-step total gene synthesis method. *Nucleic Acids Research*, **32**, e59–e59.

264. Ehrlich,P. and Doty,P. (1958) The Alkaline Denaturation of Deoxyribose Nucleic Acid. *J. Am. Chem. Soc.*, **80**, 4251–4255.

265. Pakhomov,A.A. and Martynov,V.I. (2008) GFP family: structural insights into spectral tuning. *Chem. Biol.*, **15**, 755–764.

266. Treynor,T.P., Vizcarra,C.L., Nedelcu,D. and Mayo,S.L. (2007) Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 48–53.

267. Grote,A., Hiller,K., Scheer,M., Münch,R., Nörtemann,B., Hempel,D.C. and Jahn,D. (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Research*, **33**, W526–31.

268. Bjourson,A.J. and Cooper,J.E. (1992) Band-stab PCR: a simple technique for the purification of individual PCR products. *Nucleic Acids Research*, **20**, 4675.

269. Wilson,R. (2011) Preparation of single-stranded DNA from PCR products with streptavidin magnetic beads. *Nucleic Acid Ther*, **21**, 437–440.

270. Holmberg,A., Blomstergren,A., Nord,O., Lukacs,M., Lundeberg,J. and Uhlén,M. (2005) The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. *Electrophoresis,* **26**, 501-510.

271. Ramirez,C.L., Foley,J.E., Wright,D.A., Müller-Lerch,F., Rahman,S.H., Cornu,T.I., Winfrey,R.J., Sander,J.D., Fu,F., Townsend,J.A., *et al.* (2008) Unexpected failure rates for modular assembly of engineered zinc fingers. *Nat Meth*, **5**, 374–375.

272. Bae,K.-H., Do Kwon,Y., Shin,H.-C., Hwang,M.-S., Ryu,E.-H., Park,K.-S., Yang,H.-Y., Lee,D.-K., Lee,Y., Park,J., *et al.* (2003) Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nature Biotechnology*, **21**, 275–280.

273. Segal,D.J., Beerli,R.R., Blancafort,P., Dreier,B., Effertz,K., Huber,A., Koksch,B., Lund,C.V., Magnenat,L., Valente,D., *et al.* (2003) Evaluation of a Modular Strategy for the Construction of Novel Polydactyl Zinc Finger DNA-Binding Proteins †. *Biochemistry*, **42**, 2137–2148.

274. Nam,Y., Branch,D.W. and Wheeler,B.C. (2006) Epoxy-silane linking of biomolecules is simple and effective for patterning neuronal cultures. *Biosens Bioelectron*, **22**, 589–597.

275. Garcia-Cordero,J.L. and Maerkl,S.J. (2015) Mechanically Induced Trapping of Molecular Interactions and Its Applications. *J Lab Autom*, 10.1177/2211068215578586.

276. Persikov,A.V., Osada,R. and Singh,M. (2009) Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics,* **25**(1), 22-29.

277. Kim,J.-S. and Pabo,C.O. (1998) Getting a handhold on DNA: Design of poly-zinc finger proteins with femtomolar dissociation constants. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 2812–2817.

278. Siggers,T., Reddy,J., Barron,B. and Bulyk,M.L. (2014) Diversification of Transcription Factor Paralogs via Noncanonical Modularity in C2H2 Zinc Finger DNA Binding. *Mol. Cell*, **55**, 640–648.

279. Garton,M., Najafabadi,H.S., Schmitges,F.W., Radovani,E., Hughes,T.R. and Kim,P.M. (2015) A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity. *Nucleic Acids Research*, 10.1093/nar/gkv919.

280. Beerli,R.R., Segal,D.J., Dreier,B. and Barbas,C.F. (1998) Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14628–14633.

281. Choo,Y. and Isalan,M. (2000) Advances in zinc finger engineering. *Curr. Opin. Struct. Biol.*, **10**(4), 411-416.

282. Main,E.R.G., Phillips,J.J. and Millership,C. (2013) Repeat protein engineering: creating functional nanostructures/biomaterials from modular building blocks. *Biochm. Soc. Trans.*, **41**, 1152–1158.

283. Holliger,P. and Hudson,P.J. (2005) Engineered antibody fragments and the rise of single domains. *Nature Biotechnology*, **23**, 1126–1136.

284. Plückthun,A. (2015) Designed Ankyrin Repeat Proteins (DARPins): Binding Proteins for Research, Diagnostics, and Therapy. *Annu. Rev. Pharmacol. Toxicol.*, **55**, 489–511.

M A T T H E W C . B L A C K B U R N

Avenue Édouard Dapples 36, Lausanne, 1006
Mobile: +41 (0)78.874.0930
Email: blackburn.matthew@gmail.com

Date of birth: 28.06.1985
Nationality: American
Permit B valid until 08.2016
Civil status: single

Education:

Doctor of Philosophy Candidate in Biotechnology and Bioengineering **2010 – Present**
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, CH (anticipated December 2015)

Master of Science in Chemical Engineering **2008 – 2010**
Massachusetts Institute of Technology (MIT), Cambridge, MA, USA – GPA: 4.3/5.0

Bachelor of Science in Chemical Engineering, *summa cum laude* **2003 – 2008**
University of Florida (UF), Gainesville, FL, USA – GPA: 3.93/4.0

Professional Experience

EPFL, Protein-DNA Interactions Research and Synthetic Gene Assembly **09.2010 – Present**
Institute of Biotechnology & Bioengineering, Laboratory for Biological Network Characterization; EPFL, Lausanne
- Project Title: Integrating Gene Synthesis and Microfluidics for Rapid Protein Engineering
- Method development and process optimization for solid-phase assembly of genes from short synthetic DNA fragments and integration of this technique with a high-throughput microfluidic device for quick and economical expression and bioanalytical characterization of designed proteins.

MIT, Environmental & Human Microbiome Research **02.2009 – 08.2010**
Department of Biological Engineering; MIT, Cambridge, MA; USA
- Project Title: Targeted High-Throughput Sequencing of Human Gut and Lake Microbiomes
- Development and optimization of technique for DNA extraction from environmental and human samples for the creation of genomic libraries for bacterial diversity studies using next-generation DNA sequencing platforms (metagenomics).

Kimberly-Clark Corporation, internships in various business units; Roswell, GA and Neenah, WI; USA **2006 – 2008**
- **Global Nonwovens (06 – 08.2008)**: Performed process optimization trials on elastic melt spun composite pilot line and characterized physical properties (tensile testing, hydrostatic head testing) of new diaper ear materials in preparation for a consumer use test.

- **Skin Care & Infection Control (06 – 08.2007)**: Conducted market/patent research on applications of probiotic bacteria in consumer goods, evaluated the release of chemical actives from polymer films, and explored the use of dyes as biological sensors.

- **Healthcare (09 – 12.2006)**: Researched interaction of InteguSeal* Microbial Sealant with preoperative skin preparation products and created prototypes for improved variations of the product (i.e. color change).

- **Partnership Products (01 – 05.2006)**: Designed and constructed an electrospinning apparatus to produce polymeric nanofibers for air filtration research and characterized prototype nanofiber webs produced from different polymers for filtration efficiency.

Undergraduate Research Experience

Department of Materials Science Engineering, Particle Engineering Research Center; UF, Gainesville, FL; USA

Bacterial Biofilm Adhesion Research **09.2007 – 06.2008**
- Project Title: Response of *S. aureus* Biofilm Formation to Fibronectin Pre-Conditioned, Engineered Microtopography
- Produced silicone elastomer surfaces incubated with fibronectin and evaluated the extent of bacterial biofilm growth on smooth and topographically modified (Sharklet™) samples for biomedical implant applications.

Protein Adhesion Research **01 – 06.2007**
- Project Title: Evaluation of Engineered Topographies for Bacterial Adhesion Resistance: Effect of Protein Adsorption on Substrates
- Produced silicon and silicone elastomer surfaces coated with lyophilized fibronectin and characterized the effect of protein adsorption on microfabricated topography using SEM analysis.

Cell Adhesion Research **01 – 12.2005**
- Project Title: Evaluation of Engineered Topographies for Bacterial Adhesion
- Produced and evaluated silicone elastomer surfaces for investigating marine and biomedical fouling; studied effects of nanotopography on porcine vascular endothelial cells through contact guidance.

Department of Environmental Engineering Sciences, Particle Engineering Research Center; UF, Gainesville, FL; USA

Bioaerosol Research **09 – 12.2004**
- Project Title: Evaluation of an Iodinated Resin Filter for Capture Efficiency and Disinfection
- Prepared and modified a bioaerosol testing apparatus for experimental runs, culture of microorganisms and data collection.

Capabilities

Languages:	**English**	**mother tongue**
	French	**Level B2**
	Spanish	**Level A1**

Laboratory Skills:
Mammalian and bacterial cell culture, PCR/real-time PCR, size selection of DNA fragments using solid phase reversible immobilization, Solexa sequencing library preparation, DNA isolation and purification, BioAnalyzer operation and data analysis, emulsion PCR, microfluidic device design and operation, photolithography and fabrication in PDMS, protein engineering, microarray printing

Computer Software:
MATLAB, CLC DNA Workbench, COMSOL Multiphysics, Aspen, GaussView, Clewin, ImageJ, IgorPro, Prism, MS Office (Excel, Powerpoint, Word), Adobe Illustrator, iWork (Pages, Keynote)

Personal Skills:

Project management: coordination of experiments and resources using diverse techniques from molecular biology, engineering and microfabrication (multidisciplinary) following programmed deadlines and objectives, while working within a collaborative laboratory or industrial environment

Communication: presentation of project outcomes at conferences (posters and talks), summarization of results in scientific reports

Team leadership: assisted and managed small teams of undergraduate student research projects (iGEM 2011 and 2012); trained and mentored master's students and visiting summer student research projects; key event organizer during visits of prospective doctoral students for the EPFL 'Hiring Days' (winter and summer 2012)

Team-working: substantial contributor to several multidisciplinary research projects in academia and product development projects in industry

Advanced Bacterial Genetics Course (06.2009), Cold Spring Harbor Laboratory (CSHL); Cold Spring Harbor, NY; USA
A prestigious, 3-week laboratory and seminar series offered by CSHL since 1976. Only ~15 applicants are accepted each year.

Publications

Blackburn MC, Petrova E, Correia BE, Maerkl SJ. *Integrating gene synthesis and microfluidic protein analysis for rapid protein engineering*. (in review; submitted August 2015)

David LA, Materna AC, Friedman J, Campos-Baptista MI, **Blackburn MC**, Perrota A, Erdman SE, Alm EJ. *Host lifestyle affects human microbiota on daily timescales*. Genome Biology, Vol 15, (2014).

Rodrigue S, Materna AC, Timberlake SC, **Blackburn MC**, Malmstrom RR, Alm EJ, Chisholm SW. *Unlocking Short Read Sequencing for Metagenomics*, PLoS ONE, Vol 5(7), (2010).

Ratnesar-Shumate S, Wu C-Y, Wander J, Lundgren D, Farrah S, Lee J-H, Wanakule P, **Blackburn M,** Lan M-F. *Evaluation of Physical Capture Efficiency and Disinfection Capability of an Iodinated Biocidal Filter Medium*. *Aerosol and Air Quality Research,* Vol 8(1), (2008).

Hobbies

Sports: road cycling, hiking, indoor rock climbing, alpine skiing, alpine touring

Music: cello, piano