April 2012

Characterizing the DNA-Binding Site Specificities of Cis2His2 Zinc Fingers

Heather Anne Bell
Worcester Polytechnic Institute

MQP-ID-DH-UM1

CHARACTERIZING THE DNA-BINDING SITE SPECIFICITIES OF CIS_2HIS_2 ZINC FINGERS

A Major Qualifying Project Report

Submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degrees of Bachelor of Science

in

Biochemistry

and

Biology and Biotechnology

by

Heather Bell

April 26, 2012

APPROVED:

Scot Wolfe, PhD Gene Function and Exp. UMass Medical School MAJOR ADVISOR	Destin Heilman, PhD Biochemistry WPI Project Advisor	David Adams, PhD Biology and Biotech WPI Project Advisor

ABSTRACT

The ability to modularly assemble Zinc Finger Proteins (ZFPs) as well as the wide variety of DNA sequences they can recognize, make ZFPs an ideal framework to design novel DNA-binding proteins. However, due to the complexity of the interactions between residues in the ZF recognition helix and the DNA-binding site there is currently no comprehensive recognition code that would allow for the accurate prediction of the DNA ZFP binding motifs or the design of novel ZFPs for a desired target site. Through the analysis of the DNA-binding site specificities of 98 ZFP clones, determined through a bacterial one-hybrid selection system, a predictive model was created that can accurately predict the binding site motifs of novel ZFPs.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank Dr. Scot Wolfe and the University of Massachusetts Medical School for sponsoring this project and providing guidance along the way. I would also like to thank Ankit Gupta for his guidance in carrying out this product as well as The Joung Lab and The Stormo Lab for their collaboration with the project. Lastly, I would like to thank Dr. Dave Adams and Dr. Destin Heilman, my WPI major advisors, for advising this project and helping to edit the final report.

PROJECT PURPOSE

The purpose of this project was to create a comprehensive ZFP predicative model that could be used to both accurately predict the DNA-binding sites of ZFPs, as well as to aid in designing ZFPs for a desired target site. This predictive model was developed based upon Bacterial One-Hybrid (B1H) binding site selections on 116-engineered ZFP clones based upon the structural framework of Zif268 and was created in collaboration with the laboratories of Keith Joung and Gary Stormo.

BACKGROUND

THE DISCOVERY OF ZINC FINGER PROTEINS

Through the work of Robert Roeder and Donald Brown on the 5S RNA genes of *Xenopus laevis*, it was discovered that the binding of a 40-kDa protein factor, called transcription factor IIIA (TFIIIA), is required for the initiation of transcription. Through further experiments Roeder and Brown determined that TFIIIA interacts with a 50-nucleotide long region within the gene, called the internal control region, resulting in TFIIIA being the first eukaryotic transcription factor ever described. In 1982, further studies of TFIIIA by Miller, Roeder, and Brown showed that the transcription factor contained a high concentration of zinc and could be broken down into nine tandemly repeating units of 30-amino acids in length with each unit containing a similar pattern of cysteines and histidines. The remarkable repeating motif within TFIIIA was later termed a zinc finger (ZF). The protein was given this name due to the presence of the zinc ion and the manner with which the zinc finger gripped the DNA (Klug, 2010).

The ability to modularly assemble Zinc Finger Proteins (ZFPs) as well as the wide variety of DNA sequences they can recognize, make ZFPs an ideal framework to design novel DNA-binding proteins. Of the 30 amino acids in each ZF, 25 fold around a Zn ion to form an independent structural domain, or "the finger", and the remaining 5 amino acids serve a linker between consecutive fingers. In later studies by Neuhaus, it was shown that this linker is extremely flexible and can vary depending on the organism. The discovery of the flexibility of the linker also showed that tandemly linked ZFs are structurally independent of one another and thus opened the door to a whole new world of DNA recognition and DNA binding proteins (Klug, 2010).

Since the discovery of TFIIIA, three major classes of ZFs have been defined. The first class is characterized by the presence of a Cys_6-Zn cluster motif and is found in metabolic regulators of Fungi. The second class consists of Cys_2Cys_2 (or Cys_4) ZFs with a conserved Zn-binding consensus of Cys-X_2-Cys-X_{13}-Cys-X_2-Cys and is found mainly in nuclear steroid or hormone receptors. The third class contains the most common ZF, the Cys_2His_2 ZF or the "classical" ZF domain, and can be found in a large number of regulatory proteins spanning almost all branches of the evolutionary tree (Papworth et al, 2006).

DNA BINDING OF CYS_2HIS_2 ZINC FINGER PROTEINS

Cys_2His_2 Zinc Fingers are the most common DNA-binding domain in metazoan genomes and are naturally occurring in animals, plants, and fungi. They are the best understood out of the three major classes of ZFs and are therefore the most widely used for designing proteins with novel DNA-binding specificities (Wolfe et al., 2000). Cys_2His_2 Zinc Fingers contain a conserved Cys-Cys … His-His pattern with the zinc ion tetrahedrally coordinated between the two cysteines at one end of the β-strands and the two histidines in the C-terminal portion of the α-helix to stabilize its folds (figure 1). In addition to the conserved Cys-Cys and His-His pattern, each ZF also contains three conserved large hydrophobic residues: typically Tyr6, Phe17, and Leu23, which help to stabilize the finger structure. Therefore, the conserved sequence of Cys_2His_2 ZFs can be written as (Tyr, Phe)-X-Cys-

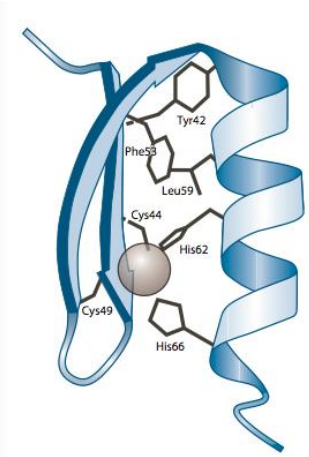


Figure 1: The structure of a Zinc Finger from a two-dimensional NMR study of two-finger peptide in solution (Klug, 2010).

X_{2-5}-Cys-X_3-(Tyr, Phe)-X_5-Leu-X_2-His-X_{3-5}-His, with X representing any amino acid (Klug, 2010). Additionally, ZFs can be linked tandemly in a linear, polar fashion to recognize DNA (or RNA) sequences of different lengths as well as proteins (Wolfe et al., 2000). However, while each finger domain has a similar structural framework, it can achieve chemical distinctiveness through variations in the 4 key amino acid positions used in DNA-binding (Klug, 2010).

In 1991, Pavletich and Pabo conducted an in depth study of the crystal structure of a DNA oligonucleotide bound to a three-finger DNA-binding domain of the mouse transcription factor Zif268. From this study, Pavletich and Pabo showed that the primary sequence-specific contacts are made by the α-helix, which binds in the major groove of the DNA. This binding occurs through specific hydrogen-bond or hydrophobic interactions from amino acids in the recognition helix at positions -1, 3, and 6 with the three positions contacting three consecutive basepairs on one strand of the DNA. Later studies showed that the amino acid in the recognition helix at position 2 also plays an important role in DNA binding, contacting a single basepair on the opposite DNA strand as positions -1, 3, and 6. These studies also demonstrated the importance of cross-strand interaction between the α-helices of tandemly linked ZFs due to the fact that the complement of the basepair contacted by position 2 of the first finger is contacted by position 6 of the second finger and so on (figure 2) (Klug, 2010).

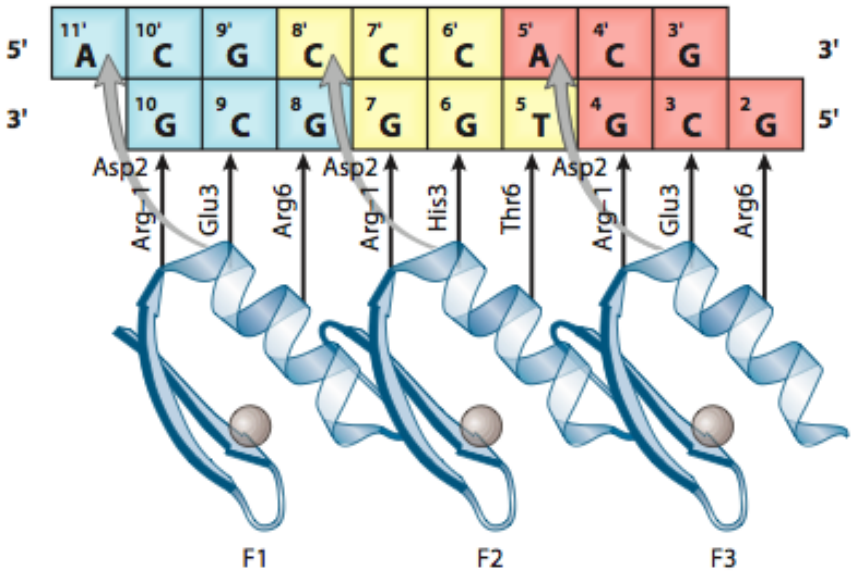


Figure 2: Schematic of the DNA-binding of the mouse transcription factor, Zif268. Contacts with the DNA are made from positions -1, 3, and 6 of the recognition helix on one strand of the DNA with the residue at position 2 contacting the complement strand. Positions -1, 3, and 6 contact three consecutive basepairs while position two contacts the complement of the basepair contacted by position 6 of the next finger. The three finger module binds with position 6 of finger three contacting the 5' end of the DNA and position -1 of finger 1 contacting the 3' end of the DNA. (Image taken from Klug, 2010).

The most common linker arrangement found in ZFs contains five residues between the final histidine of one finger and the first conserved aromatic of the next. Roughly half of the known ZFs contain a linker with the consensus sequence TGEKP. This particular linker is flexible in the free ZFP but becomes rigid upon binding of the ZFP to DNA. The docking of adjacent fingers in a ZFP are stabilized further by a contact involving the side chain of the residue at position 9 of the preceding finger's helix, along with the backbone carbonyl or side chain at position -2 of the subsequent finger. This contact using position 9 has been observed to correlate with the use of a canonical linker such as TGEKP. Due to this interfinger contact at position 9 and the highly conserved TGEKP linger, it can be implied that the interfinger organization is important in DNA recognition as well (Wolfe et al., 2000).

The mode of DNA recognition can be considered a one-to-one interaction between individual amino acids from the recognition helix to individual DNA bases (Wolfe et al., 2000). This combined with the fact that ZFs function as independent modules, allows for the design of ZFPs that will recognize longer DNA sequences through combining several ZFs with different triplet specificities. It is for this reason that zinc finger motifs are an ideal building block for the de novo design of proteins that will recognize any given DNA sequence (Klug, 2010).

ZIF268: THE MODEL Cys_2His_2 ZINC FINGER

The crystal structure of Zif268 discovered by Pavletich and Pabo, has served as the prototype for understanding DNA recognition by Cys_2His_2 Zinc Fingers (Klug, 2010). Zif268, a three-fingered protein, contacts the major groove of the DNA with the α-helix portion of each finger. The binding of each successive finger causes the protein to wrap around the DNA helix with each finger docking the DNA in the same manner. Neighboring fingers are located three basepairs apart but due to a helical motion are shifted, superimposing neighboring fingers by one basepair. The three fingers are oriented so that position 6 of third finger contacts the 5' end of the DNA while position -1 of the first finger contacts the 3' end of the DNA strand (figure 2). The DNA demonstrates a similar conformation to that of B-form DNA with the exception that the major groove is both wider and deeper than normal. This change in the major groove has been shown to be a common feature in the structures of many zinc-finger-DNA complexes (Wolfe et al., 2000).

As mentioned above, the base contacts of Zif268 are made by amino acids in the N-terminal portion of each of the recognition helices (positions, -1, 2, 3, and 6). Each helix docks with the DNA at an angle of 45° in the major groove, with each α-helix at an angle of 45° to the double helix of the DNA. This results in the amino acids of each finger at position -1 (the amino acid immediately before the α-helix), 3 and 6 being well aligned to make basepair contacts with the primary DNA strand and the amino acid at position 2 being well aligned to make contacts with a basepair on the complementary DNA strand (figure 3). The basepair coordinated by position 2 sits just outside the triplet recognized by positions -1, 3, and 6, overlapping with its complement coordinated by position 6 of the consecutive finger (Wolfe et al. , 2000).

Figure 3: Secondary Structure of the mouse transcription factor Zif268, bound to DNA.

All three of Zif268's fingers contain the same residues at positions -1 and 2 (Arg and Asp). These two residues make coordinated DNA contacts with the arginine at position -1 making a pair of hydrogen bonds to guanine at the 3' position of the primary DNA strand for each finger, and the aspartate at position 2 making two hydrogen bonds to the guanidinium group of arginine helping to stabilize the interaction of position -1. The remaining basepair contacts of Zif268 are mediated by positions 3, and 6 of the α-helix. The recognition helices for Zif268 can be seen bolded in the sequences below:

Finger 1: PYA*C*PVES*C*DRRFSRSDELTR*H*IRI*H*TGQK

Finger 2: PFQ*C*RI--*C*MRNFSRSDHLTT*H*IRT*H*TGEK

Finger 3: PFA*C*DI--*C*GRKFARSDERKR*H*TKI*H*LRQK

The residues italicized and underlined in the sequences above indicate the conserved cysteine and histidine residues while the residues in red indicate the conserved hydrophobic residues (Wolfe et al., 2000).

BACTERIAL ONE-HYBRID SELECTION SYSTEM

Many different selection strategies have been developed to identify individual fingers that posses particular DNA-binding site specificities from a randomized library. Some existing methods include Systematic Evolution of Ligands through Exponential enrichment (SELEX), protein-binding microarrays, ELISA, and yeast two-hybrid reporter assays (Meng et al., 2005). However, many of these methods are not ideal for high-throughput applications and often require specialized reagents and equipment. One method that allows for the rapid characterization of the DNA-binding specificities of Cys_2His_2 Zinc Fingers, which requires only basic molecular biology expertise to perform, is the bacterial one-hybrid (B1H) system (figure 4) (Meng et al., 2005).

In the B1H system, selective markers *HIS3* and *URA3* from yeast-based interaction trap systems are transferred into corresponding bacterial based trap systems. The bacterial based trap system can then be used to search combinatorial libraries (prey) for members that interact with a desired ZFP target (bait). The transferred *URA3* reporter allows for self-activating sequences to be removed from the library through counter-selection, while the *HIS3* reporter allows sequences that are recognized by the ZFP to be isolated from the library due to transcriptional activation mediated by the fusion of the ZFP module to the ω-subunit of RNA polymerase (Meng and Wolfe, 2006).

The B1H system used here contains three key components: a 1352 ω-UV2 ZFP plasmid (bait- transcription factor vector), a pH3U3 library plasmid (prey- reporter vector), and a bacterial selection stain (USOΔ*hisB*ΔpyrFΔrpoZ). The ZFP is expressed as a carboxy-terminal fusion to the ω-subunit of RNA polymerase, which is controlled by a mutant *lpp/lacUV5* promoter (UV2), induced by IPTG in NM medium (Meng and Wolfe, 2006). The reporter vector contains the NotI and EcoRI restriction sites for cloning a library of randomized nucleotides upstream of a weak *lac* promoter that drives the expression of the *HIS3* and *URA3* markers. If the ZFP recognizes a DNA sequence within the library reporter vector, it will recruit RNA polymerase by means of its fusion with the ω-subunit to the promoter. This will activate transcription of the reporter genes (Meng and Wolfe, 2006). Bacteria that then contain the activated reporters can be isolated from the general population by growing the selections on minimal media that lacks, histidine and contains 3-amino-triazole (3-AT), a competitive inhibitor of *HIS3*. The stringency of the positive selections are thus controlled by the concentration of 3-AT and IPTG in the medium and can be used to determine the quality of the binding sites isolated. Another advantage to the B1H system over other selection methods is once the binding-site library has been constructed and purified of self-activating sequences, only a single round of selections is required to isolate the set of ZFP binding sites (Meng and Wolfe, 2006).

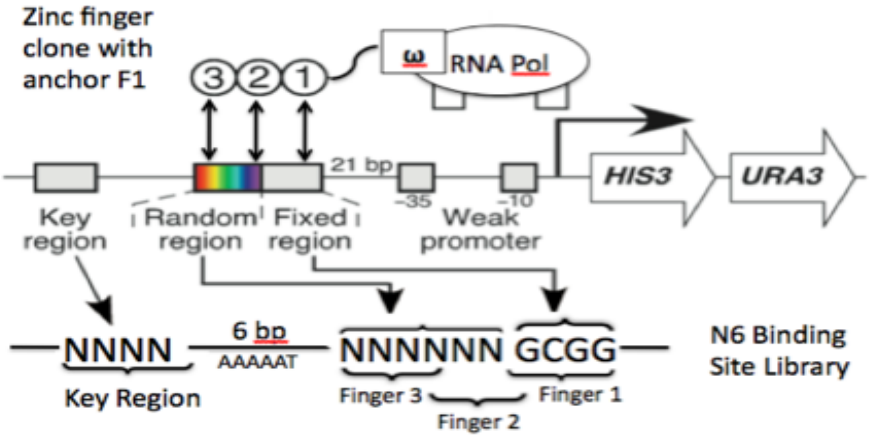


Figure 4: Schematic of the B1H Binding Site Selection System and the N6 library employed in the selections of the 98 KJ ZFP clones.

After obtaining positive colonies off of minimal media plates, the binding sites of these selections can be analyzed via PCR with specified primers for each particular plasmid. Further analysis of the ZFP's DNA-binding site specificities can be done through sequencing and data analysis.

APPLICATIONS OF ENGINEERED ZIC FINGER PROTEINS

Zinc finger proteins can be engineered with a wide variety of effector domains fused to polyzinc finer peptides, which can be designed to recognize any DNA sequence with high affinity and specificity. Therefore it is not surprising that engineered zinc finger proteins have an increasing use in research and medicine from both a biochemical and gene therapy standpoint. Perhaps the most famous case of ZFPs used in medicine was the inhibition of human immunodeficiency virus (HIV) expression using a ZFP that targeted the HIV promoter long terminal repeat (Klug, 2010).

Additionally, attaching addition domains for activation, repression, or enzymatic activity, to a ZFP allows for the protein to carry out a desired function in a site-specific manner. The development of a recognition code would allow not only for the design of

novel DNA-binding proteins, but would also provide an accurate prediction of the DNA-binding site specificities of already known ZFPs furthering the advancement of engineered ZFPs (Wolfe et al., 2000).

ZINC FINGER PROTEIN DESIGN USED IN THIS PROJECT

There is currently no comprehensive recognition code that would allow for the accurate prediction of the DNA binding motifs of naturally-occurring ZFPs, or the design of a ZFP for a desired target site. The aim of this project was to bridge this gap and create a predictive model based upon the analysis of the DNA-binding site specificities determined for 98 ZFP clones.

The zinc fingers used in this project were created based on adaptations of Zif268 and contain 3 fingers that make tandem contacts along the DNA. Each finger consists of a conserved β-β-α structure with the amino acids on the surface of the α-helix contacting bases in the major groove of the DNA at positions -1,2,3 and 6. The first finger of each clone is considered the 'anchor finger' and has a recognition helix RSDELTR (Zif268 F1 recognition helix). The linker between each finger is the conserved TGEKP linker. In order to analyze the binding-site specifies of the ZFP clones, the bacterial one-hybrid system (described above) was used along with Sanger and Illumina sequencing. A predictive model was created in collaboration with the laboratory of Gary Stormo, using the analyzed data. The predictive model was next tested using naturally occurring Human Zinc Finger Proteins identified using a SCAN domain.

METHODS

Bacterial One-Hybrid Selections

Approximately 2 μg of the ZFP bait plasmid (1342 omega-UV2 ZFP) and 50 ng of the N6 library reporter plasmid were electroporated in 60 μl of the selection strain. The two plasmids were first mixed with the selection strain cells on ice then transferred to pre-chilled 1 mm cuvettes. The cell and plasmids suspension was electroporated at 1.75V and immediately resuspended in 5 ml of SOC then recovered while rotating for 50 minutes at 37°C. The cells were then pelleted by centrifugation at 3000 rpm for 10 minutes and resuspended in 5 ml NM medium containing 200 μM uracil and 0.1% histidine. The cells were recovered at 37°C for 45 minutes then pelleted by centrifugation at 3000 rpm for 10 minutes. The cells were washed 4 times with NM medium – histidine + uracil by sequential pelleting and resuspension in 1 ml NM medium with a final suspension in 600 μl. 20 μl of the final resuspension was titrated by 10-fold serial dilutions on rich media 2xYT + 25 μg/ml Kanamycin, and 100 μg/ml Carbenicillin plates to determine the total number of transformants. The titration plates were grown for 14-16 hours at 37°C and cell counts were determined. 1×10^6 cells containing the ZFP bait and N6 library reporter plasmids were plated on selective NM minimal media plates (100 x 15 mm) containing 50 μM IPTG and 1 or 2 mM 3-AT as well as uracil (all clones were plated on both 1 and 2 mM 3-AT) and grown at 37°C for 20-30 hours. The surviving colonies were pooled and the binding site plasmids were isolated for analysis using a QIAGEN mini prep kit. The binding site region was then PCR amplified using HU100 and OK181 (50 μM each) primers and sequenced via Sanger DNA sequencing to

rapidly obtain binding site profiles for each ZF module. The primer sequences for HU100 and OK181 can be seen below:

HU100 5'-GAAATATGTATCCGCTCATGAC-3'

OK181 5'-CCAGAGCATGTATCATATGGTCCAGAAACCC-3'

Electro-competent USO selection strain cells

The *E. coli* selection strain used contained a deletion in the *hisB, pyrF,* and *rpoZ* genes as well as an F' episome encoding the *lacI* repressor (Meng, et. al., 2005; Noyes, et al., 2008).

N6 library

A randomized 6 basepair library as described in Christensen, et al. was used in selections with the KJ Zinc Finger Proteins. The 6 bp library was constructed in the pH3U3 plasmid (Meng et al., 2005) and was introduced between EcoRI and NotI sites with the following oligonucleotide sequence: 5' GCGGCCACTGGGCAGCTG GCCANNNNAAAAATNNNNNNGCGGTACCTAGG TTCTTCGAATTC 3'. The 6 bp library contained two different randomized regions, a 6bp region that corresponded to the four bases of DNA contacted by each of the three Zinc Fingers and a 4bp randomized region that served as an internal control to identify sequences that may have been enriched in the selections or preparation for sequencing due to jackpot effects. The N6 library was counterselected with 5-FOA to remove self-activating sequences as described in Noyes, et al.

N28 library

A randomized 28 basepair library as described in Noyes et al. was used in selections with the Human Zinc Finger Proteins. The 28 bp library was constructed in the pH3U3

plasmid (Meng et al., 2005) and was introduced between Not1 and EcoRI sites with the following oligonucleotide sequence:

5'GGCGCGAATTCGNNNNNNNNNNNNNNNNNNNNNNNNNNGCGGCCGCA AGGTAGCTGATTCCGTTCTCGC-3' (Noyes, et. al, 2008). The 28 bp library placed the randomized region 7 bp away from the 5' end of the -35 box of the weak promoter and controlled expression of the *HIS3/URA3* reporter genes (figure 4). Counterselections were performed on the 28 bp library in order to remove self-activating sequences as described in Noyes, et. al.

Plasmids

Library Reporter Plasmid: The N28 and N6 libraries were constructed in a pH3U3 plasmid that contained kanamycin resistance as well as *HIS3* and *URA3* genes (each with independent Shine-Dalgarno sequences) under the control of a weak *lac* promoter (Meng, et al., 2005). The plasmid also contained a multiple cloning site for inserting the randomized DNA sequences (prey) upstream of the *HIS3/URA3* promoter, a phage f1 origin, and a pSC101 origin of replication that limited the plasmid copy number to roughly ten per cell (Meng et al., 2005).

ZFP Bait Plasmid: A 'GCG' binding anchor finger with the recognition helix RSDTLAR, was fused at the N-terminus to a 2F-module (from the laboratory of Keith Joung) via overlapping PCR. The newly fused 3F ZFP was then cloned into a 1352-omega-UV2-kpn1-bamH1-stop vector containing ampicillin resistance between the Acc65I and BamHI sites for expression as an omega fusion (Gupta, et. al, 2012). A 1352 UV2 omega-Zif268 plasmid was used as a positive control for selection experiments while a 1352 UV2 omega-odd^{ori} plasmid was used as the negative control. The 1352

UV2 omega-odd^{ori} served as the negative control due to the fact that this plasmid only contains the omega subunit and does not contain any transcription factor so will not bind to the library plasmid.

Illumina Sequencing (solexa)

For quantitative modeling, the binding site pools for the ZFP modules were barcoded according to and sequenced via Illumina sequencing. The binding site specificities were then modeled from this data using both W log-odds and GRaMS methods as described in Gupta, et. al.

HZFP Determination

Eighteen three finger ZFPs were chosen based on the presence of interesting combinations of recognition sequences within the ZFPs, from those identified using a SMART database search for Cys_2His_2 ZFPs that contain SCAN domains in the human genome. Once chosen, the ZFPs from these genes were isolated via PCR and cloned into the 1352 omega-UV2 vector described above between the Acc65I and BamHI sites for expression as an omega fusion. The BIH selections for these 18 human zinc finger proteins (HZFPs) were then performed as described below.

HZFP Selections

Approximately 2 μg of the ZFP bait plasmid (1342 omega-UV2 ZFP) and 50 ng of the N28 library reporter plasmid were electroporated in 60 μl of the selection strain. The two plasmids were first mixed with the selection strain cells on ice then transferred to pre-chilled 1 mm cuvettes. The cell and plasmids suspension was electroporated at 1.75V and immediately resuspended in 5 ml of SOC then recovered while rotating for 50 minutes at 37°C. The cells were then pelleted by centrifugation at 3000 rpm for 10

minutes and resuspended in 5 ml NM medium containing 200 μM uracil and 0.1% histidine. The cells were recovered at 37°C for 45 minutes then pelleted by centrifugation at 3000 rpm for 10 minutes. The cells were washed 4 times with NM medium – histidine –uracil by sequential pelleting and resuspension in 1 ml NM medium with a final suspension in 600 μl. 20 μl of the final resuspension was titrated by 10-fold serial dilutions on rich media 2xYT + 25 μg/ml Kanamycin, and 100 μg/ml Carbenicillin plates to determine the total number of transformants. The titration plates were grown for 14-16 hours at 37°C and cell counts were determined. 2×10^7 cells containing the HZFP bait and N28 library reporter plasmids were plated on selective NM minimal media plates (150 mm diameter) containing 50 μM IPTG and 2 or 5 mM 3-AT (all clones were plated on both 2 and 5 mM 3-AT) and grown at 37°C for 40-48 hours. Sixteen colonies from the 5mM 3-AT plates were chosen for colony PCR with the binding site region PCR amplified using HU100 and OK181 (50 μM each) primers and sequenced via Sanger DNA sequencing. The chromatogram from each sequence read was inspected for quality and accuracy before using the MEME motif discovery tool to analyze the sequences between the Not1 and EcoRI sites of the N28 bp library. The settings used in MEME were such that on zero or one motif per sequence could be discovered, a motif could be discovered on either strand of the DNA and the search with was set from 6 to 30 bases (Noyes et al, 2008).

RESULTS

There is currently no comprehensive recognition code that would allow for the accurate prediction of the DNA binding motifs of existing ZFPs, or the design of a ZFP for a desired target site. The purpose of this project was to bridge this gap through creating a comprehensive ZFP predicative model that could be used to both accurately predict the DNA-binding sites of ZFPs, as well as to aid in designing ZFPs for a desired target site. This predictive model was created based upon the analysis of the DNA-binding site specificities, determined through a bacterial one-hybrid selection system, for 98 ZFP clones in collaboration with the laboratories of Keith Joung and Gary Stormo.

The zinc fingers used in this project were created based on adaptations of Zif268 and consist of 3 fingers that make tandem contacts along the DNA; two ZFs with varied recognition helix sequences and an anchor finger with a conserved recognition helix sequence. Each finger consists of the conserved β-β-α structure with the amino acids on the surface of the α-helix contacting bases in the major groove of the DNA at positions -1,2,3 and 6. The first finger of each clone is considered the 'anchor finger' and has a recognition helix RSDELTR (Zif268 F1 recognition helix). The linker between each finger was the conserved TGEKP linker. (A complete list of each ZFP clone, the recognition helix sequence, and the expected and actual DNA-binding sites can be found in supplemental 1.)

In order to analyze the binding-site specifies of the 98 ZFP clones, the bacterial one-hybrid system was used. A 1352 ω-UV2 ZFP bait plasmid that contained the Keith Joung (KJ) ZFP fused to the ω-subunit of RNA polymerase, a pH3U3 library reporter plasmid containing an N6 library and an electrocompetent bacterial strain

(USOΔ*hisB*ΔpyrFΔrpoZ) were used in the B1H selections. After electroporation, the selections were plated on selective NM minimal medium plates that contained 50 μM IPTG & 1 mM 3-AT as well as plates that contained 50 μM IPTG & 2 mM 3-AT at $1x10^6$ cells and left to grow for 20-24 hours. The selection plates were then counted and washed in order to collect all positive selection colonies. The library reporter plasmids were obtained via QIAGEN mini prep kits and were PCR amplified using HU100 and OK181 primers to obtain only the N6 library binding sites. The PCR products were sequenced through Sanger DNA sequencing and the reads were analyzed for the KJ ZFP binding sites. Of the 98 KJ ZFP clones selections, the DNA-biding sites of 64 clones matched those predicted by the Joung Lab while the binding sites of 34 clones did not match those predicted. The binding sites of 2 clones were not able to be determined due to bad sequencing reads (supplemental 1). The Sanger DNA sequence for clone 49-3 as well as Zif268 can be seen in figure 5 below.

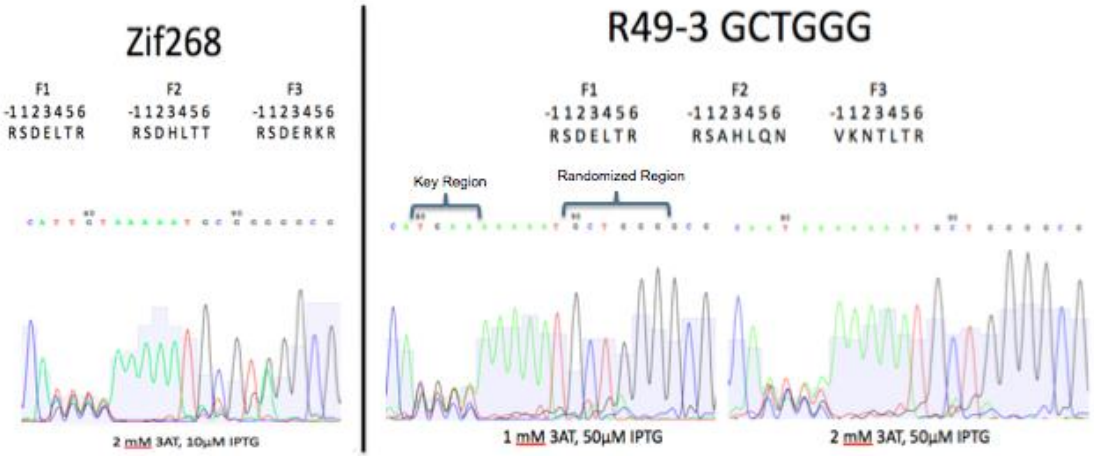


Figure 5: Sanger DNA-Binding Site Sequence Results for KJ Clone 49-3 and Zif268. The figure on the right shows the sequence results for Zif268, while the figure on the right shows the sequence results for KJ Clone 49-3. The name of the clone appears at the top of the image along with the expected binding site of the clone. The finger sequence can be seen below the clone name flowed by the sequence read. The key region as well as the randomized region of the N6 library is indicated above the biding site reads. Each clone was selected at 50μM IPTG + 1 mM 3-AT as well as 50μM IPTG + 2 mM 3-AT. The expected biding site for KJ clone 49-3 was GCTGGG, which matched with the determined binding site, GCTGGG. The anchor finger bound to GCG as was expected.

The DNA-binding site specificities of the 98 KJ ZFP clones at both 1 and 2 mM 3-AT concentration were next analyzed further through Illumina sequencing. All colonies for each of the clones at the two 3-AT concentrations were pooled and the plasmids isolated via QIAGEN mini prep kits. The library binding sites of the pooled plasmid samples were amplified by PCR and a 3 or 4 bp barcode was added to the 5' or 3' end of each pooled sample of the 98 KJ ZFPs to keep track of each sample in the Illumina Sequencing. The sequencing results were then analyzed using both W log-odds and GRaMS methods as described in Gupta et al. Sequence logos were obtained for all 98 KJ Zinc Finger Protein clones which matched the determined binding site of the clones well. It can also be observed that the higher 3-AT concentration provides a better look at the DNA binding motif due to the higher stringency of binding site selections. The corresponding W log-odds and GRaMS plots for KJ Zinc finger protein clone 49-3 can be seen below in figure 6.

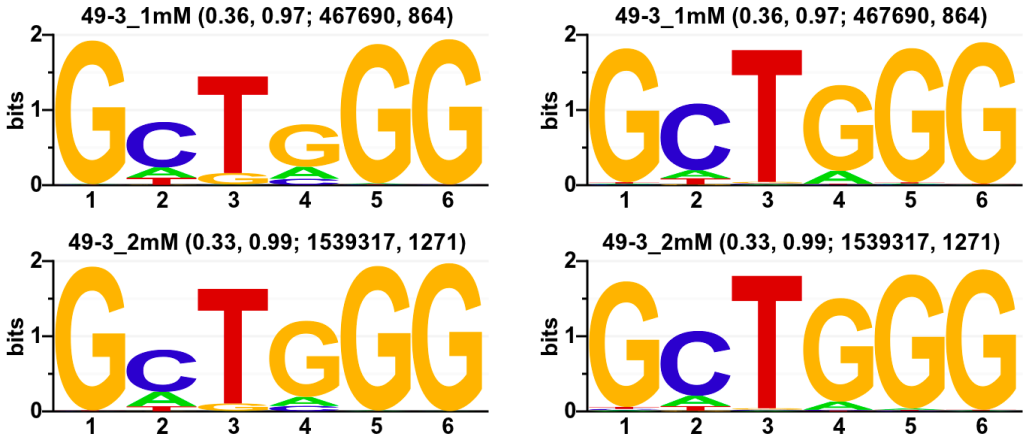


Figure 6: Illumina Data Analysis: Sequence logos for KJ clone 49-3. This figure shows a set of sequence logos for KJ Clone 49-3 at the two different stringencies used in selections, 1 mM 3-AT (top) and 2 mM 3-AT (bottom). The logos on the left are GRaMS plots while the logos on the right are W log-odds plots. The height of each letter corresponds to the frequency with which it binds in that position, with the most prominent base pair located on top. The height of the entire stack is adjusted to signify the information content of the sequence that particular position measured on a bit scale.

Using all of the selection data from the 98 KJ clones, a predictive model program was created by the laboratory of Gary Stormo and can be found at: http://stormo.wustl.edu/ZFNModels/. The model predictions the specificity of ZFs based upon a random forest model and can output a sequence logo in both a PFM and Information Content (IC) form. Screen shots of the input and output pages for the predictive model can be found in figure 7 below.

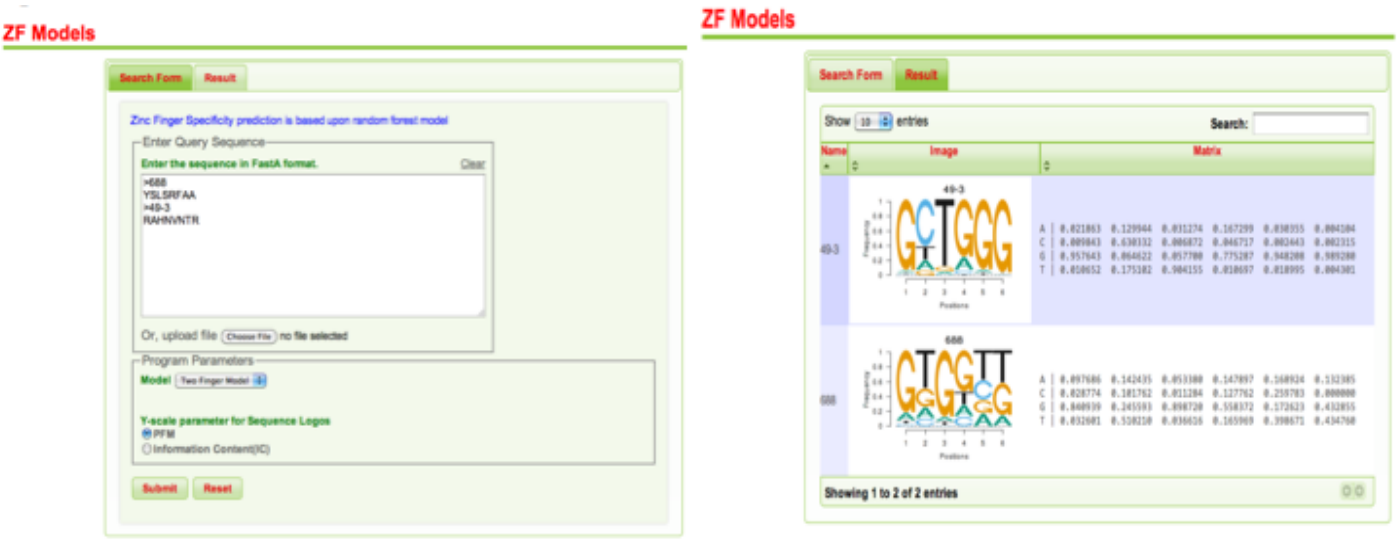


Figure 7: Zinc Finger Predictive β-Model Input (left) and Output (right) Screenshots

The predictive model was next tested using naturally occurring 3-finger Human Zinc Finger Proteins (HZFPs). The HZFPs used in this project were chosen based on the presence of interesting combinations of recognition sequences within the ZFPS from those identified, using a SMART database search for the Cys_2His_2 ZFPS that contain SCAN domains in the human genome.

The bacterial one-hybrid system was again used to determine the DNA-binding site specificities of the HZFPs. A 1352 ω-UV2 ZFP bait plasmid that contained the Keith HZFP fused to the ω-subunit of RNA polymerase, a pH3U3 library reporter plasmid containing an N28 library and an electrocompetent bacterial strain (USOΔ*hisB*ΔpyrFΔrpoZ) were used in these B1H selections. After electroporation, the selections were plated on selective NM minimal medium (-histidine, -uracil) plates that contained 50 μM IPTG & 2 mM 3-AT as well as plates that contained 50 μM IPTG & 5 mM 3-AT at $1x10^6$ cells and left to grow for 40-48 hours. Binding sites of sixteen colonies from the 5mM 3-AT selection plates were then PCR amplified using HU100 and OK181. The PCR products were sequenced through Sanger DNA sequencing and the binding site motifs were analyzed using MEME (Multiple Em for Motif Elecitation).

While 18 3-fingered HZFPs were identified for selections, only selections for clone 705A were successful. When the recognition helix of clone 705A was entered into the predictive model the output prediction results matched well with both the obtained binding site from MEME as well as the predicted binding site based upon analysis of the 98 KJ ZFP clones. The amino acid sequence for clone 705A is shown below with the residues in the recognition helices underlined.

YQCNLCEKAYTNCFHLRRHKMTHTGERPYTCHLCRKAFTQCSHLRRHEKTHTGERP
YKCHQCGKAFIQSFNLRRHERTHLGKK

The MEME binding site logo of clone 705A as well as the predictive model's sequence log can be seen below in figure 8.

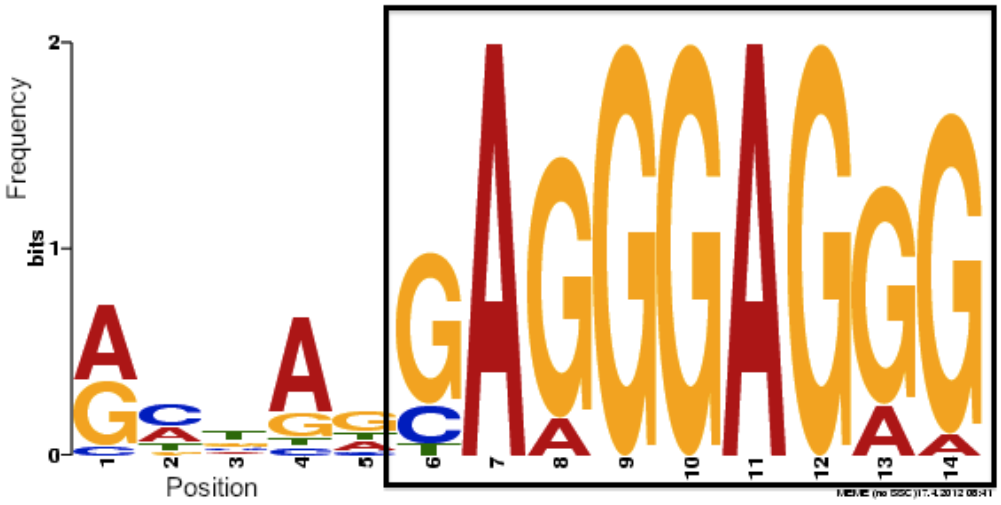
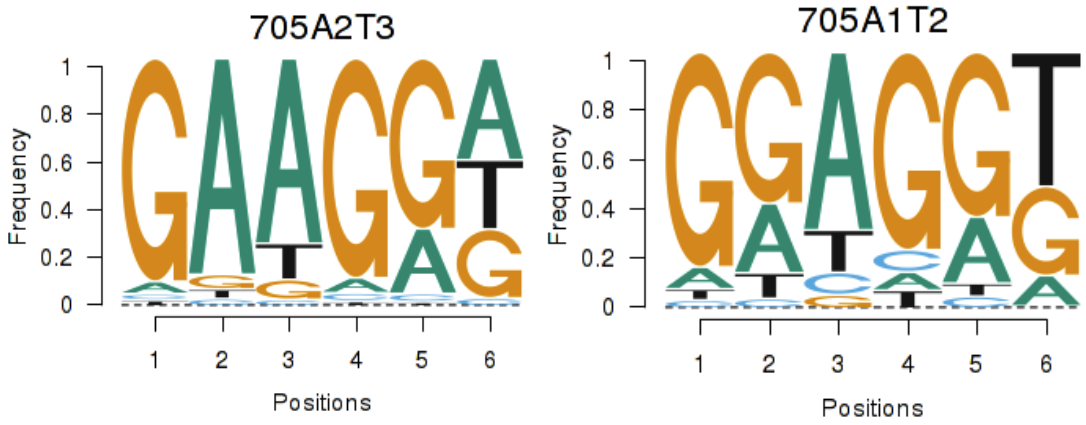


Figure 8: Binding site of HZFP clone 705A, obtained using MEME (above), Predicted binding site of HZFP clone 705A Fingers 2 through 3 (below, left) and 705A Fingers 1 through 2 (below, right).

(Above) The sequence logo for clone 705A, a three finger ZFP from a human SCAN-domain containing a ZF protein, matched the expected binding site based on the recognition helix sequences found above the figure. The motif for 705A was obtained through colony PCR of the 5 mM selection plate followed by Sanger DNA sequencing and MEME motif analysis and shows good specificity at most positions.

(Below) The recognition helix sequences for clone 705A were then entered into the Predictive Model to obtain the logos below. It can be observed that the logos below match the binding site motif from the bacterial one-hybrid selections of clone 705A. NOTE: the fingers are not in the same order as the motif above.



Due to the fact that only one selection of the HZFPs identified showed good results, further testing of the predictive model was done through inputting the recognition helices of ZFPs with known specificities into the predictive model and comparing the output results with the known binding site logos. The results showed that the predictive model was an accurate predictor of the binding motifs for the known ZFs. A table containing the 1-finger recognition helices as well as their known biding sites can be found below along with the predictive model's sequence logos generated for the already known 1-finger modules (figure 9).

Table 1: Amino Acid Sequences of ZF Clones and Their Triplet Recognition Sequences (Zhu et al., 2011).

Clone ID	Known Triplet Recognition Sequence	Recognition Helix Sequence
A1	AAG	RSDNLTQ
A2	ACG	RSDTLTQ
A3	AGG	RSDHLTQ

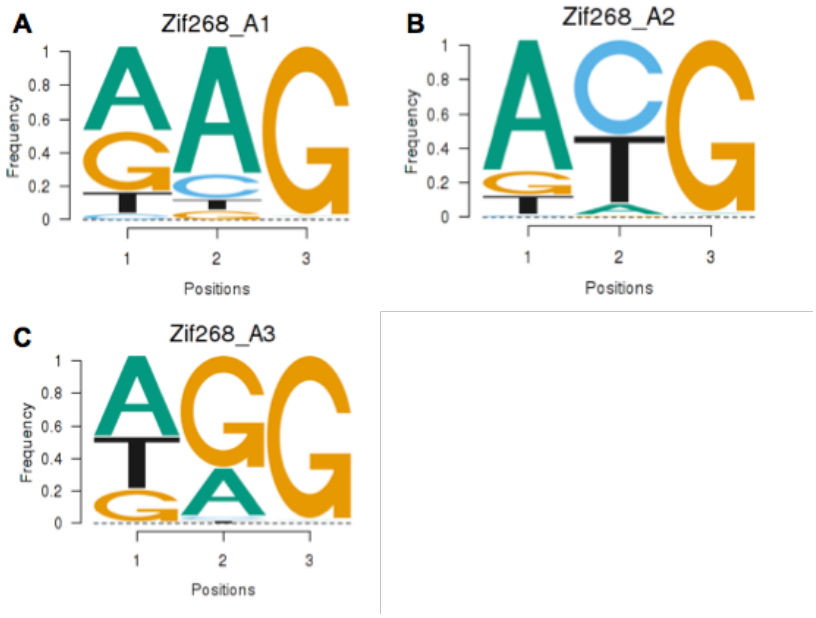


Figure 9: Predicted binding sites of ZF clones from Table 1, obtained using the ZFP Predictive model. (A) Clone A1, (B) Clone A2, (C) Clone A3. As can be seen by comparing the predictive model sequence logos to the known triplet recognition sequences, the predictive model provides an accurate prediction of the known ZFP binding sites.

DISCUSSION

Through the analysis of the DNA-binding site specificities of 98 Zinc Finger Protein clones, determined through a bacterial one-hybrid selection system, a predictive model was created that can accurately predict the binding site motifs of novel Zinc Finger Proteins. This predictive model, created in collaboration with the laboratories of Keith Joung and Gary Stormo, provides a new tool for both accurately predicting the DNA-binding sites of ZFPs, as well as aids in the designing of ZFPs for a desired target site.

Testing of the predictive model using ZFPs of known specificity showed good results with the predictive model sequence logos matching the known binding sites extremely well (figure 9). From this it was concluded that the predictive model provided an accurate binding site prediction and was tested further with a naturally occurring human zinc finger protein, clone 705A. The binding site of clone 705A was determined using the bacterial one-hybrid selection system as well and the binding site motif was created using MEME. The binding site sequence logo generated from MEME matched both what was expected as well as matched the predictive model's output (figure 8).

While eighteen HZFPs were chosen for selections, only clone 705A produced binding site results. After 48 hours of incubation at 37° C, only clone 705A and the Zif268 control plates had shown significant colony growth. Small colonies were observed on the remaining selection plates but upon colony PCR and Sanger sequencing analysis, showed no DNA-binding motifs. However, while the DNA binding motif predicted by the ZFP predictive model matched the actual DNA binding motif determined using MEME for 705A (see figure 8), further selection experiments with HZFPs should be done to test the predictive model further.

Additionally, further experiments determining the binding site specificities of larger ZFPs as well as ZFPs that do not contain the first finger of Zif268 could be done to expand the data set used to create the predictive model further enhancing the model's ability to predict the specificities of unknown ZFPs.

Currently, the model predicts the specificities of ZFs based upon a random forest model and can output a sequence logo in both a PFM and Information Content (IC) form. To take the model one step further a true recognition code with a one-to-one relationship should be created. However, due to the complexity of the relationship between the residues in the recognition helix and the basepairs they contact, one residue can have an affinity for more than just one basepair making the creation of a true recognition code more complex.

In conclusion, the predictive model created in this project provides the next step in creating novel DNA-binding proteins based upon zinc finger proteins. This tool allows for the accurate prediction of ZFP binding sites and can aid in the design of engineered zinc finger proteins, which can be designed to recognize any DNA sequence with high affinity and specificity. Additionally, attaching addition domains for activation, repression, or enzymatic activity, to a novel, engineered ZFP allows for the protein to carry out a desired function in a site-specific manner providing a new method of gene therapy.

BIBLIOGRAPHY

Gupta, et al. (April 29th, 2012) An optimized two-finger archive for ZFN-mediated gene targeting. *Nature Methods.*

Klug A (2010) The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation. *Annu. Rev. Biochem.* 79:213-231

Meng X, Brodsky MH, Wolfe SA (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature Biotechnology* 23: 988-944.

Meng X, et. al. (2007) Profiling the DNA-binding specificities of engineered Cys2His2 Zinc finger domains using a rapid cell-based method. *Nucleic Acids Research,* 28: 1-9

Meng X, Noyes MB, Zhu LJ, Lawson ND, Wolfe SA (2008) Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nature Biotechnology* 26: 695-701

Meng X, Wolfe SA (2006) Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nature Protocol* 1: 30-45

Noyes MB, et al. (2008) A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Research* 36: 2547-2560

Papworth M, Kolasinska P, Minczuk M (2005) Designer Zinc-finger proteins and their applications. *Gene* 366: 27-38

Wolfe SA, Nekludova L, Pabo CO (2000) DNA Recognition by Cys2His2 Zinc Finger Proteins. *Annu. Rev. Biophys. Biomol. Struct.* 29: 183-212

Zhu C, et al. (2011) Evaluation and application of modularly assembled zinc-finger nucleases in zebrafish. *Development* 138: 4555-4564

SUPPLEMENTAL 1

WOLFE LAB NAME	KJ NAME	Expected binding site	Binding site	F1	F2	F3
	Zif 268			RSDELTR	RSDHLTT	RSDERKR
A211	R79-1	GTTGAG	GTTGAG	RSDELTR	RQDNLGR	ARHRLIP
B2	R12-1	GAAGGG	GAAAGG	RSDELTR	RAEHLTN	QHPNLTR
			JOUNG LAB	RSDELTR	RRAHLGN	QHPNLTR
B3	R12-1	GAAGGG	GAAGGG	RSDELTR	RAEHLTN	QAPNLGR
B4	R13-1	GAAGTG	GATGTG	RSDELTR	RRNILQN	LSSNLTR
B7	R16-2	GACGGG	GAAGGG	RSDELTR	RVDHLHR	GGDNLVR
B8	R17-4	GACGTG	GACCTG	RSDELTR	RPQILIN	DPSNLRR
B9	R18-4	GAGGAG	GAGGAG	RSDELTR	RPDNLGR	RHDQLTR
B10	R18-5	GAGGAG	GAGGAG	RSDELTR	RPDNLGR	RBDNLPR
B11	R19-1	GAGGCG	GGGGTG	RSDELTR	RRESLVR	RDDHLGR
B12	R19-2	GAGGCG	GAGGTG	RSDELTR	RTEVLTN	RHDQLTR
			JOUNG LAB	RSDELTR	REDTLTR	RHDQLTR
C12	R34-4	GCGGAG	GGGGAG	RSDELTR	RQDNLGR	RREGLGR
D1	R20-1	GAGGGG	GAGGTG	RSDELTR	RKAHLKN	RRDNLLR
B1	R10-3	GAAGAG	GATGAG	RSDELTR	RRENLIR	LSSNLTR
F1	R70-1	GTAGTG	GTGGCG	RSDELTR	RADSLPR	RTDSLPR
			JOUNG LAB	RSDELTR	RKQILNN	QGGALQR
F2	R71-1	GTCGAG	GTGGCG	RSDELTR	RSDDLRR	RTDSLPR
			JOUNG LAB	RSDELTR	RGDNLGR	DLSSLPR
F3	R72-1	GTCGCG	GGTGAG	RSDELTR	RQDNLGR	LGHTLNR
			JOUNG LAB	RSDELTR	RSDDLRR	ESGALRR
D2	R20-4	GAGGGG	GAGGGG	RSDELTR	RRAHLGN	RQDNLQR
			JOUNG LAB	RSDELTR	RREHLTR	RNDKLVP
D4	R22-4	GATGAG	GATGAG	RSDELTR	RVDNLGR	ISHNLAR
D5	R23-4	GATGCG	GATGCG	RSDELTR	RQDDLTR	LSQNLGR
D7	R30-1	GCCGAG	GGGGAG	RSDELTR	RADNLAR	EHRGLKR
D8	R30-2	GCCGAG	GCTGAG	RSDELTR	RGDNLVR	GRDSLTR
D9	R30-3	GCCGCG	GACGAG	RSDELTR	RPDNLGR	DHSNLSR
D10	R34-2	GCGGAG	GCGGAG	RSDELTR	RRENLKR	RTDSLPR
D11	R34-3	GCGGAG	GGTATG	RSDELTR	RQDNLGR	RHQGLHH
A1	R48-4	GCTGCG	GCTGTG	RSDELTR	RADGLTR	LKHDLGR
A2	R48-5	GCTGCG	GGTGCG	RSDELTR	RRDDLTR	LGHTLNR
A3	R49-2	GCTGGG	GTTAGG	RSDELTR	RNDHLTN	VTNSLTR
B4	R50-1	GCTGTG	GCTGTG	RSDELTR	RVEVLTN	VRNTLTR
A5	R51-3	GGAGAG	GGTGAG	RSDELTR	RPDNLGR	KKDTLGN
A6	R54-3	GGAGTG	GAAGTG	RSDELTR	RREVLMN	QTTHLSR
A7	R55-5	GGCGAG	GGCGAG	RSDELTR	RADNLGR	DPSHLPR
A8	R57-1	GGCGGG	GGCGGG	RSDELTR	RGEHLTR	ESGHLKR
A9	R58-5	GGCGTG	GGCATG	RSDELTR	RRDLLHN	KNISLNH
A10	R59-1	GGGGAG	*BAD READ*	RSDELTR	RTDNLDR	RIDKLGG
A11	R59-3	GGGGAG	GGGGAG	RSDELTR	RPDNLGR	RVSHLQR
E2	R71-1	GTCGAG	GTCGAG	RSDELTR	RGDNLGR	DLSSLPR
E3	R72-1	GTCGCG	GTCGCG	RSDELTR	RSDDLRR	ESGALRR
F4	R73-2	GTCGGG	GTCGGG	RSDELTR	RQEHLVR	DRTPLNR
E5	R74-2	GTCGTG	GTCGTG	RSDELTR	RTDGLVR	ERRSLGR
E6	R75-4	GTGGAG	GTGGAG	RSDELTR	RDDNLQR	RPDALPR
E7	R75-5	GTGGAG	AAGGAG	RSDELTR	RQDNLGR	RDANLAT
E8	R76-2	GTGGCG	GTGGCG	RSDELTR	RPDDLRR	RPDALPR
E9	R76-3	GTGGCG	GTGGCG	RSDELTR	REDTLTR	RGANLNL
E10	R77-1	GTGGGG	GTGGGG	RSDELTR	RVEHLNN	RMDALMR
E11	R78-3	GTGGTG	GTGGTG	RSDELTR	RTEILRN	RHSTLTR
C3	R62-1	GGGGTG	GGGGTG	RSDELTR	RREVLMN	RNHGLVR
C4	R62-2	GGGGTG	GGGGTG	RSDELTR	RREVLEN	RNHGLVR
C5	R63-1	GGTGAG	GGTGAG	RSDELTR	RLDNLDR	HTHRLVS
D6	R63-5	GGTGAG	GGTGAG	RSDELTR	RRENLKR	IRHHLKR
C7	R64-4	GGTGCG	GGTGCG	RSDELTR	RPDDLRR	AGGGLAR
C8	R64-5	GGTGCG	GGTGTG	RSDELTR	REDGLHR	HTHRLVS
C9	R65-2	GGTGGG	GTTGAG	RSDELTR	RQEHLVR	HTHRLVS
C10	R66-3	GGTGTG	GGTGTG	RSDELTR	RVEVLTN	IKHHLGR
G1	R79-2	GTTGAG	GTTGAG	RSDELTR	RRDNLRR	IRTSLKR
			JOUNG LAB	RSDELTR	RRENLIR	IRTSLKR
G2	R79-3	GTTGAG	GTTGAG	RSDELTR	RADNLGR	ARHNLVP
G3	R80-2	GTTGCG	GTTGCG	RSDELTR	RADSLPR	IRTSLKR
G4	R80-3	GTTGCG	GTTGCG	RSDELTR	RADTLRK	HHNSLTR
G5	R81-3	GTTGGG	GTTGGG	RSDELTR	RAEHLTN	INHSLRR
G6	R81-4	GTTGGG	GTTGGG	RSDELTR	RAAHLDN	VNSSLGR
G7	R82-1	GTTGTG	GTTGTG	RSDELTR	RRQILSN	HHNSLTR
G8	R82-2	GTTGTG	GTTGTG	RSDELTR	RRNILQN	HHNSLTR
R15-1	R15-1	GACGCG	GACCCG	RSDELTR	RTDDLKR	DPSNLRR
R21-1	R21-1	GAGGTG	GAGGTG	RSDELTR	RRQILR	RRDNLLR
R21-1	R21-1	GAGGTG	GAGGTG	RSDELTR	RRSILAN	RGDNLAR
R22-1	R22-1	GATGAG	GATGAG	RSDELTR	RPDNLGR	VVNNLAR
R24-3	R24-3	GATGTG	GATGGG	RSDELTR	RAAHLDN	VTNNLKR
R24-4	R24-4	GATGGG	GATGGG	RSDELTR	RNTHLDN	VTNNLKR
R25-3	R25-3	GATGTG	GATGTG	RSDELTR	RTEVLTN	VVSNLRR
R27-2	R27-2	GCAGCG	GGGGAG	RSDELTR	RQDNLGR	KRVSLNL
R27-5	R27-5	GCAGCG	GCAGCG	RSDELTR	RADSLPR	QGGTLRR
R54-4	R54-4	GGAGTG	GGAGTG	RSDELTR	RREVLVN	QSQHLVR
R55-1	R55-1	GGCGAG	GGTGAG	RSDELTR	RQDNLGR	KRVSLNL
R56-1	R56-1	GGCGCG	GGCGCG	RSDELTR	RRDDLQR	ETGHLKR
R58-1	R58-1	GGCGTG	GCCGCG	RSDELTR	RADSLPR	ERRGLHR
R60-2	R60-2	GGGGCG	*BAD READ*	RSDELTR	RRDDLTR	RNDKLGP
R60-5	R60-5	GGGGCG	GTGGCG	RSDELTR	RQDDLTR	RNDKLGP
R61-3	R61-3	GGGGGG	GGGGGG	RSDELTR	RREHLTR	RNDKLVP
R69-3	R69-3	GTAGGG	GTAGGG	RSDELTR	RQEHLVR	QHSSLSR
R70-4	R70-4	GTAGTG	GTAGTG	RSDELTR	RAGILTN	QRGLSGR
R71-3	R71-3	GTCGAG	GTCGAG	RSDELTR	RRENLKR	DQTVLRR
R73-1	R73-1	GTCGGG	GTCGGG	RSDELTR	RQEHLVR	EGGALKR
R74-4	R74-4	GTCGTG	GTCGAG	RSDELTR	RPDNLGR	DRTPLQR
R77-3	R77-3	GTGGGG	GTGGAG	RSDELTR	RVDHLHR	RGDPLHR
R10-5	R10-5	GAGGAG	GAAGAG	RSDELTR	RNENLLR	QGPNLSR
R15-4	R15-4	GACGCG	GACGCG	RSDELTR	RRDDLTR	EGGNLMR
R17-1	R17-1	GACGTG	GACTTG	RSDELTR	RRQILRN	DPSNLRR
R17-4	R17-4	GACGTG	GACCTG	RSDELTR	RPQILIN	DPSNLRR
R28-1	R28-1	GCAGGG	GCAGGG	RSDELTR	RQEHLVR	QGGTLRR
R28-5	R28-5	GCAGGG	GCAGGG	RSDELTR	RREHLAR	QGGTLRR
R29-2	R29-2	GCAGTG	GCAGTG	RSDELTR	RREVLMN	QGGTLRR
R29-5	R29-5	GCAGTG	GCAGTG	RSDELTR	RSEVLAN	QGGTLRR
R46-2	R46-2	GCGGTG	GTGGTG	RSDELTR	RRQILLN	RPDGLAR
R46-3	R46-3	GCGGTG	GCGGTG	RSDELTR	RRNILQN	RLDMLAR
R47-3	R47-3	GCGGAG	GGTGAG	RSDELTR	RQDNLGR	VSNTLTR
R49-3	R49-3	GCGGGG	GCTGGG	RSDELTR	RSAHLQN	VKNTLTR
R50-5	R50-5	GCTGTG	GTTGTG	RSDELTR	RTEVLAN	VGASLKR
R51-1	R51-1	GGAGAG	GGAGAG	RSDELTR	RSDNLGK	QTTHLSR
R51-2	R51-2	GGAGAG	GGAGAG	RSDELTR	RPDNLVR	QGGHLAR