Graduate School of Biomedical Sciences

GSBS Dissertations

University of Massachusetts Medical School *Year 2012*

GETTING A TIGHT GRIP ON DNA: OPTIMIZING ZINC FINGERS FOR EFFICIENT ZFN-MEDIATED GENE EDITING

Ankit Gupta

University of Massachusetts Medical School

GETTING A TIGHT GRIP ON DNA:

OPTIMIZING ZINC FINGERS FOR EFFICIENT

ZFN-MEDIATED GENE EDITING

A Dissertation Presented

By

Ankit Gupta

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

April 27th, 2012

Department of Biochemistry and Molecular Pharmacology

Program in Gene Function and Expression

GETTING A TIGHT GRIP ON DNA: OPTIMIZING ZINC FINGERS FOR EFFICIENT ZFN-MEDIATED GENE EDITING

A Dissertation Presented
By
Ankit Gupta

The signatures of the Dissertation Defense Committee signify completion and approval as to style and content of the Dissertation

Scot Wolfe, Ph.D., Thesis Advisor

Alexander Schier, Ph.D., Member of Committee

Nathan Lawson, Ph.D., Member of Committee

Dan Bolon, Ph.D., Member of Committee

Sean Ryder, Ph.D., Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

Kendall Knight, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Biochemistry and Molecular Pharmacology
April 27th, 2012

Acknowledgement

Last five years in graduate school have been a great learning experience and I would like to thank several people who made it possible. First and foremost, I would like to thank my thesis advisor, Scot Wolfe. Perhaps, only a few people know that I wanted to join Scot's lab since I was in India. When I joined UMass and talked to senior graduate students about Scot and his work, my interest in joining his lab only rose. I am grateful that he gave me an opportunity to work with him. Although the start of my graduate research was rather bumpy, Scot not only kept patience but also was very efficient in keeping me motivated about research. One of the distinguishing qualities of Scot is that he appreciates hard work even if your experiments failed which was really encouraging for me. Soon after I joined the lab, I realized how meticulous Scot is in his thinking about experiments. This became more apparent later as we submitted our papers for publication and the reviews never came back with more experiments to be done. One of the qualities that make Scot a great leader is that he leads by setting an example. He himself is a really hard worker that sets a high standard for everyone in the lab. Moreover, he always respects his colleagues, be students or post-docs which is critical for maintaining healthy relationships. Scot is not only a gifted scientist and a great leader but also a family oriented person and exemplifies all I want to be in future.

I would also like to thank my committee members Kendall Knight, Nathan Lawson, Sean Ryder, and Dan Bolon who provided valuable suggestions on my research and

have always been supportive. Of special mention is Nathan Lawson, I will forever be grateful for your guidance and suggestion all through this time. I am also grateful to all our collaborators especially Gary Stormo and his student Ryan Christensen for performing computational analysis on our data and providing valuable input on my projects. I would also like to thank Alex Schier for taking time from his busy schedule to be an external examiner for my thesis defense. I would also like to thank all the past and present lab members from the Wolfe lab. Xiang Dong Meng, Marcus Noyes and Joseph McNulty encouraged me to join the lab and made the lab a cordial environment to work. They taught me a lot during the initial years when I needed training. I also thank Amy Rayla, Victoria Hall and Heather Bell for helping me in my experiments. I would also like to thank Stephanie Chu, Cong Zhu and Selase Enuameh for being great lab mates and sharing their reagents, knowingly or unknowingly. Finally I would like to thank all the Lawson lab members who taught me the zebrafish work and for being patient with me.

In the end I would like to thank my loving family for supporting me throughout my career and life. None of my achievements would have been possible without sacrifices and unconditional support of my parents. I am grateful to them for putting emphasis on education and hard work right from my childhood. They have always believed in me and supported my choices, however unconventional they were to them. I thank them for dedicating their lives to see me succeed. I would also like to thank my brother for being there for me always. I would also like to

thank my in-laws for letting me marry their daughter and for their unwavering support. In the end I would like to thank Pallavi, my wife. I call myself lucky that I married someone from the research field because we have so much to talk to when we go home. Although, the last few years have not been easy for her, but her support for me never waivered. I really appreciate her for not just understanding but being the best companion in lab when I had to work long days, late nights, weekends, and holidays. Whenever I was frustrated, she provided the encouragement and optimism needed to move forward. I thank her for bringing out the best in me. I am also blessed to have Soham, the newest member of our family.

In the end, I would also like to thank all my friends especially Srivatsan, Naveen, Nitya, Pranav, Srikanth, Pallavi, Venky, Samyabrata, Seemin, Tathagat and Kamna, who have been a family away from home.

Abstract

The utility of a model organism for studying biological processes is closely tied to its amenability to genome manipulation. Although tools for targeted genome engineering in mice have been available since 1987, most organisms including zebrafish have lacked efficient reverse genetic tools, which has stymied their broad implementation as a model system to study biological processes. The development of zinc finger nucleases (ZFNs) that can create double-strand breaks at desired sites in a genome has provided a universal platform for targeted genome modification. ZFNs are artificial restriction endonucleases that comprise of an array of 3- to 6- C_2H_2-zinc finger DNA-binding domains fused with the dimeric cleavage domain of the type IIs endonuclease *FokI*. C_2H_2-zinc fingers are the most common, naturally occurring DNA-binding domain, and their specificity can be engineered to recognize a variety of DNA sequences providing a strategy for targeting the appended nuclease domain to desired sites in a genome. The utility of ZFNs for gene editing relies on their activity and precision *in vivo* both of which depend on the generation of ZFPs that bind desired target sites high specificity and affinity.

Although various methods are available that allow construction of ZFPs with novel specificities, ZFNs assembled using existing approaches often display negligible *in vivo* activity, presumably resulting from ZFPs with either low affinity or suboptimal specificity. A root cause of this deficiency is the presence of interfering interactions at the finger-finger interface upon assembly of multiple fingers. In this study we have employed bacterial-one-hybrid (B1H)-based selections to identify two-finger zinc finger units (2F-modules) containing optimized interface residues that can be combined with

published finger archives to rapidly yield ZFNs that can target more than 95% of the zebrafish and human protein-coding genes while maintaining a success rate higher than that of ZFNs constructed using available methods. In addition to genome engineering in model organisms, this advancement in ZFN design will aid in the development of ZFN-based therapeutics.

In the process of creating this archive, we have undertaken a broader study of zinc finger specificity to better understand fundamental aspects of DNA recognition. In the process we have created the largest protein-DNA interaction dataset for zinc fingers to be described that will facilitate the development of better predictive models of recognition. Ultimately, these predictive models would enable the rational design of synthetic zinc finger proteins for targeted gene regulation or genomic modification, and the prediction of genomic binding sites for naturally occurring zinc finger proteins for the construction of more accurate gene regulatory networks.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
TF	Transcription factor
DBD	DNA-binding domain
ZFD	Zinc finger domain
ZFP	Zinc finger protein
2F-module	Two-finger zinc-finger module
ZFN	Zinc finger nuclease
B1H	Bacterial one-hybrid
B2H	Bacterial two-hybrid

PREFACE

The work reported in this dissertation has been published in the following articles.

Chapter II has been published previously as:

Ankit Gupta, Xiangdong Meng, Lihua J. Zhu, Nathan D. Lawson, and Scot A. Wolfe (2011). Zinc finger protein-dependent and -independent contributions to the in vivo off-target activity of zinc finger nucleases. Nucleic acids research 39, 381-392

Chapter III has been published previously as:

Christensen RG, Gupta A, Zuo Z, Schriefer LA, Wolfe SA, Stormo GD (2011) A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. Nucleic acids research 39, e83

Contents of Chapter IV have been accepted for publication

Ankit Gupta, Ryan G. Christensen, Amy L. Rayla, Abirami Lakshmanan, Gary D. Stormo, Scot A. Wolfe (2012)

An optimized two-finger archive for ZFN-mediated gene targeting, Nat. Methods, (Manuscript accepted).

CHAPTER I

INTRODUCTION

General

Complex metazoans such as humans can possess more than two hundred different cell types. Although these cell types originate from a common cell (the zygote), each one has its own gene expression profile, which helps to define its identity thus creating unique characteristics for each cell type. Each unique expression profile is the result of differential gene expression, which is controlled in large part by transcription factors (TFs). Regulation of gene expression by TFs plays pivotal roles in all cellular processes throughout development and when perturbed, can be detrimental to organismal fitness.

Transcription Factors

Transcription factors are *trans*-acting proteins that bind to *cis*-regulatory DNA elements either directly or indirectly through protein partners to influence the transcription of a gene. There are two broad categories of TFs: general transcription factors and gene-specific transcription factors. General transcription factors are components of the RNA polymerase machinery that are required for basal transcription of almost all genes and therefore are essential for an organism to survive[1,2]. Gene-specific transcription factors are required for transcription of certain genes and can act either as activators or repressors of transcription. These factors are central for differential gene expression and therefore are major determinants of phenotypic diversity and evolution of species[3].

Sequencing the human genome provided insights into total number of genes in each family of transcription factors[4]. The human genome consists of ~20,000 genes[5,6] which

seems low considering much simpler organisms, like drosophila and *C. elegans* have ~14,000[7,8] and ~20,000 genes[9] respectively. In contrast, the TFs represent ~10% of the total genes in the human genome[10], whereas they represent only ~5% in the drosophila and *C. elegans* genomes[8,11] suggesting an expansion of TF genes in more complex organisms. This increase in the TF numbers along with the increase in the non-coding genome, which includes binding sites for these TFs, may result in more entangled gene-regulatory networks imparting greater complexity to higher order organisms[12,13].

DNA binding domains

Majority of TFs consist of one or more *D*NA binding domains (DBDs) that provide sequence specificity or selectivity to the TFs allowing them to recognize their genomic targets. DNA binding domains have been categorized into different families based on their sequence homology. DBDs within a family also share basic mechanism of DNA recognition. Although TFs in the human genome can be divided into more than 20 groups based on their DBD, over 80% fall into three groups of DBDs namely, C_2H_2-Zinc finger domain (ZFD), Homeodomains (HD) and basic helix-loop-Helix domain (bHLH)[14].

Basic Helix-loop-helix proteins

Basic Helix-loop-helix proteins are the third most common type of TFs[14]. These proteins function as dimers where each monomer consists of two α-helices connected by a small

loop[15,16]. The basic region is an extension of the first helix and binds the major groove of DNA making base specific contacts. BHLH TFs typically bind to a consensus sequence CANNTG,which is known as the E-Box and play important roles in cell proliferation and differentiation[17].

Homeodomains

Homeodomains, first discovered in drosophila as homeotic genes[18], are the second most abundant DBDs in TFs[14]. These are ~60 amino acid long DBDs that fold into three α-helices preceded by a flexible N-terminal arm. Each homeodomain recognizes 3 to 8 bp of DNA where the third α-helix docks into the major groove of DNA making base specific contacts to the 3' end of the target sequence. The N-terminal arm may dock in the minor groove of DNA and contacts the 5´ bases. Recently, Noyes *et al.* determined the DNA binding specificity of 84 homeodomains from the drosophila genome and classified them into 11 different groups based on their specificities[19]. Homeodomains can bind DNA as both monomers and dimers, where the dimerization is mediated by the YPWM motif on the N-terminal arm of the homeodomain[20]. Dimerization of homeodomains allows them to bind to longer binding sites providing them additional specificity and finer control of gene expression. Homeodomain containing genes are involved in regulating developmental processes such as body axis formation, appendage formation, organogenesis and therefore it is not surprising that they are highly conserved from insects to mammals.

C_2H_2-Zinc finger domain

The C_2H_2-Zinc finger domain (ZFD) is the most frequently used DNA binding domain with more than 50% of TFs in the human genome incorporating this domain[14]. The C_2H_2-Zinc finger family has expanded considerably in mammals through gene duplications followed by functional divergence, where there is evidence of strong positive-selection pressure at the specificity determinants in paralogs generating novel recognition preference[12,21,22]. The expansion of zinc fingers has resulted in more than 700 multi-zinc finger genes in humans with ~8.5 zinc fingers per TF[4,12,22]. Since, each zinc finger in the canonical mode of DNA recognition (see below) binds to 3 basepairs of DNA, multifinger-TFs may recognize longer DNA sequences with higher affinities. However, all ZFDs in a multifinger TF may not be involved in DNA recognition. For example, TFIIIA contains 9 ZFDs, but only 6 of them bind DNA where only 4 make majority of the DNA contacts[23]. Similarly, CTCF contains 11 ZFDs but only fingers 4 through 8 are required for binding site recognition[24]. In fact, CTCF may employ different combination of fingers to bind different DNA sequences in the genome[25]. Moreover, in addition to binding DNA, C_2H_2-Zinc finger domains can bind to proteins and RNA[26-28]. For example, TFIIIA protein binds to both the DNA and the 5S rRNA[29]. Ikaros, a six finger zinc finger protein (ZFP), uses its four N-terminal fingers to bind DNA and the two C-terminal fingers for homodimerization[26].

Most zinc-finger containing TFs in the human genome are associated with either a **SCAN** (SRE-ZBP, CTfin51, AW-1 and Number 18 cDNA), **KRAB** (kruppel associated box) or

BTB (Broad complex, Tramtrack, Bric-a-Brac) effector domains or a combination of SCAN and KRAB domains. SCAN is small, leucine-rich domain that is found only in the vertebrates. It allows homo- and hetero-dimerization with other SCAN containing domains and is not associated with transcriptional regulating properties[30]. BTB domain is an evolutionary conserved domain that is found in all eukaryotic organisms. It is estimated that 5-10% of zinc finger containing TFs in the human genome incorporate the BTB domain[22]. The BTB domain allows homo- or hetero-dimerization of these TFs. Interactions of the BTB domain with corepressors Sin3A, SMRT, and N-CoR recruit histone deacetylase to the target genes which are then repressed[30]. KRAB domain is a tetrapod specific transcriptional repressive domain that is associated with almost 40% of zinc finger proteins[22]. The KRAB domain binds with its co-repressor, KAP1 (KRAB-associated protein 1) and similar to the BTB domain, recruits histone deacetylases and represses target genes through changes in chromatin architecture[31].

Although most zinc finger proteins associated with the KRAB and BTB act as transcriptional repressors, there are instances of zinc finger proteins such as Sp1[32] and Zelda[33] that act as transcriptional activators. Sp1, is a three finger ZFP that was identified from HeLa cell extracts as an activator of transcription of SV40 promoters[34]. Through biochemical experiments it was shown to bind to GC rich DNA elements[35,36]. Currently, it is known that Sp1 and its paralogs can regulate expression of more than 1,000 genes both TATA-containing and TATA-less genes[37] involved in many vital cellular functions through its interaction with transcriptional co-activators $hTAF_{II}130$ and $hTAF_{II}250$ or with other TFs[27,32]. Moreover, association of Sp1 with chromatin-

modifying factors such as p300 and histone deacetylases (HDACs) has been linked to its role in chromatin remodeling[37].

ZFD-containing TFs are involved in a wide variety of biological processes. TFIIIA is a gene specific transcription factor required for transcription initiation of the 5S rRNA gene, nuclear export and cytoplasmic storage of the 5S rRNA[27]. Egr1 (also known as Zif268), which contains a well-studied three finger ZFP is involved in differentiation, mitogenesis, and tumor progression[38]. CTCF (CCCTC binding factor) is a highly conserved TF in metazoans that is involved in chromatin remodeling, enhancer blocking, and transcriptional regulation[39]. A recent study showed that syntenic and constitutively bound CTCF sites overlap with cohesin-associated loci and may function as insulator elements[40]. Moreover, these CTCF sites preferentially flank disease-associated genes[40]. Prdm9 is a 12-finger zinc finger protein (ZFP) that is involved in specification of meiotic recombination hotspots[41,42]. Although, the DNA sequence for chimpanzee is highly similar to that of humans, the hotspot locations have changed between these species that has been linked to corresponding changes in the DNA recognition properties of Prdm9, implicating its importance in the evolution of species[41].

History of C_2H_2 zinc finger research

Zinc finger proteins (ZFPs) were first discovered in Xenopus through biochemical studies on TFIIIA which itself was the first transcription factor to be identified[29]. TFIIIA is a part of the transcription initiation complex that was known to bind RNA as well as DNA. Hanas *et al.* first reported zinc binding for TFIIIA but later Klug and colleagues reported a more correct stoichiometry for zinc binding[29,43]. Through biochemical experiments

Klug and colleagues showed that binding of zinc, and not other ions stabilizes the TFIIIA-RNA complex. They also found that each molecule of TFIIIA binds to approximately 7 molecules of zinc ions. Proteolytic digestion of TFIIIA resulted in intermediate products that differed in size by ~3KDa before finally breaking down into ~3KDa products implying that TFIIIA is composed of repeating structures. Alignment of the amino acid sequence of TFIIIA, published by Ginsberg *et al.*[44], revealed a repeating ~30 amino acid long motif that corresponded to a size of ~3KDa[29]. This motif was termed as 'zinc-finger' since it bound zinc and also gripped DNA. Since, there were 9 repeating motifs, each one binding to one zinc ion, the estimate of approximately 7 molecules of zinc per molecule of TFIIIA was almost right on target. The Klug lab also proposed a model for zinc finger binding to DNA that had some agreement with the structure of Zif268 published a few years later[29,45]. Soon after the zinc finger motif was reported in TFIIIA, two other proteins were reported to contain a zinc finger motif: the Serendipity gene and Kruppel, both from drosophila[46,47]. As additional zinc finger sequences became available, the following consensus sequence for zinc fingers was revealed[48]:

$$(F/Y)\text{-}X\text{-}C\text{-}X_{2-5}\text{-}C\text{-}X_3\text{-}(F/Y)\text{-}X\text{-}\mathbf{X}\text{-}\mathbf{X}\text{-}\mathbf{X}\text{-}\mathbf{X}\text{-}\Psi\text{-}\mathbf{X}\text{-}\mathbf{X}\text{-}H\text{-}X_{3-5}\text{-}H$$

where 'X' represents any amino acid and '**Ψ**' is a hydrophobic residue. The positions indicated in bold represent potential DNA-interacting amino acids on the recognition helix that were defined later through biochemical and structural analyses.

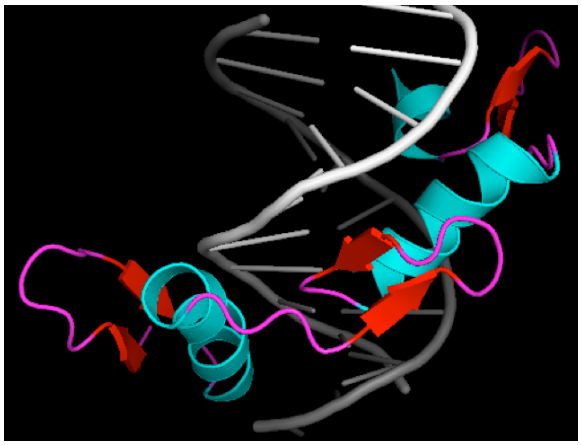
DNA recognition by zinc finger proteins

The NMR structures described by the Wright group revealed that zinc fingers fold into a ββα-fold stabilized by tetrahedral coordination of a zinc ion by two cysteines and two histidines[49]. However, significant understanding of DNA recognition by zinc finger came from the X-ray crystal structure of the DNA-bound Zif268 protein described by the Pabo lab[45,50]. Zif268, or Egr1, is a three-finger protein that has served as the framework for understanding DNA recognition and creation of zinc fingers with novel DNA binding specificity. Zif268 was crystallized with its preferred binding site (GCGTGGGCG) as determined from prior biochemical analysis[51] **(Figure 1-1a)**. The zinc finger folding, docking, and DNA recognition of Zif268 is considered as a benchmark for evaluating other zinc fingers and therefore are considered 'canonical' (for review refer to Wolfe *et al.*[48]).

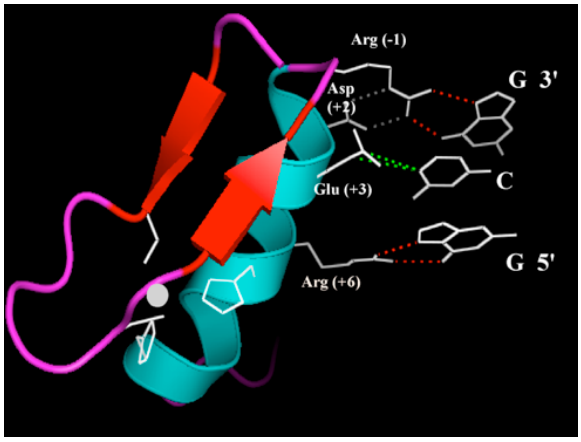
The crystal structure revealed that zinc fingers fold into a ββα-fold stabilized by tetrahedral coordination of a zinc ion by two cysteines and two histidines[45] **(Figure 1-1b)**. The α-helix docks into the major groove of DNA, making base specific contacts through residues on the 'recognition helix'. The amino acids on the recognition helix are numbered indicting their position relative to the start of the α-helix. The three fingers of Zif268 bind similarly to the DNA and make majority of the base specific contacts to one strand of the DNA called the 'primary strand' **(Figure 1-1c)**. The fingers run anti-parallel to DNA that is, the N-terminus of the zinc finger is facing the 3′ end of the binding site and vice versa **(Figures 1-1c and 1-2)**. In the canonical mode of binding,

Figure 1-1

A



B



C

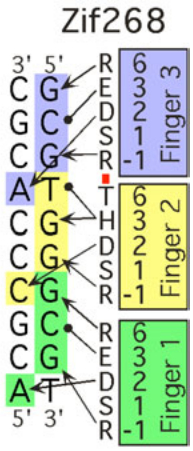


Figure 1-1: Structure of Zif268. (a) X-ray crystal structure of Zif268 bound to its binding site[45]. (b) Close-up of finger-1 of Zif268 bound to the 'GCG' subsite[45]. Amino acids that contact the 'GCG' triplet are highlighted and numbered referring to their positions relative to the start of the recognition helix. Hydrogen bonds from the arginines to guanines are shown in red and from Asp at +2 to Arg at -1 are shown in grey. Glu at +3 makes van der Waals contacts (green) to cytosine. (c) Schematic showing base contacts for Zif268[45,52]. Arrows represent the hydrogen bonds and black dots indicate van der Waals contacts. Red bar indicates inter-finger interactions.

Figure 1-2

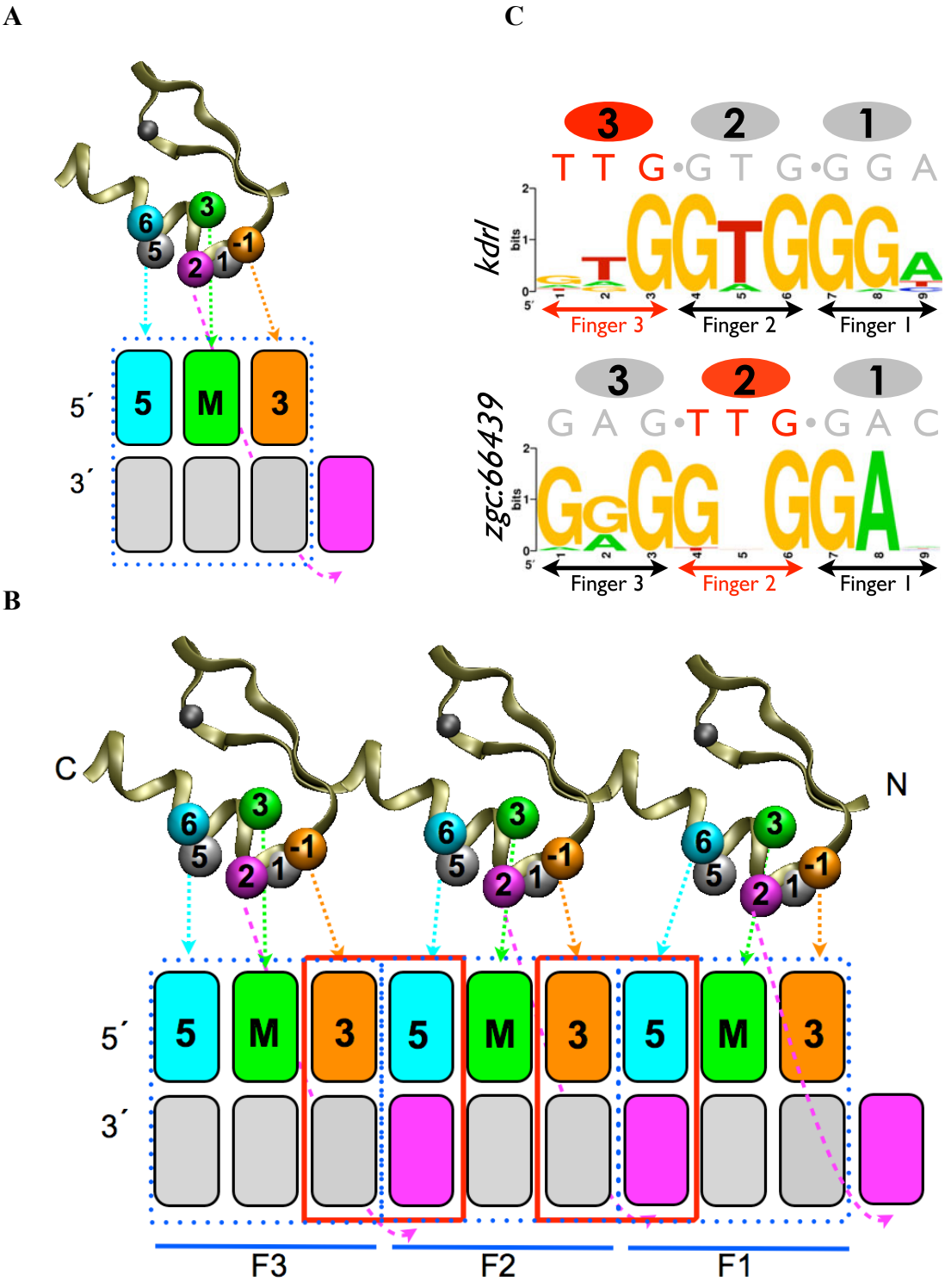

A
C
B
5´
3´
5
M
3
kdrl
zgc:66439
bits
Finger 3
Finger 2
Finger 1
C
N
F3
F2
F1

Figure 2-2: Canonical model of DNA recognition by zinc fingers and context dependent effects. (a) In the canonical mode of DNA recognition, each zinc finger binds to 3 basepairs of DNA (represented by 5,M and 3) in the anti-parallel orientation. The amino acids on the recognition helix are numbered relative to the start of the helix. Residues at positions 6, 3, and -1 bind to three basepairs 5, M and 3 respectively. Amino acid at position 2 contacts the base (pink) outside the 3 bp core site on the complementary strand of DNA. (b) In a multi-finger protein, zinc fingers bind adjacent 3bp sites. The red-boxed basepairs form the 2 bp junction that is specified by the amino acids at the interface of the two neighboring fingers. The pink base contacted by the amino acid at position 2 of the C-terminal finger is shared by the amino acid at position 6 of the neighboring N-terminal finger. (c) An example of context dependent alteration of specificity. The 'TTG' finger (red) in different contexts in two 3-finger proteins shows different specificities (red triplets)[53].

each finger binds to a core 3bp element of DNA (represented by NNN) with amino acids at positions -1, 3, and 6 making base specific contacts to the 3′ base, middle base and the 5′ base respectively in the 3bp core element **(Figures 1-1 and 1-2)**. The contacts made by the finger-1 of Zif268 to the 'GCG' core DNA sequence are shown in **figure 1-1b**. The amino acid at position 2 makes a contact to the base complementary to the base just outside (3') the 3bp core element **(Figures 1-1c and 1-2)**. Since neighboring fingers in a ZFP bind to adjacent DNA triplets, the base outside the core triplet is shared by the amino acid at position of 6 of the neighboring N-terminal finger **(Figures 1-1c and 1-2)** and therefore, the sequence preference of a finger can be influenced by the sequence of the neighboring finger.

Many of the neighboring zinc fingers in ZFPs are linked by an evolutionary conserved 5-amino acid linker (TG(E/Q)KP)[22]. This linker is flexible in solution but takes a more rigid conformation upon binding to DNA based on NMR analysis of the free and DNA-bound protein[49]. In addition to the phosphate contact made by lysine, threonine, and glycine cap the C-terminus of the neighboring N-terminal finger's helix[48]. Mutating residues T, G, K, and P in the TGEKP linker of TFIIIA resulted in atleast 6-fold reduction in its affinity to DNA showing that the linker sequence is important for binding of zinc fingers to DNA with high affinity[54,55]. Moreover, phosphorylation of Thr/Ser residue of the (T/S)G(E/Q)KP linker can modulate binding of the ZFP to DNA and might be used by nature to control activity of ZFPs during mitosis[56,57]. Also, the sequence and the length of the linker dictates the spacing between the subsites for neighboring fingers. Adjacent

fingers linked by the conserved five amino acid linker, bind adjacent DNA subsites **(Figures 1-1c and 1-2)**. One of the potential problems with polydactyl proteins with more than 3 or 4 fingers is that they might get out of phase of the DNA helical pitch[48,58]. Studies have shown that using a modified non-canonical linker, with an extra residue inserted between the canonical TG(E/Q)KP linker, to connect every 2N and 2N+1 finger in a 5 or 6 finger ZFP can help build high affinity polydactyl ZFPs that may show improved discrimination for their cognate vs mutant sites[59,60]. Moreover, using longer linkers or inserting extra spacer finger can help build ZFPs that bind neighboring subsites separated by 1 or more base pairs[59-65].

The Zif268-like canonical mode of binding is followed by many zinc fingers from other proteins such as finger 2 of tramtrack[66], finger 3 of YY1[67], and finger 2-4 of WT1[68]. However, DNA recognition by many zinc fingers deviates from this canonical mode of binding. For example, F1 of tramtrack[66], fingers 4 and 5 of Gli[69], fingers 1, 2 and 4 of YY1[67], fingers from the engineered zinc finger protein-$TATA_{ZF}$[52] make DNA contacts in a pattern different than the Zif268-like fingers and thus follow a non-canonical mode of DNA binding even though in some cases (e.g. $TATA_{ZF}$) they dock with the DNA in a Zif268-like manner[52].

Using zinc fingers to create artificial DNA binding domains

Initial site-directed mutagenesis experiments demonstrated that the specificity of zinc fingers can be altered to bind desired sequences[70-72]. The finger swapping experiments revealed the modular nature of zinc fingers[71]. Moreover, first Choo *et al.* and later Liu *et*

al. demonstrated that artificial zinc finger proteins can be constructed and used to recruit an effector domain to desired locations in the genome to modulate gene expression in a directed manner[73,74]. These characteristics of zinc fingers present the opportunity to engineer zinc finger proteins (ZFPs) to bind a desired DNA sequences where these proteins could then be used to regulate gene expression at will. Understanding the protein-DNA interactions that define sequence preference on a broad scale for zinc fingers would allow the development of recognition models that would enable not only rational design of sequence specific ZFPs but also prediction of specificities for naturally occurring zinc finger containing TFs. In this direction, zinc finger libraries based on the backbone of Zif268 or its variant with randomized DNA-contacting amino acid residues have been utilized to isolate novel zinc fingers with desired DNA-binding specificity via selection based methods.

Selections of Zinc finger proteins with novel DNA-binding specificity

The following selection techniques have been successfully utilized to isolate active zinc finger members from a randomized library.

Phage Display

In the past, phage display based selections, where each member of zinc finger library is expressed on the surface of a M13 bacteriophage, have been successfully employed to obtain zinc fingers that bind to desired DNA sequences with high affinity[75-86]. Pabo,

Barbas, and Klug labs have utilized phage display to select a large number of zinc fingers with novel specificities (discussed below). Most of the zinc finger modules that are currently being used by Sangamo BioSciences were selected using phage display. Although, this method allows the use of large unbiased zinc finger libraries with up to 10^{10} members, the selection of zinc fingers is performed *in vitro* through multiple rounds of enrichment, which sometimes drives the selections in favor of high affinity instead of specificity.

Yeast based selection method

In comparison to the phage display, the yeast based selection method allows selection of zinc finger proteins *in vivo* but owing to the low transformation efficiency of yeast, only medium complexity libraries with ~10^7 members can be searched restricting the use of this method to explore complex libraries[87]. A yeast-based reporter system has also been used to screen for zinc fingers from the human genome to isolate zinc fingers that could be employed to build artificial ZFPs[88].

Bacterial selection methods

There are two bacterial selection methods available: the bacterial-1-hybrid (B1H) and the bacterial-2-hybrid (B2H). The B1H method, a modified version of the original bacterial system described by Dove *et al.*, uses the yeast *HIS3* and *URA3* genes as reporter genes downstream of the zinc finger binding site (Note: the *URA3* gene also serves a counter-selectable marker for eliminating self-activating or false positive sequences from the library). The zinc finger library is fused directly to either the α-subunit of the RNA

polymerase[89,90] or in the more recent version, to the ω-subunit of the RNA polymerase[19]. A member of the zinc finger library and the binding site reporter plasmid are compartmentalized in a bacterial cell lacking the *hisB* (bacterial homolog of *HIS3*), *pyrF* (bacterial homolog of *URA3*) and *rpoZ* (gene encoding the ω-subunit -only for the recent ω-subunit version) genes. Binding of the ZFP to its binding site recruits the RNA polymerase (via the fused α- or ω-subunit) on the promoter controlling the reporter genes inducing their expression and ultimately allowing the bacterial cell containing the active zinc finger protein to grow on the selective media. Originally described for determining the binding site specificity of transcription factors[19,89-93], the B1H system can also be used to select for ZFPs that bind to desired target sites[94].

The B2H system, working on a similar principle as B1H, employs a reporter construct containing the reporter genes *HIS3* and *aadA* downstream of the zinc finger binding sites. When the zinc finger member fused to the Gal11p domain binds to the desired target site, dimerization of the Gal4-fused α-subunit of the RNA polymerase is induced allowing expression of the reporter genes and survival of bacterial cells on selective media[95-100]. In comparison to the B2H system, the ω-based B1H selections can be performed in a ω-deficient strain (Δ*rpoZ*) providing higher sensitivity to the system that allows weaker DNA-protein interactions to be characterized[92].

The B1H and B2H selection methods combine the advantages of phage display and yeast-based selection system. They are performed *in vivo* where the bacterial genomic DNA acts as a sink for non-specific binding of zinc finger proteins thus selecting ZFPs

for both specificity and affinity[89,91,94,95]. Moreover, bacterial based selection methods allow survey of large libraries with ~10^9 zinc finger members in a single round of selection as compared to multiple rounds of enrichment required in phage display thus making the selection process easy and rapid.

Selection Strategies

Starting from the single finger selections, the selection strategies to identify ZFPs with desired specificities were refined as the field matured. The following is a summary of the different selection strategies that have been employed, their advantages and disadvantages.

Single finger selections ignoring complications of context-dependence

The foremost selection studies involved creating libraries of zinc fingers based on three finger proteins, Zif268 or a variant of this sequence[83,101], where the base specifying contacts for one of the zinc fingers were randomized and the other two fingers were kept constant due to the limitations of the library size that can be screened by the utilized selection systems. This library was then used for selections via one of the selection methods described above to identify ZFPs that bind different 3bp subsites in a 9bp binding site[75-80,82,83,95]. The Barbas lab used this strategy to identify individual zinc fingers that bind almost all 3bp sequences[75,82,83]. These individual zinc finger modules could be combined to rapidly create multi-finger zinc finger proteins in a process termed modular assembly. Although, modularly assembled ZFPs have been shown to bind their desired targets *in vitro*, only 20% of the analyzed ZFPs contained a 'non-GNN' recognizing finger[102].

Modularly assembled ZFPs showed low activity when tested using the B2H-based reporter assay where their success rate was again dependent on the number of GNN recognizing fingers incorporated into the ZFP[100,103]. The low success rate of these modularly assembled ZFPs was attributed to their low affinity for the target sites[104] which presumably results from context dependent effects[82,83,100,103].

Problem of context dependence

One of the problems with the 'single finger selection, no context dependence' strategy is that it assumes a modular nature of zinc fingers ignoring two types of inter-finger interactions:

a) The amino acid at position 2 (and sometimes at position 1 [52]) interacts with the base outside the 3bp-core triplet on the strand complementary to the primary strand (Figures 1-1c and 1-2a)[45,105]. In the context of a multifinger ZFP, this base outside the core-3bp element is also shared by the N-terminal neighboring finger (Figure 1-2b). Therefore, when selections were performed in the context of the three-finger protein, the outcome was influenced by the neighboring fingers that were kept constant in the library of zinc fingers. Moreover, when individually selected zinc fingers are combined to create novel combinations of ZFPs the contact from the amino acid residue at position 2 to the 5' base of the subsite of the neighboring C-terminal finger can alter the DNA-binding specificity of the neighboring fingers altering[53] (Meng *et al.* unpublished data). These problems can be exacerbated if zinc fingers bind in a non-canonical mode making non-standard

contacts that can influence the specificity of the selected fingers or the neighboring fingers in an engineered protein[52].

b) Since the neighboring zinc fingers bind adjacent DNA subsites that are approximately 3.4 Å away, amino acids at the finger-finger interface, especially residue at position 6 of the N-terminal finger and residue at position -1 of the C terminal finger, can interact with each other[45,52,105]. These protein-protein interactions at the finger-finger interface can again influence the outcome of the selections and influence specificity of fingers in an engineered zinc finger protein.

Single finger selection factoring in context dependence

New strategies were developed to address the problems of context dependence. The Pabo lab developed a protocol for sequential selection of three-finger zinc fingers, where one finger is selected at a time while moving through the 9bp binding site[81,86]. This protocol although highly effective, is a labor-intensive process. Another strategy, termed OPEN (Oligomerized Pool ENgineering) involves constructing three different libraries, one for each of 3 fingers in the context of two constant anchor fingers and then searching them individually in parallel using the B2H system to select for a few active clones that bind constituent subsites of the desired 9bp site[97,99]. These active clones for each finger are then fused to yield a smaller library of three-finger proteins that is now reselected to identify for three finger ZFPs that bind the desired 9bp target site[96-99,106]. Although these selections yield highly active ZFPs, they are labor-intensive to perform and therefore they are not feasible for rapid generation of engineered zinc-finger proteins. Moreover, during the individual library

selections the finger-finger interactions still exist and can affect the outcome of the selections.

Selecting for interface residues

One of the ways to minimize effects of context dependence in multi-finger assemblies is by selecting for zinc-finger units with optimal interface-residues that are kept constant during the assembly. Isalan *et al.* created a library of three finger ZFPs where the residues at the finger-finger interface were randomized and then searched this library to identify groups of interface-residues that are optimal for binding to all 2-bp DNA junctions[84]. However, this study was not detailed enough to understand properties of interface recognition in different contexts and was never followed up to demonstrate its utility to build multi-finger proteins. The same group devised a 'bipartite' strategy where they created two overlapping three finger libraries of limited diversity each with randomized DNA contacting residues of 1.5 fingers[85]. This library was used to identify 1.5 fingers that bind overlapping half sites of the desired 9-bp sequence. The selected proteins were then combined to create active three-finger ZFPs. This strategy was further utilized to create an archive of two-finger (2F) units that could be combined to create multi-finger ZFPs. This archive, however, is proprietary to Sangamo Biosciences and thus is not available for general use. Recently, the Joung lab utilized their OPEN (Oligomerized Pool ENgineering) pools[98] to identify overlapping 2F-units that could be combined to create active ZFPs via context-dependent assembly (CoDA) method[100]. However, their archive mainly consists of modules that recognize 'GNN' triplets or 'N-G' junctions. Moreover, the artificial ZFPs and ZFNs assayed, were almost entirely

constructed from 2F-modules that recognize 'GNN-GNN' 6bp sites with 'N-G' type junctions that are not the limiting factor for advancing ZFN design[100].

Binding site selections

The artificial zinc-finger proteins selected from randomized libraries can bind to or even prefer other DNA sequences to the desired target site[82,83,86]. Therefore, determining the binding-site specificity is a useful way to validate the selected ZFPs. Moreover, identification of binding site preferences of transcription factors (TFs) would allow for the prediction of TF sites and *cis*-regulatory modules in the genome providing a better understanding of gene-regulation networks. The following is a summary of available methods for determining binding site specificities (reviewed by Stormo *et al*.[107]).

Initially, phage display and ELISA-based methods were used to determine the specificity of selected ZFPs[75,82,83,108]. However, these methods could determine specificity of only one finger at a time and were labor intensive. SELEX (Systematic Evolution of Ligands by EXponential enrichment) involves the extraction of binding-sites from an unbiased pool of oligonucleotides through multiple rounds of enrichment of sites bound to the protein[109-112]. It has been utilized to assess binding site specificities for selected ZFPs but since it involves multiple rounds of enrichment, energetic differences between different binding sites cannot be estimated[62,63,113-116]. A modified version of SELEX, HT-SELEX requires only one round of enrichment and couples this with high-throughput sequencing to develop energetic models that fit better than the regular SELEX[117]. However, like

SELEX this method is *in vitro* and requires protein purification or *in vitro* protein expression. Another *in vitro* method for identifying binding-site preferences for naturally occurring TFs or engineered ZFPs is using protein-binding microarrays (PBMs) wherein a library of binding sites is spotted onto a microarray chip. This chip contains ~40,000 spots of 60-mer oligos in which all possible 8-base sequences are represented in 32 different contexts[118]. The TF, either purified or *in vitro* expressed, is added to the chip and the TF-bound DNA oligos are identified using fluorophore-conjugated antibody that binds to the TF. The binding specificity of a TF can then be estimated using motif-finding algorithms. Although PBM is a medium throughput method to estimate binding specificities and affinities of DNA-binding proteins, it is limited by the number of sequences that can be interrogated on a chip thereby restricting its utility to TFs with short (~10mer) binding sites[107,118,119].

In comparison to *in vitro* methods, Chromatin Immunoprecipitation (ChIP) coupled with microarray (ChIP-chip) or high-throughput sequencing (ChIP-Seq) followed by computational analysis of over-represented sequence motifs provides a powerful way to determine *in vivo* binding sites for a TF in a genome[120-122]. However, this method is low throughput and dependent on the availability of antibodies specific to the protein of interest and on the condition of the *in vivo* sample. A more recent method, bacterial-one-hybrid (B1H) based selection method involves identifying ZFP binding sites from a randomized library through reporter activation[19,89-94,123]. This method has several advantages over *in vitro* methods and ChIP. This is a medium-throughput *in vivo* method

that does not require protein purification and involves only single round of selection thus making it easy to perform.

Recognition Codes

A recognition code is a set of rules describing the interactions of an amino acid with a nucleotide at a given position such that for any residue (amino acid or nucleotide) its interacting partner(s) can be accurately predicted. Although, the differences in docking mechanisms and non-independent interactions limit the possibility of an accurate universal code for all DNA-binding domains, there can be recognition codes that describe DNA-protein interactions within families and subfamilies of DNA-binding domains[124]. If such a code is available for zinc-finger proteins, it can be used to rationally design artificial ZFPs with desired specificity and also predict specificity of artificial as well as naturally occurring ZFPs. Qualitative models have been described for zinc finger-DNA recognition that provide information on the amino acids that can be used for recognizing desired nucleotides[48,108,125]. However, these models being qualitative, do not provide any information on the energetics of DNA recognition and also ignore context-dependent interactions. Benos *et al.* created quantitative models that used frequencies of contacts observed in SELEX and phage display data [126]. However, their predictive capabilities are limited mainly due to the paucity of high quality zinc finger-DNA interaction data that took into account the context-dependent effects and assumption of independence of protein-DNA interactions. Also, the DNA-protein data was biased toward proteins that

recognized 'GNN' subsites and, therefore, availability of extensive unbiased DNA-protein interaction data for zinc-finger proteins would be necessary to build accurate predictive models.

Applications of artificial zinc finger proteins

Owing to their semi-modular nature and ability to recognize a wide variety of DNA sequences, zinc-finger proteins proved advantageous for creating artificial DNA-binding domains that could recruit any given effector domain to a desired address in the genome[127]. Most common applications of ZFPs include their use as artificial transcription factors and as zinc-finger nucleases.

Artificial Transcription Factors

The first application for artificial ZFPs was to use them to mimic their natural function of gene regulation by fusing them with either an activation domain, such as the VP16 [128] or p65 [129] domain or a repression domain such as the KRAB domain[130]. Choo *et al.* first demonstrated artificial ZFPs selected using phage display could be used to repress as well as activate transcription[73]. Although the activation was mediated by the VP16 domain, repression was not induced by any effector domain but was caused by interfering with the interaction of RNA polymerase with its promoter[73]. Liu *et al*. in comparison employed the KRAB domain for repression of reporter genes in transiently transfected cells[74]. The Barbas lab showed for the first time that the individual fingers selected using phage

display could be combined to create artificial TFs that could specifically regulate expression of desired endogenous genes[82]. Since then a number of applications have been reported employing ZFPs as artificial TFs including the repression of HIV-1 5′ LTR promoter[131], the repression of herpes simplex viral (HSV) genes[132], the repression of *pparγ-1* and *pparγ-2* genes in 3T3 cells[133], activation of human *erythropoietin* gene (*EPO*)[134], the activation of *VEGFA* gene to induce angiogenesis[135] and the activation of *glial cell line-derived neurotrophic factor* gene[136]. Moreover, it has been demonstrated that the activity of these artificial TFs can be influenced by the chromatin structure at and around the target site and conversely, these artificial TFs can remodel the chromatin around the target site[82,134,137,138]. In another strategy to regulate gene expression, artificial ZFPs were fused to a chromatin modifying enzyme, v-ErbA, that repressed the targeted gene expression[139,140]. Some of these applications have potential for use as therapeutics. VEGFA activator (VEGFA-TF) is a chimeric protein with a p65 activator domain fused to the three-finger ZFP that binds a 9bp sequence GGGGGTGAC[135]. This activator upregulates expression of all isoforms of *VEGFA,* thus inducing angiogenesis resulting in the formation of new functional blood vessels[141]. Later, VEGFA-TF injection in the rat model of diabetes was shown to alleviate some symptoms of peripheral diabetic neuropathy, which causes damage to microvasculature in extremities encouraging the use of VEGFA-TF as a therapeutic[142]. Although the results of clinical trials showed that the therapy is safe but there was no clear statistically significant difference in alleviation of symptoms between patients who received the therapy versus those who got a placebo resulting in the recent cessation of the clinical trial[143]. Another artificial TF (GDNF-TF)

being evaluated in clinical trials is a six-finger ZFP fused to the p65 domain that activates the expression of the endogenous glial cell line-derived neurotrophic factor (GDNF) gene[136]. Microarray results showed that GDNF was the highest upregulated gene by GDNF-TF with only a few off-target genes being activated at lower levels. AAV2 mediated striatal delivery of GDNF-TF in rat model of Parkinsons disease resulted in improvements of certain symptoms of the disease. Pre-cinical trials are underway for this potential therapeutic.

Zinc finger Nucleases

Zinc finger nucleases are synthetic restriction endonucleases comprising of artificial zinc finger proteins fused to the cleavage domain of the *FokI* type IIs endonuclease. ZFNs can create a double strand break in DNA at targeted loci thus allowing targeted genome manipulation. The current architecture of ZFNs includes either a homodimeric or obligate heterodimeric versions of the *Fok*I cleavage domain fused by a linker to a 3- to 6-finger ZFP that binds to 9- to 18 basepairs of DNA. Since the *Fok*I endonuclease domain requires dimerization for activity, two monomers of ZFNs are designed each binding up to 18bp long DNA half-site where the two monomeric sites separated by a 5- or 6-bp spacer **(Figure 1-3)**.

***Fok*I**, isolated from *Flavobacterium okeanokoites,* is a type IIs restriction enzyme that binds to $^{5'}$GGATC$^{3'}$ sequence and non-specifically cleaves DNA 9/13-bp away leaving a 4bp 5′ overhang[144]. Through biochemical assays and later confirmed by structural

Figure 1-3

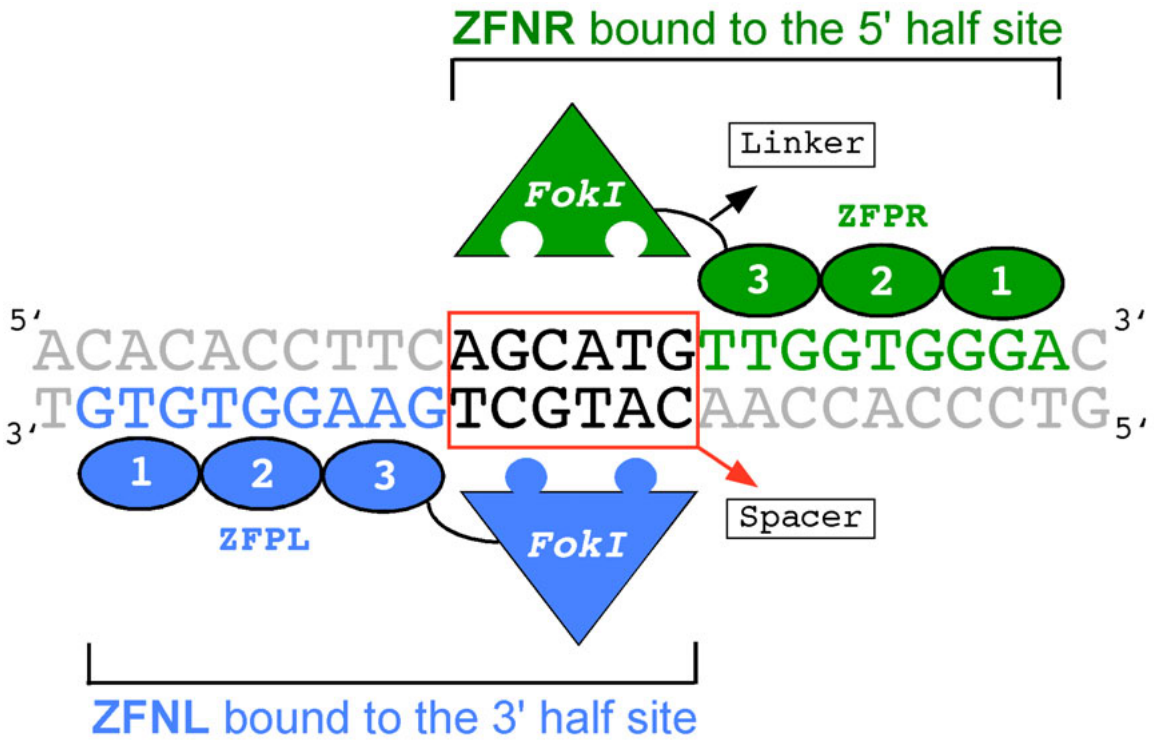


Figure 1-3: **Zinc finger nucleases.** A schematic drawing showing the ZFNs bound to the target site. The *Fok*I nuclease domain is fused at the C-terminal of the zinc finger protein. The two ZFN monomers (ZFNL and ZFNR) bind respectively to the 9 bp 5′ and 3′ half sites through the associated ZFPs (fingers indicated by numbered ovals), which position the heterodimeric nuclease domain over the 6 bp spacer between the two ZFP half sites.

studies, *Fok*I was shown to consist of two modular domains: the N-terminal DNA recognition domain and the C-terminal cleavage domain[145,146]. The modular nature of *Fok*I allowed replacement of its DNA-binding domain with other naturally occurring DNA binding domains[147-150]. Further studies, again both biochemical and structural, demonstrated that dimerization of *Fok*I cleavage domain is essential for its activity[151,152], however, at very high concentrations, only one monomer needs to bind DNA for activity[151]. As a result, dimerization interface mutants of *Fok*I (D483A and R487A) that were found to be inactive in a cleavage assay[151]. Interestingly, these same residues were engineered to create some of the obligate heterodimeric pairs (D483R in the 'RR' monomer and R487D in the complementary 'DD' monomer) of the *Fok*I nuclease domain[153]. These and other dimerization mutants force the ZFNs to heterodimerize and therefore, reduce undesired off-target activity of ZFNs[153-155]. However, these mutants of the dimerization interface show reduced on-target activity as compared to the wild type nuclease domain presumably due to the lowered dimerization potential [153-155]. To enhance the activity of these dimerization mutants, other mutations both at the dimerization interface[155] and the cleavage domain[156] of the FokI were recently described. Recently, ZFNickases were described where the cleavage domain of one of the nuclease domain was rendered inactive resulting in creating nick at the target site instead of a DSB[157,158]. These ZFNickases, although less active than ZFNs, allow DNA repair through HR and suppress NHEJ mediated repair and thus, would suppress unwanted lesions at off-target sites in the genome. These results show that the choice of the *Fok*I nuclease domain can effect on- as well as off-target activity of ZFNs.

The sequence specificity to ZFNs is provided by a tandem array of 3 to 6 zinc finger domains termed as zinc finger protein (ZFP) or zinc finger array (ZFA). Highly specific ZFPs are required both to increase the on-target activity of ZFNs and decrease their off-target activity[159,160]. ZFPs that bind desired binding site can be selected from randomized libraries as described above[81,94,99] but these selections are time consuming and require expertise. A rapid way to generate ZFPs is using modular assembly wherein pre-characterized zinc finger modules are combined to create tandem arrays of zinc fingers. Although zinc finger modules are available for most of 64 DNA triplets[75,82,83], but when combined with each other, they (especially the 'non-GNN' recognizing modules) often show context-dependent specificities (Meng *et al.*, unpublished data) which may result in low success rates of modularly assembled ZFAs as ZFNs[53,100,103]. Although archives of improved single zinc finger modules (1F-modules) exist, ZFNs assembled from them still show low success rate (~25%)[53,161]. One of the ways to minimize these context dependent interfaces is to identify two-finger modules (2F-modules) that contain optimal residues at the interface and combine them in a modular assembly fashion such that the interface residues are preserved. Using phage display and the bipartite strategy, 2F-modules have been selected and subsequently employed by Sangamo BioSciences to create active ZFNs[84,85,162]. However, these modules are a proprietary of Sangamo BioSciences limiting their use to ZFNs purchased through Sigma-Aldrich at a cost that is prohibitive for many laboratories. A recent study from the Joung lab used their OPEN selection strategy to identify 2F-modules that could be assembled using CoDA assembly method wherein the 2F-modules, F1-F2 and F2-F3 with a common F2, are assembled to

create three-finger ZFAs. ZFNs created using this assembly showed high activity in zebrafish[100]. However, their archive mainly contains 2F-modules that bind to 'GNN-GNN' 6bp sites with 'N-G' type junctions that are not the limiting factor for advancing ZFN design. In chapter IV, we describe a publicly available archive of 2F-modules that bind even 'non-N-G' type of junctions with high specificity and allow highly efficient ZFN-mediated gene targeting.

The linker joining the *Fok*I nuclease domain to the ZFP can also influence the activity and the specificity of ZFNs. Conventionally, a 4 amino acid linker (TGGS or LRGS) is used which allows for a 5- or 6- bp spacing between the two half-sites of a ZFN. Changing the length and composition of this linker can impact the spacing between the half-sites[64] and therefore can influence specific and non-specific activity of ZFNs.

Mechanism of Action of ZFNs

ZFNs upon binding to their half-sites in proper orientation create a double-strand break (DSB) at the spacer between the two half-sites which stimulates DNA repair pathways namely non-homologous end joining (NHEJ) and homologous repair (HR).

NHEJ primarily involves simple religation of the DNA ends created by the DSB but occasionally the ends are modified through insertion or deletion (InDel) of a few base pairs before ligation. These insertions or deletions can result in a frameshift of the open reading frame that might cause gene disruption. Since, ZFN cleavage predominantly leaves 4bp overhangs, majority of InDels are 4bp insertions that result from simple blunting of the overhangs followed by ligation[163,164].

HR is a more complex but precise DNA-repair pathway where a second copy of DNA that has homology to the region flanking the target site is required for repair. This copy can either be the second allele in the genome or an exogenously supplied double-stranded DNA or a single stranded oligonucleotide[165]. Unlike NHEJ, HR can use an exogenously-supplied DNA that allows precise genomic modifications including site-specific gene insertions, gene deletions, and gene correction. Although low frequency spontaneous HR has been used in ES cells of mice for genomic editing, HR in most other organisms is not feasible. However, DSB creation stimulates the rates of HR by ~5,000 fold[166]. ZFNs thus, have allowed high frequency genome modification through HR in organisms such as flies[167], mice[65] and rats[65].

Applications of ZFNs

The first use of ZFNs was reported by Chandrasegaran lab wherein they fused the *Fok*I cleavage domain to engineered zinc finger proteins[147]. Since then, ZFNs have been used for targeted gene editing in a variety of cell lines[62,63,113,115,168-171] and model organisms such as Drosophila[167], zebrafish[94,114,172], C. elegans[173], rats[174], mice[65,175], pigs[116,176], Arabidopsis[177], maize[178] and tobacco[179,180] (reviewed by Urnov *et al.*[162]). Although NHEJ mediated gene editing is reported in all these studies, HR mediated gene repair has been limited to only cell lines[162] and a few organisms such as flies[167,181], mice[65], and rats[65,174]. Orlando *et al.* demonstrated that NHEJ-mediated repair pathway following ZFN-induced DSB can be employed to insert double stranded DNA containing compatible overhangs at the target site[164]. Moreover, employing two pairs of ZFNs can

result in targeted large chromosomal deletions[182]. Since ZFNs allow targeted gene editing, they show tremendous potential for use as therapeutics and are being evaluated in clinical trials. Sangamo Biosciences demonstrated that ZFN mediated disruption of the CCR5 gene (the receptor required for HIV entry) in human CD4+ T cells and CD34+ hematopoietic stem cells allows HIV-1 control[113,183]. These ZFNs are currently undergoing the Phase II clinical trials (Sb-728). In another study they demonstrated that ZFN-induced DSB creation and homology mediated or homology independent repair allows for replacement or insertion of the blood coagulation factor IX gene in a hemophilia mouse model[63]. These ZFNs are currently undergoing pre-clinical trials.

Off-target effects of ZFNs

Similar to any other technology or drug, ZFNs also can have side effects. Like on-target activity, the off-target activity of ZFNs is influenced mainly by the specificity of the incorporated ZFPs, the choice of the nuclease domain and the sequence and composition of the linker as described above and will be described in detail in chapter 2. In majority of studies, the off-target activity of ZFNs is assessed by *in silico* identification of a few potential off-target sites (<20 heterodimer sites) followed by InDel detection either by deep sequencing or Cel1 nuclease assay[62,94,113,114]. Some of these studies detected infrequent off-target events in their cursory analysis. In the chapter 2 of this thesis we describe a more in-depth analysis of *in vivo* off-target activity for *kdrl* ZFNs and its dependence on the dose of ZFNs, specificity of the incorporated ZFPs and the choice of the *Fok*I nuclease domain[160]. Recent studies report a more genome-wide analysis of off-

target activity and are discussed in Chapter 2 [184,185]. A recently published study showed that the ZFN activity can also be affected by the chromatin structure around the target site which might also impact the off-target activity of ZFNs[186].

Gene targeting in zebrafish

Zebrafish, *Danio rerio*, is a tropical freshwater fish that has become an attractive model to study vertebrate development. Some of the attractive features of zebrafish that have facilitated research are its large clutch size (>100 embryos) and rapid development, which makes large forward-genetic screens and drug screens feasible. Moreover, since the zebrafish embryos develop *ex utero*, they can be viewed and manipulated at all stages of development. Most importantly, zebrafish embryos remain transparent through much of their early embryonic stages, allowing detailed microscopic observations. Although forward genetic approaches for manipulating the zebrafish genome have existed since early 1980s, development of reverse genetic approaches has been challenging. Reverse genetic approaches for gene manipulation prior to the use of ZFNs includes include Targeting Induced Local Lesions in Genomes (TILLING) or retroviral or transposon mediated insertional mutagenesis[172]. Since both of these approaches are random mutagenesis methods, a significant effort is required to carry out these screens. Morpholino mediated gene knockdown provides another pseudo-reverse genetics tool for studying gene function in zebrafish. Although it is gene-targeted approach, the knockdown is only transient lasting up to 72 hours post fertilization, thus limiting its applicability to genes that are expressed early in the development. Recently, ZFNs have

been employed in zebrafish for gene disruption. Since their first use by Meng *et al.*[94] and Doyon *et al.*[114], they have become a popular tool for gene inactivation in zebrafish[53,96,100,187-189]. ZFNs create a DSB at the desired target site in the genome that gets repaired by the NHEJ pathway resulting in small insertions or deletions at the target site where a subset of these lesions can result in gene disruption. The primary limitation on the use of ZFNs is the absence of a method to efficiently and reliably create highly specific ZFPs for use in ZFNs (see chapter IV).

Other Chimeric Nucleases

In addition to ZFNs, meganucleases[190] and the recently introduced Transcription Activator-Like Effector Nucleases (TALENs) allow targeted gene editing[173,191-193]. The utility of meganucleases is limited due to their inability to recognize a wide range of sequences. TALENs on the other hand have minimum limitations in terms of range of sequences that can be targeted and thus have a tremendous potential for gene editing[194,195]

Summary

Zinc fingers are naturally occurring DNA binding domains that can be engineered to recognize a variety of DNA sequences. Engineered zinc fingers can be utilized to create zinc finger nucleases that have demonstrated tremendous potential for targeted gene editing in cell lines and model organisms. The activity and precision *in vivo* of ZFNs is dependent on the specificity and affinity of the incorporated zinc fingers. We characterized the off-target effects of ZFNs in zebrafish and demonstrate their dependence on the specificity of the incorporated zinc fingers. We also employed

bacterial-one-hybrid based selections to identify and characterize zinc finger modules with optimal inter-finger interactions that allow efficient ZFN-mediated gene targeting *in vivo* and increased the targeting range of ZFNs by ~5-fold over existing modules. Our results not only advance ZFN-mediated gene targeting but also provide crucial understanding of zinc finger-DNA recognition that will in future facilitate development of predictive models of recognition that would allow the prediction of genomic binding sites for naturally occurring zinc finger proteins for the construction of more accurate gene regulatory networks.

CHAPTER II

ZINC FINGER PROTEIN-DEPENDENT AND -INDEPENDENT CONTRIBUTIONS TO THE *IN VIVO* OFF-TARGET ACTIVITY OF ZINC FINGER NUCLEASES

Contents of Chapter II have been published previously as:

Ankit Gupta, Xiangdong Meng, Lihua J. Zhu, Nathan D. Lawson, and Scot A. Wolfe (2011) Zinc finger protein-dependent and -independent contributions to the in vivo off-target activity of zinc finger nucleases. Nucleic acids research 39, 381-392

Xiangdong Meng selected the old and new *kdrl* ZFPs. Lihua J. Zhu performed the computational analysis of the data. Nathan Lawson injected the ZFNs in fish and isolated the genomic DNA. He also prepared the sequencing sample for the first Illumina run.

INTRODUCTION

Zinc Finger Nucleases (ZFNs) are artificial restriction enzymes that hold tremendous potential for the manipulation of genomes in a wide variety of plants and animals[162,196]. These enzymes generate a site-specific Double Stranded Break (DSB) that can abrogate gene function through imprecise repair (via generation of a frameshift) or can introduce tailor-made changes by stimulating homology directed repair from an exogenously supplied DNA template. The utility of ZFNs for gene inactivation and genome editing has been demonstrated in a wide variety of cell lines[170,171], including human ES cells and iPS cells[62,169], as well as in the germline of plants[179,197-199] [178,180]and animals[94,96,114,167,174,200]. Due to their demonstrated utility, ZFN-based therapies are being evaluated in clinical trials[113,201].

ZFNs are composed of two modular domains: a tandem array of Cys_2His_2 zinc fingers (ZFP) tethered to the cleavage domain of *FokI* endonuclease **(Fig 1-3)**[147]. The incorporated ZFPs can be engineered to recognize a specific DNA sequence[94,98,100,104,202], thereby targeting the attached nuclease domain to a desired location within the genome. Dimerization of the cleavage domain is required for enzymatic activity[151]. As a consequence, a pair of ZFNs must bind with the proper orientation and spacing to generate a DSB[64,203]. ZFN-mediated gene inactivation/modification is sufficiently robust to generate cell lines with multiple biallelic knockouts[204] and, when applied directly *in vivo*, founder animals that transmit mutant alleles to their offspring with high frequency[94,96,114,174,200]. However, in many instances cytotoxicity is observed as a side

effect of ZFN treatment, which presumably results from ZFN-generated DSBs at off-target sites within the genome[94,159,168,205].

Efforts to improve the *in vivo* precision of ZFNs have focused primarily on properties influencing DNA recognition. For each ZFN, the number of binding sites within a genome is primarily dictated by the number and quality of the incorporated zinc fingers. Consequently, utilizing ZFPs with higher specificity can reduce the cytotoxicity of ZFNs[159]. The type of nuclease domain dictates the active ZFN configurations. ZFNs bearing engineered nuclease variants that preferentially heterodimerize display reduced toxicity *in vivo* by disfavoring homodimeric DNA recognition[153,154]. The number of functional target sites is also defined by the composition and length of the linker joining the ZFP and nuclease domain, which determines the required spacing between ZFN half-sites for activity[64,203]. Finally, restricting the *in vivo* half-life of ZFNs can also attenuate their cytotoxicity[206].

Although the *in vivo* precision of ZFNs has been analyzed via the characterization of off-target lesion events, an in depth analysis of ZFN properties that influence these effects has not been performed. Potential off-target sites are typically defined by using the DNA-binding specificity of the incorporated ZFPs to scan the genome for sites most similar to these recognition sequences with the appropriate spacing for nuclease activity [62,94,113,114]. In the majority of these studies, ZFN-induced lesions are identified at these off-target loci by *Cel1* endonuclease or Restriction Fragment Length Polymorphism (RFLP) assays[62,114,174]. Most of these studies did not detect lesions at their predicted off-

target sites, however they typically examined only a small number of off-target sites (<10). Moreover, these assays are not sensitive enough to detect lesion frequencies at ≤ 1%[113]. In two studies[94,113], massively parallel sequencing technology has been used to characterize ZFN-induced off-target lesions with greater sensitivity. Both of these studies revealed that, although infrequent, lesions were present at a subset of the analyzed sites. However, only a small number of off-target sites were analyzed between these two studies: 7 heterodimeric sites for the CCR5 ZFNs and 17 for the *kdrl* ZFNs. Moreover, the influence of ZFN properties on *in vivo* precision was not examined in either of these studies.

The *kdrl* ZFNs, which display a low but measurable frequency of off-target events[94], provide an excellent system for exploring the parameters that affect ZFN precision *in vivo*. In our present study, we performed an in-depth analysis of ZFN precision by assaying lesion frequencies at 141 potential off-target sites in the zebrafish genome. The *kdrl* ZFNs generate lesions at a small subset of these sites and demonstrate greater promiscuity with increasing dose. Unexpectedly, we found that both the ZFP specificity and dimerization interface of the nuclease domain can influence the precision and activity of ZFNs. These results provide a broader picture of factors that influence the precision of ZFNs with implications for the best compositions to employ for genome manipulations in both model organisms and clinical gene therapy.

RESULTS

Off-Target analysis of Meng *et al*. *kdrl* ZFNs

In our previous study we demonstrated the efficacy of ZFNs targeting the *kdrl* gene in zebrafish, which incorporated ZFPs optimized through bacterial one-hybrid (B1H) selections and the "DD/RR" engineered heterodimeric nuclease domain[94]. As an initial assessment of the off-target lesions produced by these nucleases, we assayed the presence of lesions at 41 off-target sites (17 heterodimeric and 24 homodimeric), which revealed four heterodimeric off-target sites that accumulated lesions at a low frequency (~1%). However, the small number of off-target sites examined and the small number of sequences analyzed per site (~250) provided only a limited overview of the off-target activity in the genome.

In order to assess ZFN off-target activity in greater depth, we determined lesion frequencies generated by the *kdrl* ZFNs at the target site and off-target sites. The off-target sites were chosen based on the DNA-binding specificities of the ZFPs (ZFPL & ZFPR), which we previously determined using the B1H system[94]. To provide greater complexity to these motifs, we repeated B1H selections and sequenced the binding sites from the pool of surviving colonies by Illumina sequencing, where more than 1000 unique sequences were used to generate the binding site logo for each ZFP **(Figure 2-1)**. Based on these more informative motifs, we found that position 3 in the ZFPL motif and positions 1 and 2 in the ZFPR motif provide limited discrimination in DNA recognition. Consequently these positions were not considered when identifying the most favorable potential off-target sites within the genome based on matches to the target sequence. Based on the ZFPL and ZFPR binding specificity, we chose to characterize 141 putative off-target sites in the zebrafish genome that contained from one to five mismatches

Figure 2-1

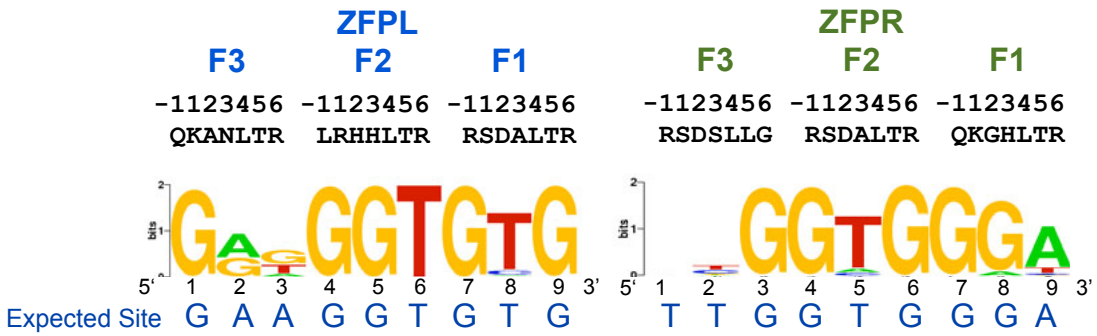


Figure 2-1: DNA binding specificities of *kdrl* ZFPs[94]. The DNA-binding specificities of the *kdrl* ZFPs determined at high stringency (5 mM 3-AT) using the B1H system displayed as a Sequence logo [207,208]. The expected binding site is provided at the bottom of the logo.

relative to the target site **(Figure 2-2 and Online Table[a])**. Twenty-eight of these sites represent potential recognition sequences for homodimeric ZFNs to examine the exclusivity of the engineered DD/RR nuclease domains. Among the remaining 113 heterodimeric sites, 59 contained the conventional 5 or 6 bp spacer between the two ZFN half-sites. The remaining 54 heterodimeric off-target sites contain a 14, 15 or 16bp spacer, as previous studies have indicated linker-dependent ZFN activity at sites with longer gaps between the half-sites[64,203].

To assess activity of ZFNs at these sites, zebrafish embryos were injected with two different doses of mRNAs (10 or 20 pg) encoding the *kdrl* ZFNs. These embryos were scored for viability and morphology at 24 hours post fertilization (hpf) to provide an overt assessment of toxicity **(Figure 2-3)**. At the 10 pg dose, ~50% of the surviving embryos were morphologically normal whereas the remainder displayed developmental abnormalities ('deformed' henceforth). Separate pools of ~25 injected embryos were prepared from morphologically normal and deformed embryos for lesion analysis. At the 20 pg dose, the majority of embryos were deformed or dead. Consequently, only deformed embryos were characterized at this dose. Restriction Fragment Length Polymorphism (RFLP) analysis confirmed the activity of *kdrl* ZFNs at the target site **(Figure 2-4)**.

[a] Online Table available at: http://nar.oxfordjournals.org/content/39/1/381/suppl/DC1

Figure 2-2

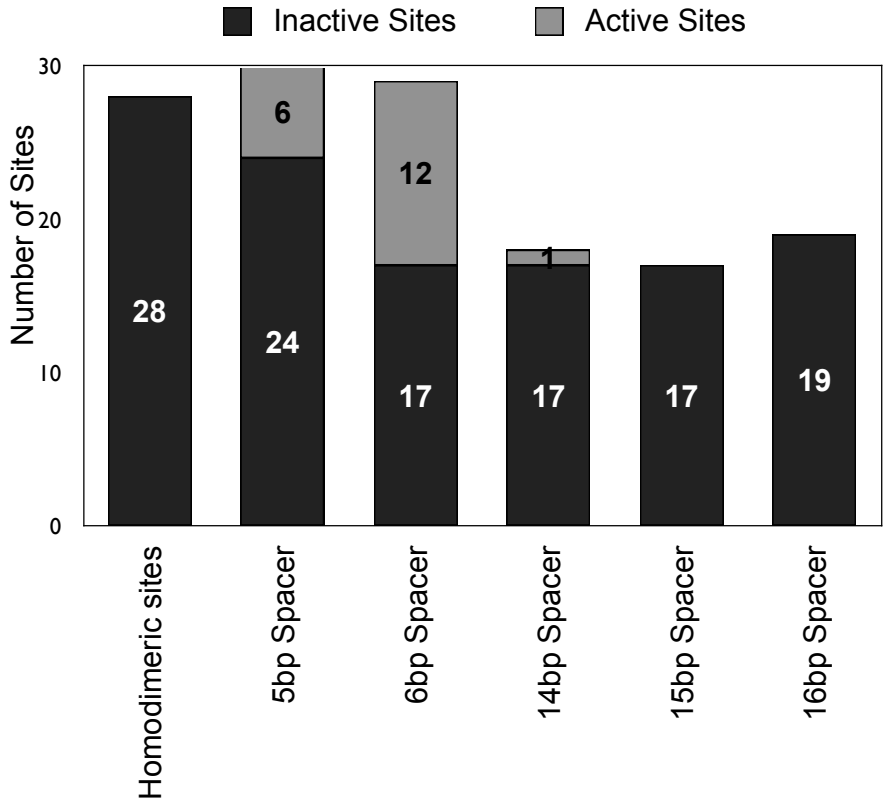


Figure 2-2: Overview of the off-target analysis for the original *kdrl* ZFNs. The number of active (grey) and inactive (black) off-target sites is depicted in the graph. The sites are subdivided according to the type of site (homodimeric or heterodimeric), where heterodimeric sites were divided into 5 different groups based on the spacing between the two half-sites. A total of 8 active off-target sites (see text for criteria) were found in normal embryos from 10 pg ZFN dose (Table1), and 11 additional off-target sites were active either only in deformed embryos from 10 pg or 20 pg ZFN dose or in one of the two biological replicates, as described in the text.

Figure 2-3

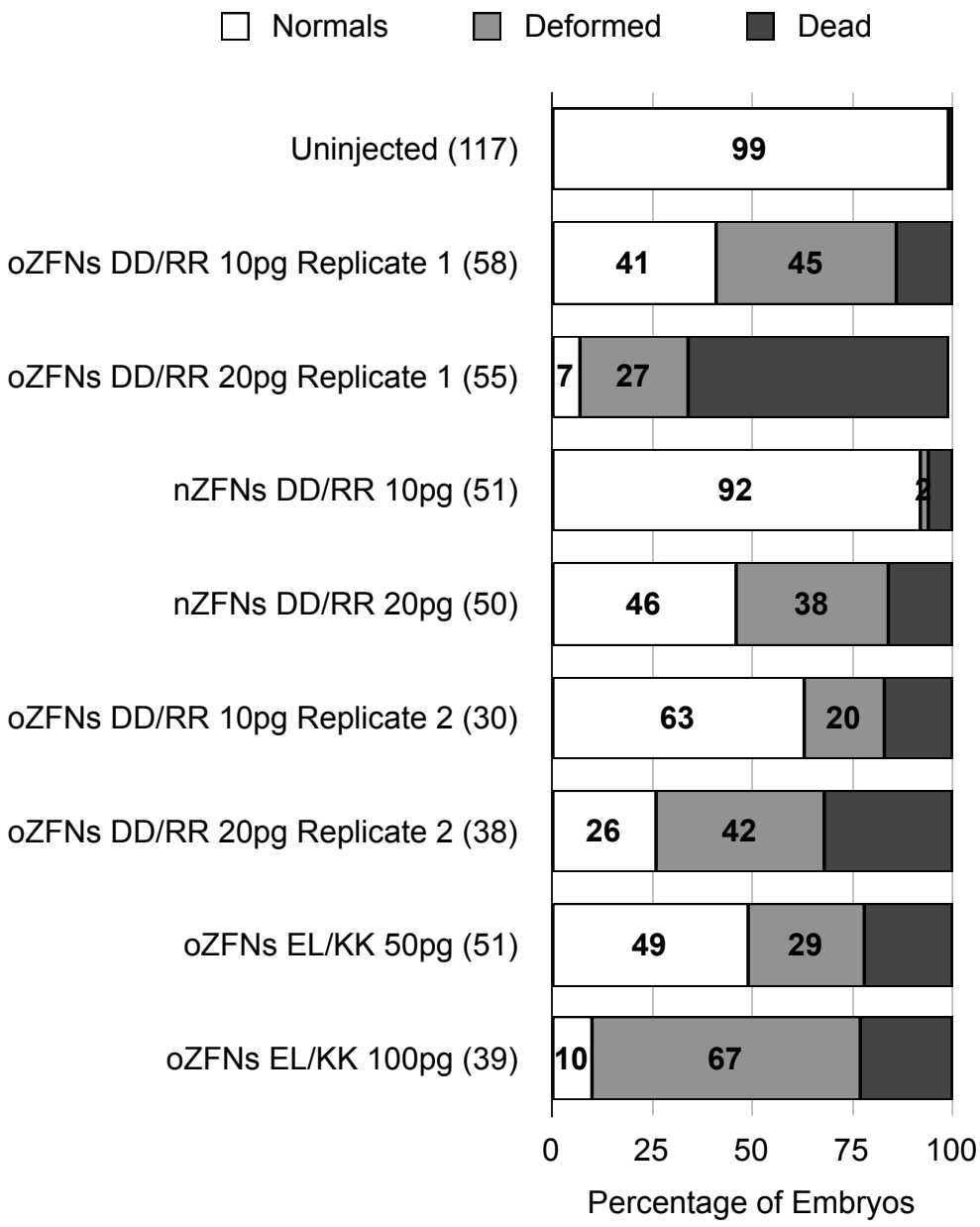


Figure 2-3: **Proportion of ZFN-treated embryos with different morphology at 24 hpf**. The number in parentheses represents the total number of embryos used for analysis. The activity of ZFNs containing two different sets of ZFPs were compared: the original ZFNs (oZFNs)[94] and new ZFNs (nZFNs) generated in this study.

Figure 2-4

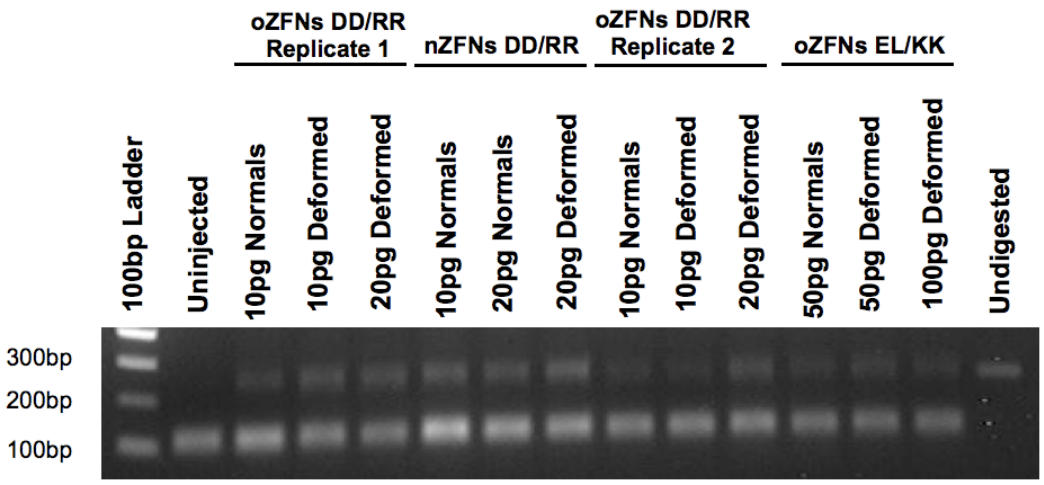


Figure 2-4: The RFLP analysis performed for ZFN-treated embryos. Genomic DNA from untreated or ZFN-treated embryos was pooled. The DNA flanking the *kdrl* target site was PCR-amplified[94] and digested with *Nsp*I enzyme which cleaves the unaltered spacer region within the ZFN target site resulting in two bands each ~110 bp in size. Indels within the spacer region at the target site result in the loss of *NspI* site, which is detected as the presence of undigested band of 220 bp in size.

Figure 2-5

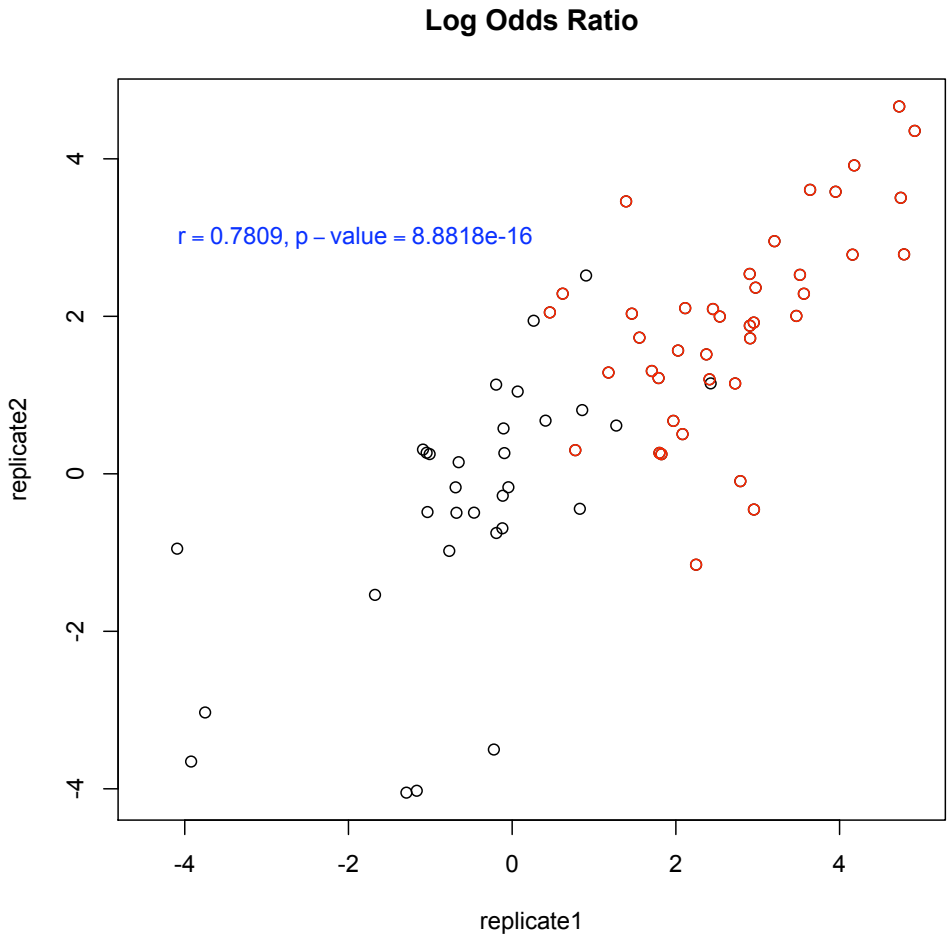


Figure 2-5: **High correlation between the replicates.** Pearson correlation analysis of the lesion log odds ratio for 41 identical sites between three identical treatment groups from different biological replicates. After excluding sites where the odds ratio was 0, there are total of 71 dots in the plot with each dot representing a pair of log odds ratios from replicate 1 and replicate 2 that have the same ZFN dose at the same site. The three treatment groups are Original ZFNs DD/RR 10pg Monsters, Original ZFNs DD/RR 10pg Normals and Original ZFNs DD/RR 20pg Monsters. A significant correlation exists between the replicates (r = 0.78, p-value= 8.9e-16) where the red dots indicate sites with a significant increase in lesion frequency in either replicate 1 or replicate 2 (BH adjusted p-value <0.05) as compared to the uninjected control.

The presence and frequency of lesions at each site was determined by Illumina-based sequencing of PCR amplicons spanning each genomic locus. On average approximately 8,000 reads per site were obtained, which allowed confident assessment of combined insertion and deletion (indel) frequencies ≥ 0.1%. Owing to the short read length of Illumina sequencing (~36bp), our analysis was limited to the detection of small insertions or deletions. Only indels that were >1bp in length were counted to avoid the bulk of the sequencing artifacts. To ascertain the consistency of the data, lesions at a subset of off-target sites were analyzed from a second independent biological replicate of the ZFN injections. Analysis of the site-specific lesion frequency between the biological replicates shows that they are significantly correlated **(Figure 2-5)**. The presence of indels at each site was considered significant only if the following criteria were fulfilled: a) indels occurred at a significantly higher frequency (Benjamini-Hochberg (BH) adjusted p-value < 0.05) in the injected sample relative to the uninjected control to account for noise in the sequencing data at some sites, which leads to a small fraction of sequences that appear to contain lesions even in the uninjected control[209]; b) indels constituted ≥ 0.1% of the sequence reads (in the average of the two replicates where available); and c) more than one different indel sequence was observed (to avoid potential jackpot effects). We believe these criteria constitute a conservative assessment of activity, and may assign sites as inactive that actually incur indels at a low frequency. Consistent with the RFLP analysis of the *kdrl* ZFNs, the lesion frequency at the target site was ~7% in normal embryos at the 10pg dose, which increased to ~15% at the 20pg dose **(Table 2-1)**.

Table 2-1

Name	Sequence of the on/off target site	Matches to the Target site	Spacer (bp)	Frequency of Indels (%)					
				10pg oZFNs DD/RR Normal	10pg oZFNs DD/RR Deformed	20pg oZFNs DD/RR Deformed	10pg nZFNs DD/RR Normal	50pg oZFNs EL/KK Normal	100pg oZFNs EL/KK Normal
Target	**CACACCTTC**AGCATG**TTGGTGGGA**	**18**	**6**	**7.3**	**11.7**	**15.8**	**7.4**	**9.4**	**8.5**
OT1	**TCCCACCcg**AGTCCT**GcAGGTGTG**	**15**	**6**	**1.6**	**4.5**	**3.4**	**0.3***	**ND**	**ND**
OT2	**CACACCaTC**CTACCT**TTGGTGGGt**	**16**	**6**	**1.3**	**2.5**	**2.7**	**0.1**	**ND**	**ND**
OT3	**CACACCTTC**ACAGAC**gTGGgGGGA**	**16**	**6**	**0.8**	**1.5**	**2.6**	**1.4***	**0^**	**0.3**
OT4	**cCCgACCAg**ATTGT**GAAGGTGTG**	**15**	**5**	**0.4**	**0.4**	**1**	**0.2***	**0.1***	**0.2***
OT6	**aCCCACCgA**GATAC**GcgGGTGTG**	**14**	**5**	**0.3**	**0.5**	**0.7**	**0**	**0.0**	**0.2***
OT7	**CcCACCcTC**GTGATG**TTGGaGGGA**	**15**	**6**	**0.3**	**0.4**	**0.7**	**0**	**0.2***	**0.0**
OT8	**CACACCggC**AGACT**gcGGcGGGA**	**13**	**5**	**0.2**	**0.3**	**0.4**	**0**	**0.2***	**0.1***
OT9	**CACACCcaC**AAAAG**aTGGTGGGt**	**14**	**5**	**0.2**	**0.4**	**0.8**	**0.1**	**0^**	**0.0**
OT5	**TCCCACCcA**GGAAGT**GAtGGTGaG**	**15**	**6**	**0.2**	**0.4**	**0.8**	**0**	**0.1**	**0.1**
OT11	**TCCCACCgg**AGCGGT**GAtGGTGaa**	**13**	**6**	**0.2**	**0.2**	**0.6**	**0**	**0.0**	**0.1**
OT12	**TCCCgCCAA**CAAAT**GAcGGaGTG**	**15**	**5**	**0.1**	**0.5**	**0.8**	**0**	**0.1**	**0.1**
OT13	**CgCACCgcC**AGACAT**aTGGTGGGA**	**14**	**6**	**0.1**	**0.1**	**0.6**	**0**	**0.1**	**0.0**
OT10	**TCCCcCCtg**CCATGA**GgAGGTGTG**	**14**	**6**	**0.2**	**0.7**	**0.4**	**0.2**	**0.4***	**0.8***
OT14	**aCCCACCcA**CTACT**GAgGGTGaG**	**14**	**5**	**0.1**	**0**	**0.3**	**0**	**0.1**	**0.1***
OT15	**CACACCTcC**AATTA**gaGGcGGGA**	**14**	**5**	**0.1**	**0.3**	**0.1**	**0.1**	**0.1**	**0.1**
OT16	**TCCCtCCct**AAGGGT**GAtGGgGTG**	**13**	**6**	**0**	**0.2**	**0.3**	**0**	**0.1**	**0.0**
OT18	**CACACCagC**TGCATT**TTGGTGGGt**	**15**	**6**	**0**	**0**	**0.1**	**0**	**0.0**	**0.0**
OT20	**TtCCACCAA**GTATCA**GAAGGTGTa**	**16**	**6**	**0**	**0**	**0.1**	**0**	**0.0**	**0.1**
OT22	**TCCCACCAg**GATATCCGGGTTAC **GcAGGTGTG**	**16**	**14**	**0**	**0**	**0.3**	**0**	**0.0**	**0.0**

Table 2-1: Sequences and lesion frequencies for each ZFN pair and dose at the target and 19 active off-target sites. Off-target sites show significant lesion frequencies either for normal embryos (green) or only in deformed embryos (yellow) injected with original ZFNs^{DDRR} (oZFNs^{DDRR}). Off-target sites in orange displayed significant activity in only one of the biological replicates. The red off-target site has a 14 bp spacer separating the ZFP recognition sequences. Asterisks in the final three column indicate off-target sites that meet our significance criteria. Only one off-target site showed increased lesion frequency in nZFNs as compared to oZFNs-DD/RR (OT3 in Blue). "0^" in the 50 pg oZFN^{ELKK} column indicates two off-target sites which had insufficient sequencing reads to determine the lesion frequency.

Overall, only 19 off-target sites were "active" (i.e., displayed indels at a significant frequency based on the criteria above) even at the higher ZFN dose **(Figure 2-2)**. All of the examined homodimeric sites were inactive, which is consistent with previous studies indicating that the DD/RR nuclease domain suppresses activity at homodimeric sites[94,153]. In ZFN-treated embryos with normal morphology, only 8 of the 113 heterodimeric sites were active (across both biological replicates where available, **Figure 2-6 and Table 2-1**). Notably, all of these sites contain a 5 or 6 bp spacer between the two half-sites. At the higher ZFN dose an additional 4 off-target sites were actively cleaved. Moreover, 7 other off-target sites were found to be active in one of the two biological replicates. Since these 7 sites contained hallmarks of ZFN induced lesions (multiple types of lesions in the spacer region between the two ZFN half-sites), we included them in our analysis of active sites. One of these sites **(OT22 in Table 2-1)** contains a longer spacer (14 bp) between the ZFN binding sites. Among the examined off-target sites, those containing a 6 bp spacer were the most likely to be active, both based on the fraction of active sites (38%) and the indel rates at active sequences **(Figure 2-2 and Table 2-1)**. Off-target sites containing a 5 bp spacer were the only other group where multiple sites (23%) were actively cleaved. These results are consistent with a previous study indicating that ZFNs with a 'TGGS' linker connecting the ZFP and the nuclease domain are most active on target sites separated by a 6 bp spacing followed by sites with a 5 bp spacing, whereas sites with longer spacers are inefficiently cleaved[64]. With regards to the types of observed lesions 4 bp insertions, which represent a simple fill-in and religation of the 5′

Figure 2-6

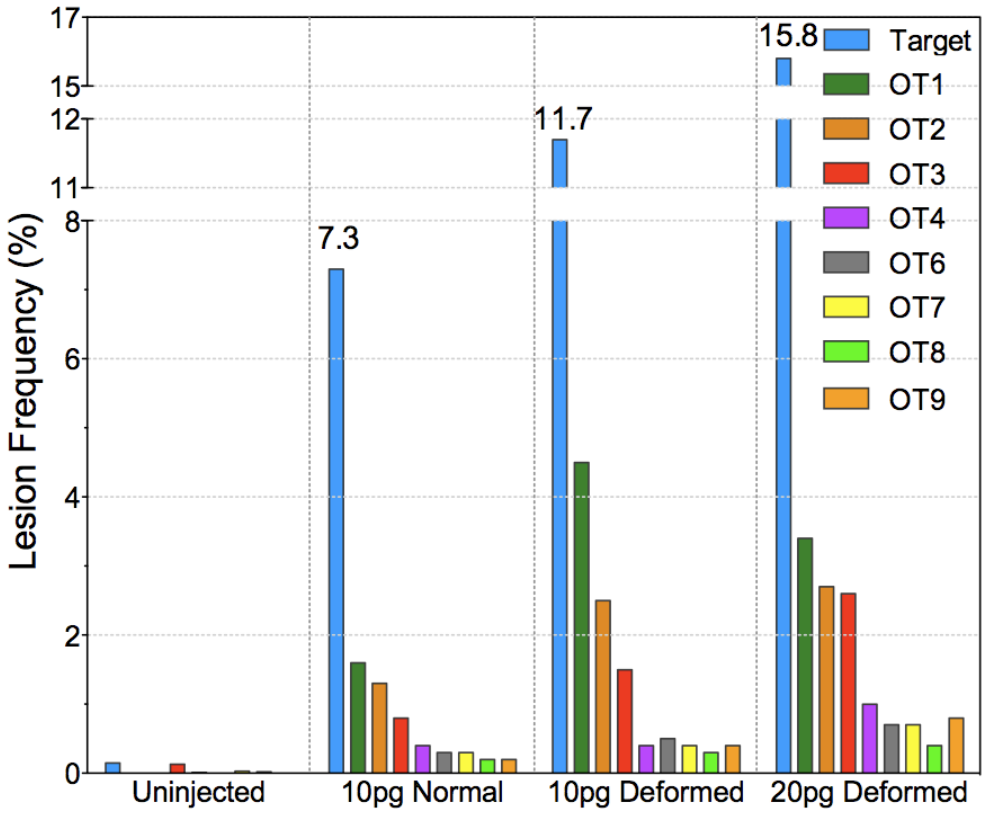


Figure 2-6: Dose dependent effects of *kdrl*-ZFNs on its *in vivo* activity and precision. The lesion frequency was plotted for the on-target (blue) and 8 off-target sites active in the morphologically normal embryos at the 10 pg dose.

overhangs generated by the *Fok*I nuclease domain, are the most common events **(Figure 2-7)**.

Concomitant with greater on-target lesion frequency, increasing the ZFN dose increased the degree of off-target cleavage within the genome **(Figure 2-6)**. However, preferential activity at the target site was maintained at both ZFN doses, as the on-target lesion frequency exceeded that of any off-target site by at least 4-fold. Notably, animals treated with the higher ZFN dose were more likely to be deformed suggesting that increased collateral damage within the genome may contribute to their abnormal development. Consistent with this hypothesis, normal embryos at the 10 pg dose displayed fewer active off-target sites than the deformed embryos at the 20pg dose (8 vs. 12 considering active sites in both the replicates or 12 vs. 18 considering activity in one replicate). Moreover, these normal embryos also exhibited significantly lower frequencies of off-target lesions at the 8 common active off-target sites with the median lesion frequency increasing from 0.6% in normal embryos to 1.5% in deformed embryos at the 20pg dose (p-value < 0.0001). Thus, increased off-target lesion frequency is associated with the presence of developmental abnormalities.

Common features of active off-target sites

We next sought to identify common characteristics of active off-target sites that distinguish them from inactive sites. Since active sites could simply share greater homology to the *kdrl* target sequence, we compared the total number of matches to the target site for active versus inactive off-target sequences containing a 5- or 6-bp gap

Figure 2-7

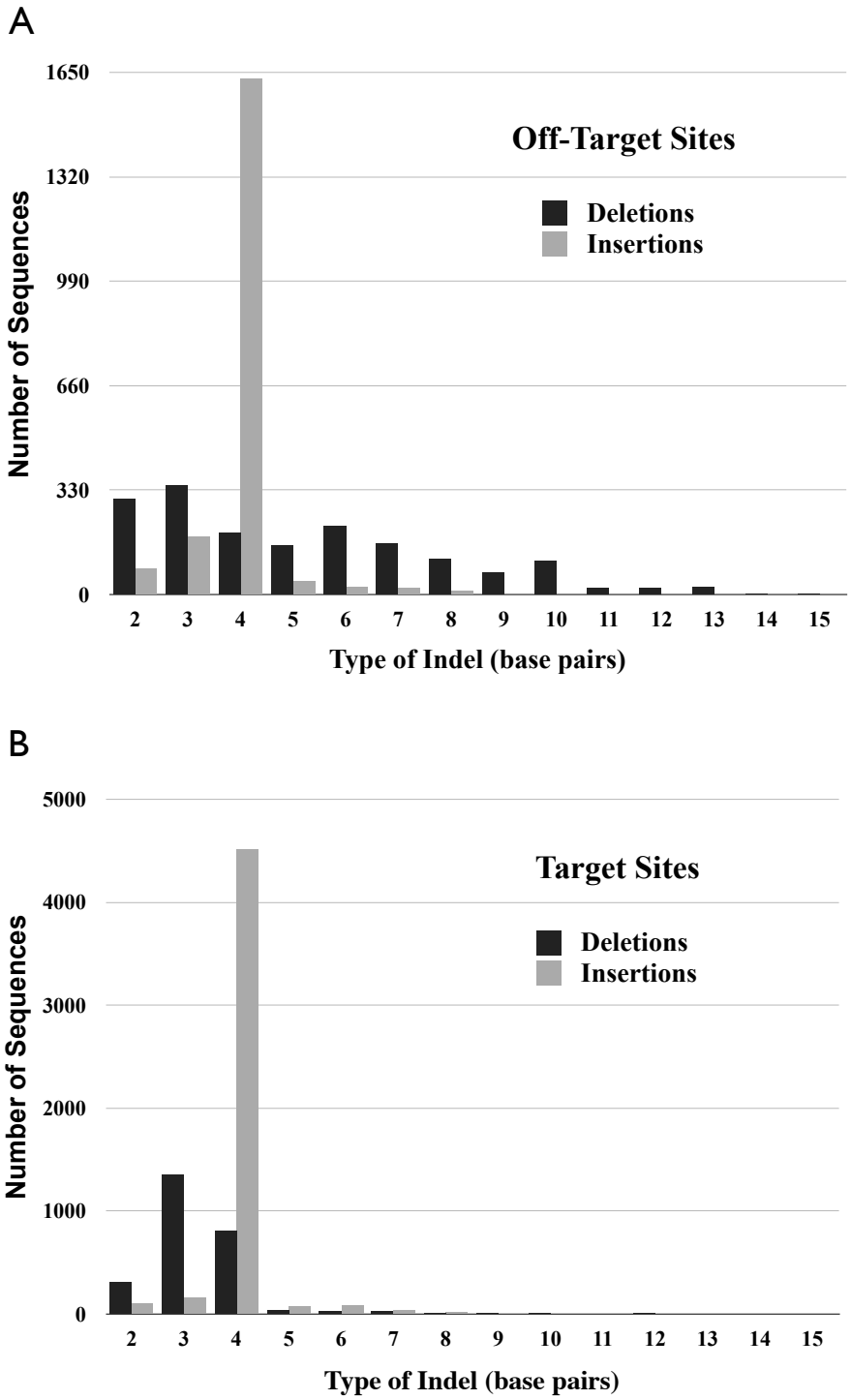


Figure 2-7: **The distribution of the type of indels observed at the off-target sites (A) and the target site (B).**

between the ZFN half-sites **(Figure 2-8)**. Surprisingly, there was no significant correlation between the degree of identity and off-target activity (p-value = 0.48). Furthermore, the distribution of active sites with regard to homology to the target site simply reflected the general distribution of all prospective off-target sites (τ = 0.89, p-value = 0.0367). Thus, in this population of sites that are highly similar to the *kdrl* ZFN target sequence, the degree of identity is not a defining feature of activity.

To better identify attributes that distinguish active from inactive off-target sequences, we constructed a frequency plot of the bases at each position in the sites from the active group **(Figure 2-9a)**. One striking characteristic of the active off-target sites is the complete conservation of a number of the guanines (7 of 10) in the composite ZFP recognition sequences. These positions are typically more diverse in the inactive sequences **(Figure 2-9a)**, suggesting that they represent critical features that define activity. Examining this trend in greater depth, we find that 15 out of 23 off-target sites with all ten 'G' contacts were active, whereas only 3 out of 36 off-target sites lacking one or more of these 'G' contacts were active **(Figure 2-9b)**. This correlation was highly significant (p-value = 5.6e-6). Using a B1H-based activity assay[92], we directly determined the importance of the 'G' contacts in the *kdrl*-ZFPL and *kdrl*-ZFPR recognition sequences by mutating each base independently to cytosine and assaying the effect on ZFP-dependent cell growth. Consistent with the *in vivo* data, we observed that mutation of any of the conserved Gs in the ZFPL or ZFPR binding site strongly reduced ZFP-dependent cell growth even at low stringency (1 mM 3-AT), where only the most important recognition positions should be detected **(Figure 2-10)**. Other positions within

Figure 2-8

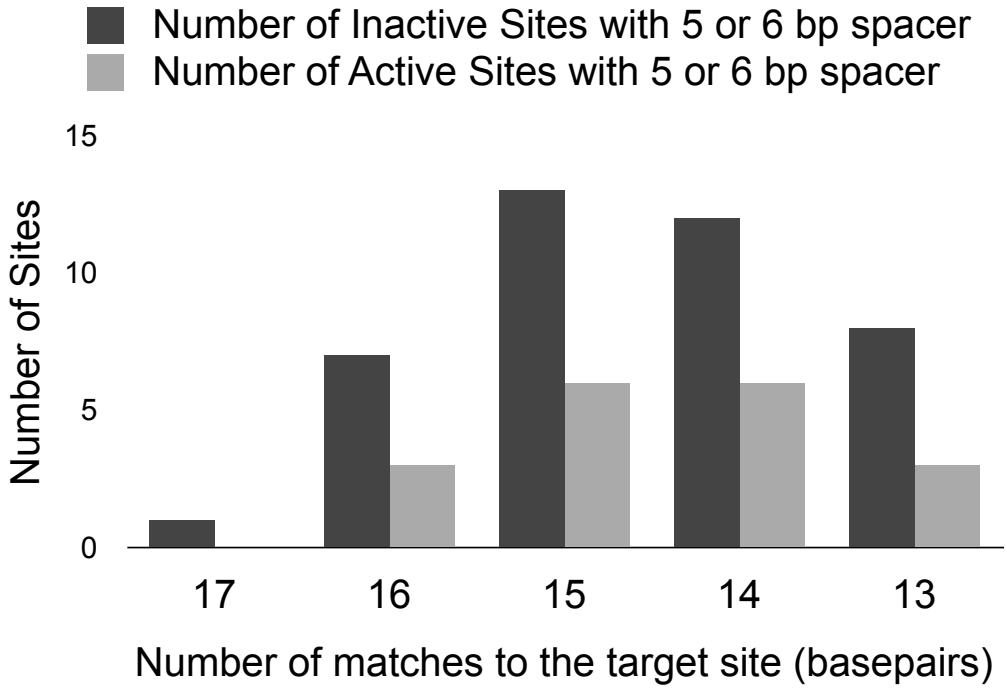


Figure 2-8: **Characteristics of active off-target sites**. The distribution of the number of matches to the target site for active (Grey) and inactive (black) off-target sites with 5- or 6-bp spacing is shown.

Figure 2-9

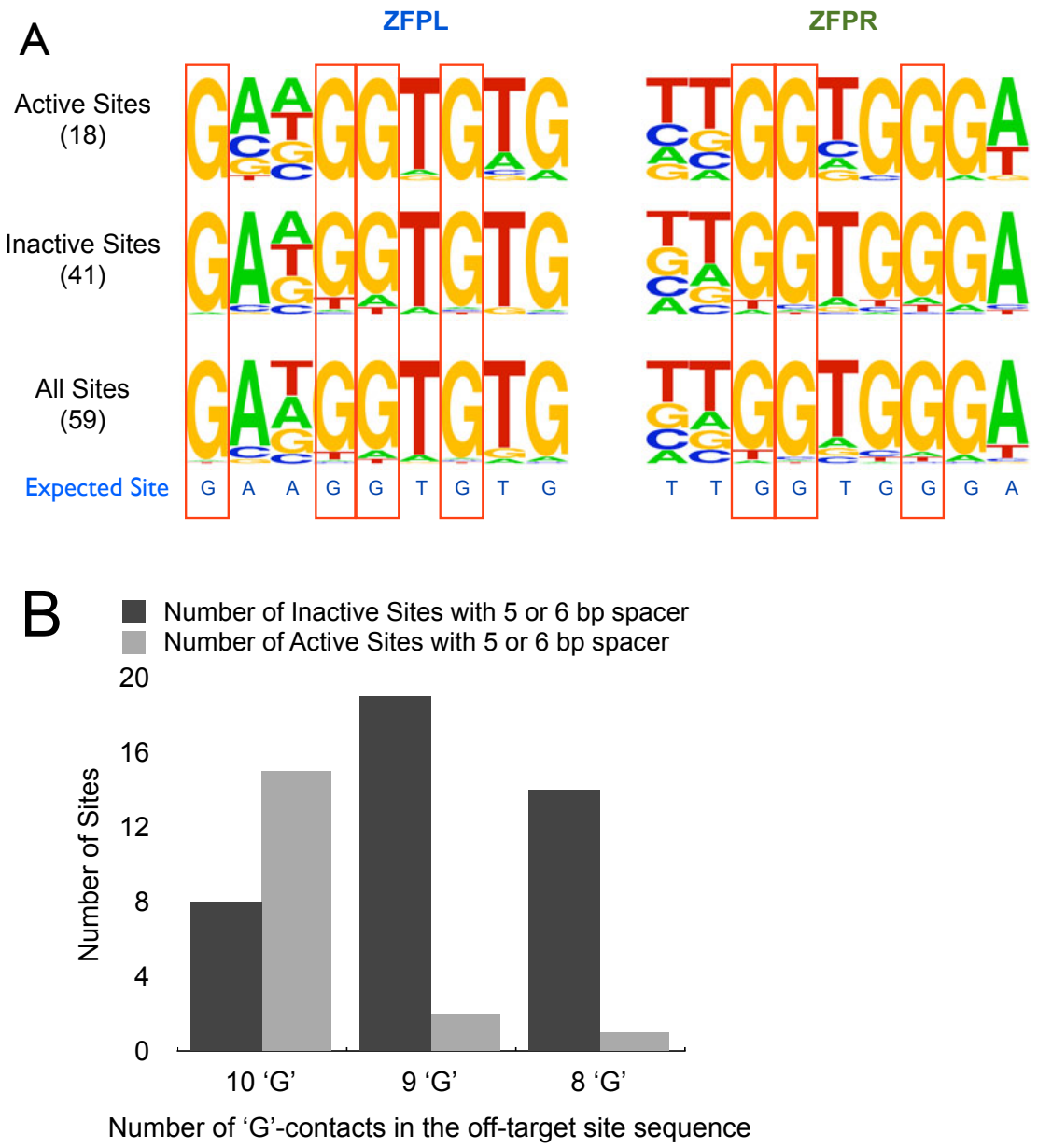


Figure 2-9: **Enrichment of Guanine-contacts in the active off-target sites**. (a) Base frequency at each position in the ZFPL and ZFPR binding sites are displayed as a logo for (top) the group of 18-active off-target sites, (middle) the group of 41-inactive off-target sites, and (bottom) all 59 sites together. Guanines at seven positions (red boxes) in the active off-target sites were absolutely conserved within the active sequences, but are more variable in the inactive sites. (b) The distribution of the number of active (grey) and inactive (black) off-target sites as a function of the number of guanines preserved in each recognition sequence.

Figure 2-10

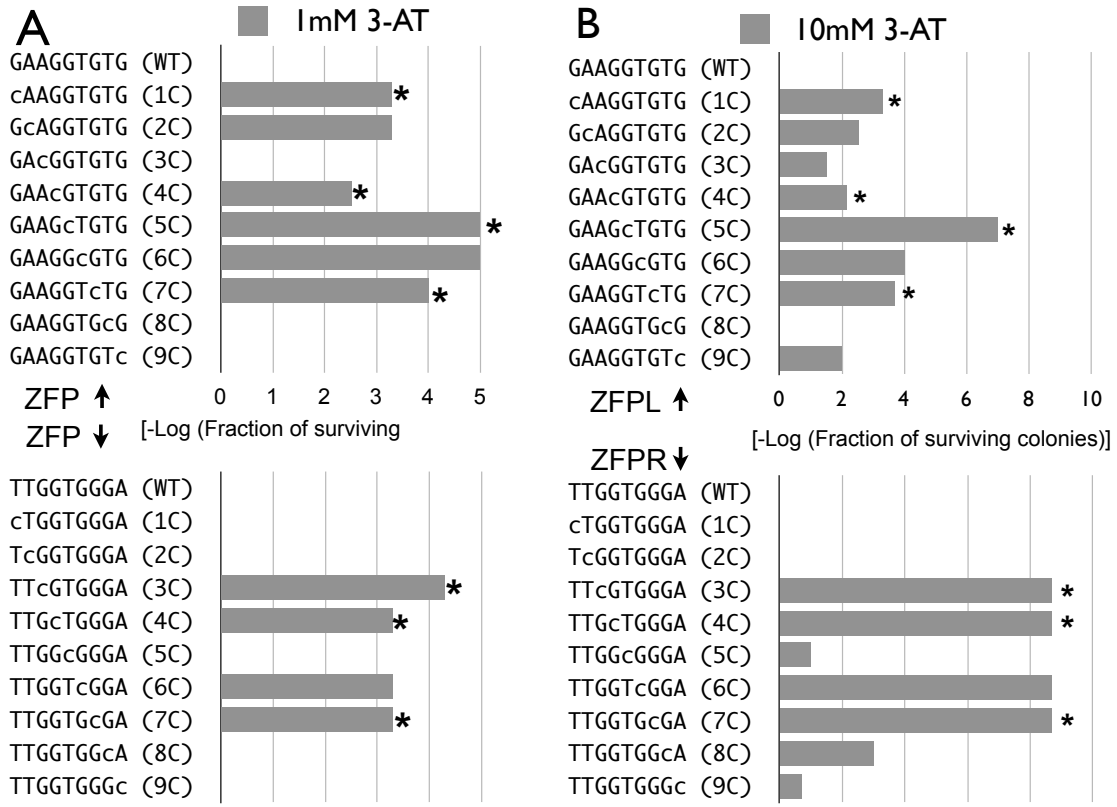


Figure 2-10: Bacterial-one-hybrid based analysis of importance of positions within each ZFP binding site. Each base in the binding site of the ZFPL and ZFPR was independently mutated to cytosine, which is not found at any position in either of the ZFP recognition sequences. Its influence on ZFP binding was assayed using B1H-based activity assay performed[92] at 1 mM 3-AT (a) to detect only the most important positions for recognition, and at 10 mM 3-AT (b) to detect other important positions for recognition, where a reduction in cell survival (plotted as the -log of surviving colonies) indicates a position important for recognition. All of the absolutely conserved guanines - indicated by an asterisk (*) - are critical for activity.

each ZFP binding site also influence recognition, however the impact of mutations at some of these positions is only detected at higher stringency within the activity assay **(10 mM 3-AT, Figure 2-10)**. Thus, for these ZFPs the conserved G contacts identified in this analysis appear to be necessary but not sufficient for efficient recognition of their sub-sites.

Reduced off-target effects when employing ZFNs with improved specificity

Having established a baseline of off-target events with our original *kdrl* ZFNs, we investigated the influence of the two distinct functional domains within the ZFN (the ZFP and the nuclease domain) on *in vivo* precision. We focused initially on further optimization of the *kdrl* ZFPs since improving their DNA-binding specificity would be expected to have the greatest impact on off-target events. Based on the determined specificity of these ZFPs **(Figure 2-1)**, each ZFP displayed a strong preference for the desired base pair at ~7 of 9 positions within their target site. However, in both ZFPs 2 positions within the C-terminal finger (finger-3) recognition site were relatively poorly specified, which became our focus for improvement. To identify ZFPs with improved specificity, additional clones generated from our original B1H selections were characterized using B1H-based binding site selections followed by sequencing of binding sites from a few surviving clones[94]. This yielded an improved clone (nZFPL) for the left recognition site, but no obviously improved clone was identified for right recognition site. Instead, a modestly improved clone (nZFPR) was generated by introducing previously defined specificity determinants that are compatible with T recognition at

positions 3 and 6 of the finger-3 recognition helix[86]. Subsequent efforts to reselect finger 3 of the ZFPR in a different context yielded an identical finger sequence (RSDALRK) (Zhu, C., Lawson and Wolfe, *unpublished results*). Comprehensive binding motifs for the new ZFPs (nZFPs) were determined by B1H binding site selections followed by Illumina sequencing. More than 1000 unique sequences were used to generate each recognition motif. Comparison of the recognition motifs indicates improvements in the specificity of both nZFPs **(Figure 2-11)**. nZFPL displays a dramatic increase in the preference for adenine at position 2 (rising from 56% to 99%) and position 3 (rising from 13% to 36%) within the recovered sequences. Likewise, nZFPR displays a modest increase in the preference for thymine at position 2 (rising from 42% to 67%) within the recovered sequences.

Although the improvements in specificity of the nZFPs appear modest, we assessed whether these differences would translate into improved ZFN precision *in vivo*. We compared the *in vivo* activity and toxicity of the nZFNs (incorporating the new ZFPs) with original ZFNs (oZFNs). mRNAs (either 10 or 20 pg dose) encoding each set of ZFNs were injected into zebrafish embryos. After 24 hpf, treated embryos were scored as morphologically normal or deformed. The nZFNs displayed markedly lower toxicity: ~45% of the nZFN-treated embryos displayed normal morphology at the 20 pg dose whereas only ~17% of the oZFNs-treated embryos were normal **(Figure 2-3)**.

We reasoned that the reduced toxicity of the nZFNs was a consequence of decreased off-target cleavage. Therefore, we compared off-target lesion frequencies at the same 141

Figure 2-11

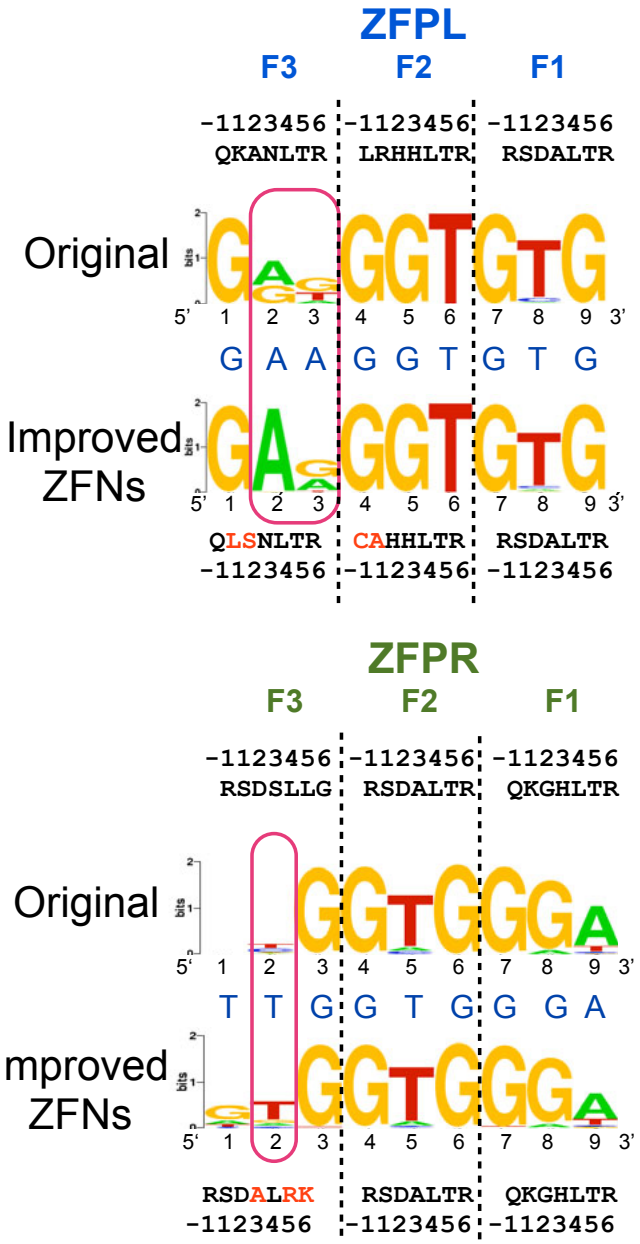


Figure 2-11: Improved binding site specificities of the new ZFPs. (a) Binding site specificities of the new and old ZFPs determined using the B1H system are displayed as Sequence logos (Schneider et al sequence logo and Weblogo reference). The recognition helix sequences for each finger are displayed where the amino acids that differ in the nZFPs are indicated in red. Red rectangles highlight the positions where information content of the desired base was higher in the improved ZFNs.

sites characterized for the oZFNs in genomic DNA isolated from embryos treated with the nZFNs. The nZFNs, at a dose of 10 pg, showed an on-target lesion frequency of ~7.4% which was similar to that observed with an analogous dose of the oZFNs. Notably, even with similar on-target activity, nZFNs displayed significantly lower rates (p-value < 0.0001) of off-target cleavage at the majority (7 out of 8) of the active off-target sites for the oZFNs in normal embryos **(Figure 2-12 and Table 2-1)**. Among the 59 heterodimeric off-target sites with a 5 or 6 bp spacer, only 3 displayed lesions at a significant frequency based on our criteria **(Table 2-1)**, which represented a reduction compared to the 8 active sites for the oZFNs. Only one off-target site (OT3) showed an increase in the lesion frequency with the nZFNs. This may be due to the presence of a 5' guanine in the nZFPR OT3 half-site, as the nZFP recognition motif indicates a slight preference for 'G' at this position in the recognition sequence, which is absent in the oZFPR recognition motif. Thus, based on this analysis even a modest improvement in ZFP specificity can result in dramatic reduction in ZFNs promiscuity.

Examining the influence of the nuclease domain variant on ZFN promiscuity

Although the primary determinant of ZFN specificity is the incorporated ZFP, there is ample evidence that the nuclease domain can also influence the cytotoxicity of ZFNs[153,154]. Consequently, the influence of the *Fok*I nuclease dimerization interface on ZFN activity and precision *in vivo* was investigated. We compared the on- and off-target activity of the original *kdrl* ZFNs containing the engineered heterodimeric DD/RR nuclease domains ($ZFNs^{DDRR}$) to the same ZFPs fused to the heterodimeric EL/KK

Figure 2-12

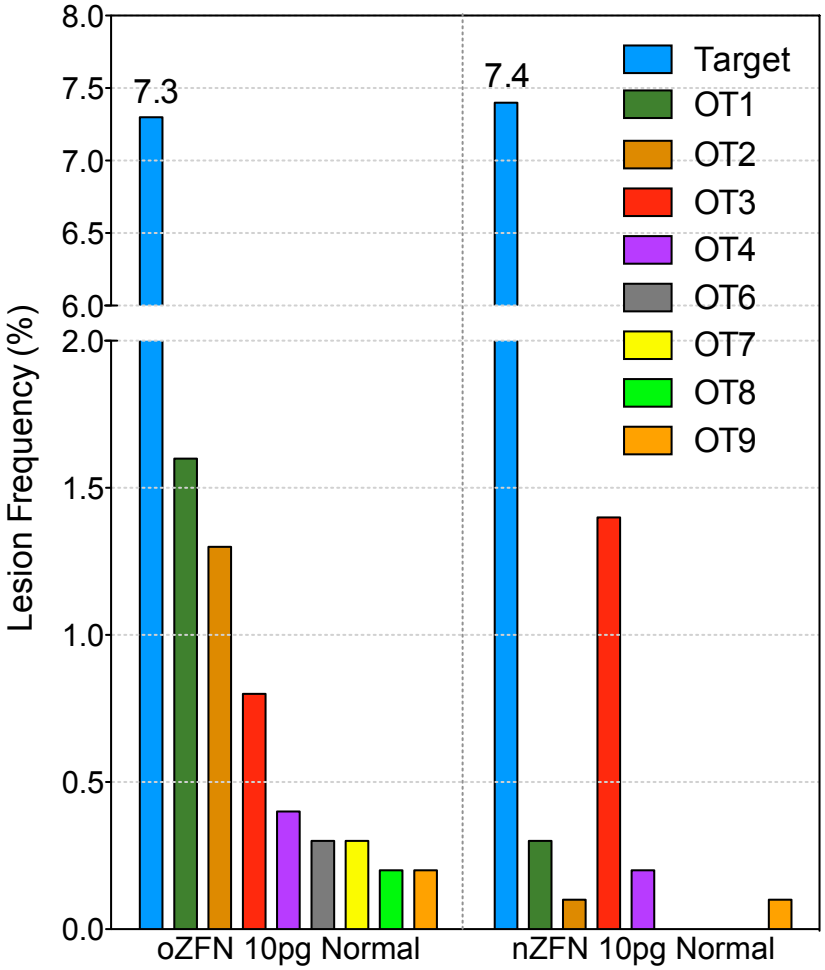


Figure 2-12: **The specificity of the ZFP domains influences the precision of ZFNs.** Comparison of lesion frequencies at the on-target site and the 8 active off-target sites were plotted for oZFN- and nZFN-treated embryos. The color scheme remains same as in Figure 2. Lesion frequency at the off-target sites in nZFN treated embryos was reduced at all but one off-target site.

versions of the *Fok*I nuclease domain ($ZFNs^{ELKK}$)[154]. Although both nuclease variants have been successfully used on chromosomal targets *in vivo*[153,154], there has not been a detailed study comparing their activity and their potential influence on ZFN specificity *in vivo*. Notably, we found that the $ZFNs^{ELKK}$ had a markedly lower activity such that injection of five to ten times more mRNA (50 pg and 100 pg doses) was required to achieve on-target lesion rates similar to the $ZFNs^{DDRR}$ **(Figure 2-4)**. Consequently we performed lesion analysis for the EL/KK ZFNs at these higher doses, where we examined a subset of the previously characterized off-target sites (96 out of 141). Unexpectedly, we found that the $ZFNs^{ELKK}$ displayed reduced off-target lesion frequencies compared with the $ZFNs^{DDRR}$ **(Figure 2-13)**. At previously defined active off-target sites, normal embryos treated with 100 pg of $ZFNs^{ELKK}$ displayed a significantly lower average off-target lesion frequency (0.13%) than normal embryos treated with 10 pg $ZFNs^{DDRR}$ (0.37%, p-value < 0.0001). Only one other off-target site **(OT10, Table 2-1)** displayed significant lesions in the $ZFNs^{ELKK}$ treated embryo. Notably, for this site all 10 of the target site guanines are retained. Thus, for the original *kdrl* ZFNs, the choice of the engineered nuclease domain has a surprising impact on the ratio of on-target to off-target lesions *in vivo*.

DISCUSSION

Although ZFNs have been used to create genetically engineered organisms[210,211] and initial clinical trials employing them as therapeutics are underway[113,201,212], the characterization of ZFN-induced collateral damage to the genome of treated cells has

Figure 2-13

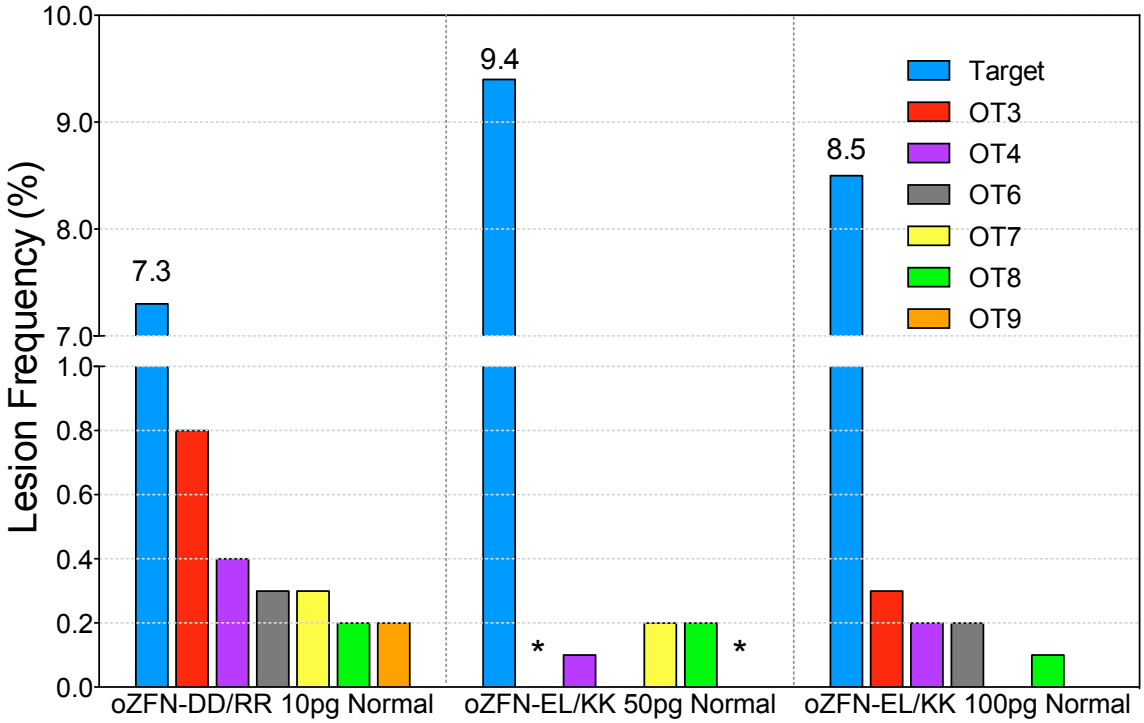


Figure 2-13: **Influence of the type of the engineered nuclease domain (DD/RR[153] or EL/KK[154]) on the precision of the original ZFNs.** The lesion frequencies for normal embryos treated with 10 pg dose of oZFNs^{DDRR}, 50 pg dose of oZFN^{ELKK} or 100 pg dose of oZFN^{ELKK} were plotted for the on-target site and a subset (6 of 8) of the active off-target sites for oZFNs^{DDRR} that were assayed in this experiment. Asterisks indicate 2 positions in the 50 pg oZFN^{ELKK} where there were insufficient sequencing reads to provide a confident assessment of the lesion frequency (none were observed). Normal embryos treated with 50 pg of oZFNs^{ELKK}, like the 100 pg of oZFNs^{ELKK} embryos, displayed a significantly lower average off-target lesion frequency (0.1%) than normal embryos treated with 10 pg oZFNs^{DDRR} (0.3%, p-value < 0.0001) among the active off-target sites with sufficient number of sequencing reads.

been limited primarily to indirect assays of toxicity[168] and DSB foci[113,159] or lesion analysis at a small number of potential off-target sequences[114,174]. In this study, we have performed the most detailed analysis to date of the off-target effects of ZFNs by characterizing lesion frequencies at 141 potential off-target sites from the genomes of ZFN-treated zebrafish embryos. Using the *kdrl* ZFNs as a model, we show that the B1H-selected three-finger ZFNs preferentially cleave their target site to any assayed off-target site and thus, are sufficient for relatively precise gene modification. We also probed the influence of the components of *kdrl*-ZFNs on their precision. Surprisingly, both the choice of the nuclease domain and the specificity of the component ZFP domains dictate the accuracy of these ZFNs.

Not unexpectedly, the thermodynamics of DNA recognition appear to dominate the impact of binding site mutations on ZFN activity. Simply assessing the likelihood of ZFN activity at an off-target site based on the number of matches to the target sequence was a poor predictor, as evidenced by the absence of correlation within the data for our three-finger ZFNs **(Figure 2-8)**. Off-target sites with as many as five mismatches to the target site contained indels at a statistically significant frequency, whereas other sites with just one or two mismatches were inactive. Data from binding site selections provides a much better metric for defining critical positions for recognition. The relative importance of individual positions within each ZFP binding site was initially defined by our high stringency B1H binding site selections **(Figure 2-1)**, which provided a consensus recognition sequence for each ZFP. The most critical positions were identified using the low stringency B1H activity assay, where we could examine the importance of

individual positions by mutating them independently **(Figure 2-10)**. In principle, information on the most critical positions could also be obtained through B1H binding site selections performed at low stringency. In the case of *kdrl* ZFNs, the preservation of a subset of the arginine-guanine interactions in off-target sites was strongly correlated with ZFN activity at these sequences. Arginine-guanine interactions are typically important specificity determinants at the zinc-finger-DNA interface: abrogating similar contacts in the Zif268 recognition sequence results in a 100- to 400-fold decrease in its binding affinity[213]. Based on these observations, we speculate that engineering ZFNs with specificity determinants that distribute the binding energy more uniformly over the entire recognition sequence - instead of employing a few critical Arginine-Guanine contacts - will result in ZFNs with improved functional specificity. Achieving this goal may require increasing the number of fingers per ZFP as well as the use of appropriate linkers to attenuate ZFP affinity[59], a hallmark of many of the ZFNs currently employed by Sangamo BioSciences[113,114]. However, a recent report demonstrated that subsets of fingers in a polydactyl ZFP can bind independently to target sites in the genome thus decreasing their on-target activity and possibly increasing their off-target effects[214].

The influence of ZFP specificity on the *in vivo* activity and toxicity of ZFNs was first demonstrated by Cornu *et al*. where they compared the activities of ZFNs containing modularly-assembled ZFPs to ZFNs containing ZFPs selected for the identical target sequences[159]. The selected ZFPs displayed higher specificity as measured by the ratio of the affinity of each ZFP for its target site relative to bulk non-specific DNA. When incorporated into ZFNs, the resulting nucleases generally showed higher activity and

lower toxicity in human cells than the nucleases containing their modularly-assembled counterparts. In this study, we have performed a more in-depth analysis by defining the base-preferences at each binding site position for the employed ZFPs, which, unlike the bulk specificity, provides information about key sequence features likely to be shared by potentially active off-target sites. This information coupled with a broad assessment of the frequency of ZFN-induced lesions at a number of off-target sites in the genome of zebrafish embryos reveals that even modest changes in the ZFP specificity can decrease off-target activity leading to improved functional specificity and reduced toxicity. Thus, detailed specificity analysis of candidate ZFPs provides not only an estimate of key sequence features of potentially active off-target sites but also an assessment of the relative fitness of the candidate for utilization in ZFNs. In cases where the DNA-binding specificity is sub-optimal, this information can be employed for focused optimization of suspect specificity determinants to obtain ZFPs with higher specificity and superior *in vivo* performance.

Surprisingly, in addition to the influence of the ZFP specificity on ZFN activity, we observed that the type of the engineered nuclease domain influences ZFN precision. We examined the influence of two pairs of *Fok*I variants DD/RR and EL/KK on ZFN activity, both of which favor heterodimerization over homodimerization[153,154] and display lower *in vivo* toxicity. Although, Miller *et al.* reported that ZFNs incorporating these engineered *Fok*I nuclease variants show two- to three- fold less activity than the wild type domain, there has been no detailed study comparing their relative precision. In fact, conflicting data exists regarding the precision of these engineered nucleases. Kim *et al.*

found that only the DD/RR nuclease variant appeared to reduce cellular toxicity relative to the WT nuclease domain[161]. In our study the *kdrl* ZFNs harboring the EL/KK variant ($ZFNs^{ELKK}$) consistently show lower activity than the ZFNs harboring the DD/RR nuclease ($ZFNs^{DDRR}$). Consequently, a five-fold higher dose of $ZFNs^{ELKK}$ was required to obtain an on-target lesion frequency similar to the $ZFNs^{DDRR}$. This result differs from a recent report by Guo *et al.* that the EL/KK-containing ZFNs are more active than DD/RR-containing ZFNs on an integrated target in 293 cells[156]. We cannot explain this discrepancy, however our observation of reduced activity for the EL/KK variants in zebrafish has been confirmed for a number of other ZFNs targeting different genomic loci (Smith, T., Wolfe, S. & Lawson N., *unpublished observations*) and in the recently published study[155]. For the *kdrl* ZFNs, even though EL/KK variants were injected at an elevated dose the toxicity of $ZFNs^{ELKK}$ and $ZFNs^{DDRR}$ was similar **(Figure 2-3)** and genomic analysis confirmed that the $ZFNs^{ELKK}$ generate fewer off-target lesions than the $ZFNs^{DDRR}$. The decreased activity and toxicity of the $ZFNs^{ELKK}$ could be the result of lower dimerization potential for the EL/KK nuclease domain, which would reduce the degree of cooperative binding between the two EL/KK monomers[155]. As a result, stronger interactions between each ZFP monomer and its binding site would be required to achieve residence times necessary to generate a DSB. Reduced cooperativity has been previously proposed as an explanation for the decreased toxicity of EL/KK variant as compared to the wild type nuclease domain[154]. However, the reduced toxicity of EL/KK variant in this study could have been associated with its limited homodimeric activity. By directly comparing the EL/KK and DD/RR nuclease variants, neither of which

displays significant homodimeric activity based on our analysis, it is readily apparent that the cooperativity between the nuclease monomers is an important feature of ZFN activity. These results suggest that further reduction in the dimerization potential of the nuclease domain coupled with specific zinc fingers with distributed binding affinity may lead to additional improvements in the precision of ZFNs. Recent studies have identified mutations at the dimerization interface that when coupled with DD/RR or EL/KK nuclease variants can increase the dimerization energy thereby enhancing the activity of these nucleases[155]. Moreover, mutations in the cleavage domain of the *Fok*I nuclease have been identified that increase its activity[156]. However, the off-target activity of these nuclease variants with enhanced activities has not been carefully examined.

Another factor that may influence the off-target activity of ZFNs is the chromatin structure around the potential off-target loci. A recent study demonstrated that the ZFN activity can also be influenced by the chromatin structure around the target site[186]. Therefore, similar to the on-target activity, off-target activity for ZFNs would be influenced by the chromatin structure making it difficult to predict the potential off-target sites based on just the binding site specificity data for ZFNs. Including the DNaseI hypersensitivity data and the Micrococcal nuclease hypersensitivity data might help develop better prediction models for ZFN on- and off-target activity.

ZFNs have been used to create genetically engineered organisms like zebrafish and rats where generating gene modifications with conventional homologous recombination based methods has not been feasible. We and others have shown that these genetic

modifications created using ZFNs can be transmitted through the germline. However, the degree of germline transmission of off-target lesions is an unaddressed concern for these ZFN-modified animals. To assess this possibility, we outcrossed one founder fish generated in Meng *et al.* and examined the progeny for the presence of lesions at the active off-target sites identified in this report[94]. Although we found lesions at the target *kdrl* site in ~50% of 35 offspring analyzed, we did not find evidence of lesions at any of the off-target sites (data not shown). This result, although merely representing a single founder, suggests that using ZFNs generated via B1H-based selections, one can obtain lines of genetically engineered animals relatively free of background mutations without the need for extensive outcrossing of founder animals.

Although this is the still the most detailed study where off-target activity of ZFNs has been examined in an organism but numerous questions remain to be addressed. One of the key limitations of this study is its characterization of ZFNs specific for a single target sequence. Although this study has improved our understanding of the activity of ZFNs within the genome, further analysis of the activity of other ZFNs pairs will allow a more comprehensive understanding of ZFN activity *in vivo*. This study is also biased by our choice of genomic sites for analysis based on the characterized specificity of our ZFPs. A more comprehensive survey of active ZFN targets could be obtained by performing a genome-wide analysis of ZFN occupancy using ChIP-seq in combination with lesion analysis, which might identify classes of active target sites (such as alternate spacings or registers of binding) that were uncharacterized in our survey. Ultimately, understanding the parameters that influence the precision of ZFNs *in vivo* will lead to improved designs

facilitating the ease of creating genetically modified organisms as well as improved therapeutics for gene therapy.

Comparison to recently published off-target studies

Recently, two studies were published that characterized off-target cleavage of zinc finger nucleases in the K562 cell line[184,185].

Gabriel *et al.* following the ZFN treatment, mapped the integration sites of the co-transfected IDLV-DNA (Integrase deficient lentivral virus) by (nr)LAM-PCR (non-restrictive linear amplification-mediated PCR) coupled with deep sequencing to identify genomic locations of off-target activity for their CCR5 and IL2RG ZFNs[185]. Pattanayak *et al.* on the other hand, first identified potential off-target sites for CCR5 and Vegf ZFNs from a library of ZFN sites using an *in vitro* selection system that identifies sequences that can be functionally cleaved at a specific ZFN concentration and then determined lesion frequency in K562 cells at the *in vitro*-identified potential off-target sites that were also present in the human genome[184]. In general, results from both the studies were in concordance with our findings. The off-target activity for ZFNs was lower than their on-target activity and increased with the ZFN dose. Although, the active off-target sites showed high homology to the target site but gross matches to the target site was not sufficient to predict off-target sites for ZFNs corroborating our findings. Although, both studies identified off-target sites for CCR5, not many active off-target sites were common between the two studies suggesting that identification by both methods are inefficient, which may result from the inherent limitations of the methods they employed. The

IDLV-integration method used by Gabriel *et al*. relies on the NHEJ-mediated integration of the IDLV DNA into the site of the double strand break that may represent only a fraction of total NHEJ events resulting in low sensitivity of this method due to which it may not capture off-target sites that show low but significant activity[185]. Pattanayak *et al*. did not use a completely randomized library for their *in vitro* selection method to restrict the library size and the ZFN activity *in vitro* may differ from its activity *in vivo*[184]. Thus, neither of these methods provides a complete assessment of off-target activity *in vivo* suggesting that the development of a better method to detect off-target activity for ZFNs *in vivo* would have useful application.

METHODS

Zebrafish husbandry

Zebrafish adults and embryos were handled according to standard method[215]. These studies were approved by the UMass Medical School IACUC. The wild-type line used in this study (referred to as Crawfish) was established through several incross generations of wild-type fish originally obtained from Scientific Hatcheries.

ZFN mRNA injections and on-target lesion analysis

ZFPs were cloned into the pCS2 vector containing either DD/RR (R487D (DD) and D483R (RR))[153] or EL/KK (Q486E; I499L (EL) and E490K;I538K (KK))[154] variants of *Fok*I nuclease as described[94]. pCS2-ZFN constructs were linearized with *Not*I enzyme and mRNAs were transcribed using the mMessage mMachine SP6 kit (Ambion) followed

by DNAse treatment. ZFN mRNAs were injected into one-cell-stage zebrafish embryos according to standard methods[215]. ZFN-induced on-target lesions at the *kdrl* locus were detected by *Nsp*I digestion as described previously[94].

Identifying ZFPs with improved specificity

The DNA-binding specificity of additional clones obtained from B1H-selections for ZFPL in Meng *et al.* were previously characterized using the 28 bp randomized library via omega-based B1H-selections[94]. The improved ZFPR clone was generated by design incorporating specificity determinants into finger 3 that would be compatible with the desired DNA-binding specificity[86] and its binding specificity was characterized with the B1H system using the 28 bp randomized library. Binding sites from a few surviving clones were sequenced and motifs were generated using MEME[216]. Clones (nZFPs) for ZFPL and ZFPR showing improved specificity over the original ZFP (oZFPs) were used for further analysis.

Selection of Off-Target Sites Stage I

The zebrafish genome (Zv7 repeat masked) was scanned using a perl algorithm to identify off-target sites containing half-sites similar in sequence to the determined binding site specificities of the two ZFPs (oZFPL: GANGGTGTG; and oZFPR: NNGGTGGGA where N allows all bases) in proper orientation with either 5- or 6-bp spacing between the two half-sites. The sites were ranked based on the number of matches to the target site with a score of 15 matches being the maximum. Heterodimer off-target sites that match the target site at 14 of 15 positions were chosen for analysis

(*kdrl* exon 2 is the only 15-of-15-bp match). Homodimeric sites were derived from sites that match either the 'GAXGGTGTG' composite site at 14 or 15 out of 16 bp or the 'XXGGTGGGA' composite site at 14 of 14 bp. For the pilot scale analysis, a total of 20 heterodimeric and 28 homodimeric sites were chosen. The details for these sites are provided in the online table and are marked as "**Off-Target Sites I**" (online table available at http://nar.oxfordjournals.org/content/39/1/381/suppl/DC1).

Selection of Off-Target Sites Stage II

Computational analysis was performed to bin additional off-target sequences (identified as described above) based on the number of conserved guanines (maximum = 10) within the potential off-target sites. A total of 47 sites that contain all 10 guanines were chosen for analysis with a range of total base matches to the target site (13 to 16 for 5- and 6-bp spacing sites and 14 to 17 for 14-, 15- and 16-bp spacing sites). Another 47 sites that are missing one or more guanines were chosen for analysis with a range of total base matches to the target site (13 to 16 for 5- and 6-bp spacing sites and 14 to 17 for 14-, 15- and 16-bp spacing sites). Both groups are designated in the **Online Table**[a] ("**Off-Target sites II – 10g**" & "**Off-Target sites II – non-10g**"). One off-target site is identical between the sites analyzed in Stage I and Stage II.

Solexa data analysis for off-target site lesions

36bp sequence reads from the Illumina run (both stage I and stage II runs) were binned to different ZFN treatments based on the barcode sequence. For each ZFN treatment,

[a] Online Table available at http://nar.oxfordjournals.org/content/39/1/381/suppl/DC1

sequences for different off-target sites and the target site were classified using a unique 9bp "prefix" following the adapter sequence **(Online Table[a])**

For each off-target site, insertions or deletions in the spacer region were defined based on the distance between the 9bp "prefix" at the 5' end of each off-target site and a 6 bp (8 bp in one case) "suffix" at the 3' end of each off-target site, where a more proximal suffix was employed to identify insertions and a more distal suffix for deletions. In some cases single nucleotide polymorphisms were present within the suffix sequences requiring a more relaxed suffix sequence definition. If the distance between the prefix and any suffix pair in each sequence matched the expected distance these sequences were binned as "correct (W)", where a secondary distal suffix was also employed to identify sequences of the appropriate length. Distances that were greater than expected were binned as "insertions (I)", and distances that were shorter were binned as "deletions (D)" with the exception that 1bp insertions or deletions were ignored because of the noise in the sequencing data associated with 1bp frameshifts in sequences evident in uninjected samples. Reads that did not contain the suffix sequence were marked as undefined (U). This analysis will miss long insertions or deletions that alter either the prefix or suffix but it is robust to the bulk of sequencing errors yielding high-confidence indels. The number of sequencing reads that are correct and the number of reads containing indels (insertions plus deletions) at each analyzed site for each ZFN dose were computed for the subsequent statistical analysis.

---

[a] Online Table available at http://nar.oxfordjournals.org/content/39/1/381/suppl/DC1

All statistical analyses were performed using R, a system for statistical computation and graphics[217]. The lesion frequency and its 95% confidence interval for each off-target site and the target site within each treatment were estimated based on a binomial distribution. The Fisher Exact Test was applied to assess whether there is a significant difference between each individual ZFN treatment and the uninjected control in the indel rate for the on-target and off-target sites. The odds ratio and its 95% confidence interval were computed for each ZFN dose using the fisher.test function based on conditional maximum likelihood estimation. To adjust for multiple comparisons, p-values were adjusted using the Benjamini-Hochberg (BH) method[209].

Criteria for defining an active off-target site

An off-target site was considered active only if the following criteria were fulfilled: a) indels occurred at a significant frequency in the injected sample relative to the uninjected control (BH adjusted p-value < 0.05)[209]; b) indels constituted $\geq 0.1\%$ of the sequence reads in the average of the two replicates (when applicable); and c) more than one different indel sequence was observed (to avoid potential jackpot effects).

Comparing the reproducibility between the two biological replicates of oZFN treatments

To examine the reproducibility of the data, the Pearson correlation test was applied to common sequences between the replicate datasets (oZFN DD/RR replicate 1 and 2: 10 pg normal, 10 pg deformed, and 20 pg deformed) on the log odds ratio of ZFN treated sample vs. control sample **(Figure 2-5)**. The indel rates for the two replicates were

averaged for further analyses. If for any off-target site, there were <1000 sequences in one of the replicate, the frequency of lesions from the other replicate was used for analysis.

Bacterial one-hybrid based Activity Assay

To assay the importance of Guanine contacts for the binding of oZFPL and oZFPR to their respective binding sites, a B1H activity assay was performed. The ZFP binding sites (wild type or mutant) were cloned in the pH3U3 reporter vector. The oZFPL and oZFPR were cotransformed with plasmids bearing their binding sites in US0 cells and the activity assay was performed as described in Noyes *et al*.[92]. From a stock of 10^9 cells/ml, the 10-fold serial dilutions were placed as 5 ul drops on 2XYT or NM selective media plates containing kanamycin (25ug/ml), carbenicillin (100ug/ml), 3-aminotriazole (1mM or 10mM) and IPTG (10uM). The colonies were grown at 37°C for 20hrs (1mM 3-AT plates) or 48hrs (10mM 3-AT plates). The number of colonies were counted for the NM selective plates and reduction in colony counts was calculated as -log (number of colonies for the wild-type or mutant binding site/ number of colonies for the wild-type sequence).

Binding site analysis using Illumina Sequencing

Binding site selections were performed using the 28 bp randomized library as described previously[89,92,94]. The colonies surviving on selective media plate containing 5mM 3-AT, 10uM IPTG were counted and then washed off the plate. The plasmid DNA from the pooled colonies was isolated and the binding site was PCR amplified using Phusion

(NEB) enzyme starting with 50ng plasmid DNA as template. The PCR reaction conditions were as follows: 98°C 3 min; 25 cycles (98°C, 20 sec.; 60°C, 20 sec.; 72°C, 30sec); 72°C 5min. The primer sequences employed were:

Forward:

CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGTGAACGCTCTCCTGA GTAGG

Reverse: CTGCTCTGTCATAGCTGTTTCC

One of the Solexa adapter sequences (adapter-P2) was incorporated into the forward primer so that only single adapter ligations were required (see below). The PCR product (~1ug) was digested with 40 units of *Eco*RIHF (NEB) enzyme at 37°C and gel purified. The purified DNA was treated with Klenow Exo$^-$ (NEB) in the presence of 0.1mM dNTPs for 45min at 37°C. The DNA was spin purified using QIAquick PCR purification kit (Qiagen) and barcoded adapters **(Table A-5)** were ligated using 20 units of T4 DNA ligase (NEB) at room temperature for 2hrs. **Important: the barcoded adapters should <u>not</u> be 5' phosphorylated so that they will ligate to only the *Eco*RI digested end of the DNA molecule.** Following ligations, the DNA was PCR amplified (Phusion polymerase (NEB)) using in-house Illumina primers, where 20% of the ligation reaction was used as the starting template. The reaction conditions are as follows: 98°C 3 min; 6 cycles (98°C, 20 sec.; 60°C, 20 sec.; 72°C, 30sec); 72°C 5min. The primer sequences employed were:

In-house Illumina P1 (Invitrogen):

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC
GATCT

In-house Illumina P2 (Invitrogen):

CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT

Following PCR of the barcoded-adapter-ligated DNA, 20% of the reaction was run on an agarose gel containing ethidium bromide to estimate DNA quantity. An equal amount of PCR-amplified DNA (~80ng) from different barcoded-adapter ligated samples was pooled. The pooled DNA was run on gel and the adapter-ligated fraction of the DNA was gel purified. This purified DNA (~25ng) was amplified by PCR using the Illumina Genomic DNA primers (1.1 and 2.1) and Phusion polymerase (NEB) using the follows conditions: 98°C 3 min; 9 cycles (98°C, 20 sec.; 60°C, 20 sec.; 72°C, 30sec); 72°C 5min. The PCR sample was gel purified and sequenced at 4pM concentration. The sequences were binned according to their barcode. For each ZFP, unique sequences were isolated and ranked according to their frequency of occurrence. The top most frequent sequences (equivalent to the number of surviving colonies on the selective media plates) were analyzed by MEME to discover the recognition motif[216]. The aligned sequences were then used to generate a Sequence logo[207] using WebLogo[208].

Illumina sample preparation for lesion analysis at "Off-Target Sites I"

For the **"Off-Target Sites I"** group and the on-target site (target site run 1)**,** the Illumina sample was prepared as described previously [94]. Genomic DNA from the following 7

samples were used: uninjected, oZFPs DD/RR 10pg normal Replicate 1, oZFPs DD/RR 10pg deformed Replicate 1, oZFPs DD/RR 20pg deformed Replicate 1, nZFPs DD/RR 10pg normal, nZFPs DD/RR 20pg normal, nZFPs DD/RR 20pg deformed. The number of embryos used for isolating the genomic DNA is given in Figure 2-3. The sequences of the primers used to PCR DNA flanking the off-target sites as well as the number of sequences with wild type sequence and indels recovered at each site are provided in the **Online Table** available at http://nar.oxfordjournals.org/content/39/1/381/suppl/DC1.

Primer design for amplifying off-target sites

The sequences for primers used to amplify the DNA flanking the off-target sites are listed in the **Online Table**[a]. For each primer pair, the proximal primer binds neighboring the off-target site while the distal primer binds ~150bp away. The proximal primer contains an *Acu*I restriction site that following cleavage allows the Illumina adapter to be ligated within a few basepairs of the center of the putative off-target site. The distal primer contains the Illumina adapter (P2) sequence, and consequently Illumina adapter ligation is only necessary at the proximal end.

Solexa Sample preparation for the Off-Target Sites II

Lesion analysis at "Off-Target sites II" group (10g and non-10g) and the on-target site (target site run 2) was performed for the following 14 samples: uninjected, oZFPs DD/RR

[a] Online Table available at http://nar.oxfordjournals.org/content/39/1/381/suppl/DC1

10pg normal Replicate 1, oZFPs DD/RR 10pg deformed Replicate 1, oZFPs DD/RR 20pg deformed Replicate 1, oZFPs DD/RR 10pg normal Replicate 2, oZFPs DD/RR 10pg deformed Replicate 2, oZFPs DD/RR 20pg deformed Replicate 2, nZFPs DD/RR 10pg normal, nZFPs DD/RR 20pg normal, nZFPs DD/RR 20pg deformed, oZFPs EL/KK 50pg normal, oZFPs EL/KK 50pg deformed, oZFPs EL/KK 100pg normal, oZFPs EL/KK 100pg deformed. Genomic DNA was isolated from the ZFN-injected or uninjected embryos 24-hpf using DNeasy Blood and Tissue Kit (Qiagen). The number of embryos used for each group is listed in Figure 2-3. Using the isolated genomic DNA as template, the DNA flanking the off-target sites was PCR amplified with the primers listed in the **Online Table**[a]. For each of the 14 conditions mentioned above, PCR-amplified DNA for all 96 off-target sites was pooled. The pooled DNA for the off-target sites was digested with 25 units of *Acu*I restriction enzyme for 4hrs at 37°C and gel purified using QIAquick Gel Extraction kit (Qiagen). The purified DNA was treated with 1 unit of T4 DNA polymerase at 12°C for 15min in the presence of 0.1mM dNTPs to polish the 3' overhangs (from *Acu*I digestion). The reaction was stopped immediately after 15min by adding EDTA (final concentration of 10mM) and heating it to 75°C for 20min. DNA was then spin purified using Qiaquick PCR purification kit (Qiagen) and was treated with Klenow Exo$^-$ for 45min at 37°C in the presence of 0.1mM dATP to add 3' A overhangs. The DNA was spin purified as described above and the barcoded adapters were ligated as described above (**Binding site analysis using Illumina Sequencing)**. The 14 off-target samples were ligated to the 2bp-barcoded-adapters and PCR amplified with in-house

[a] Online Table available at http://nar.oxfordjournals.org/content/39/1/381/suppl/DC1

Illumina primers as described above. 20% of the PCR reaction was run on an agarose gel containing ethidium bromide to estimate DNA quantity. An equal amount of the PCR-amplified DNA (~80ng) from 14 barcoded-adapter ligated samples was pooled (~1ug total).

For the on-target site lesion analysis, the DNA flanking the *kdrl* ZFN site was PCR-amplified from the genomic DNA samples from the uninjected or ZFN treated embryos (14 samples in total) using the following primers:

kdrex2 solexa on-site 5p: CCTGATCCACAACTGCTTCCTGATGGATATCCAC

kdrex2 solexa on-site-P2 3p:
CGGCATACGAGCTCTTCCGATCTATAAAGTGGCCATTGAACGTAGATGCAC

The PCR amplified DNA was digested with *Eco*RV restriction enzyme and gel purified. The purified DNA was treated with Klenow Exo$^-$ as above and spin purified. The 14 on-target samples were ligated to the 2bp-barcoded-adapters, PCR amplified with in-house Illumina primers and pooled as described above. ~10ng of the pooled on-target DNA was added to ~1ug of the pooled off-target DNA and was run on gel. The adapter-ligated DNA was gel purified and used as a template for PCR amplification with Illumina primers (1.1 and 2.1) as described above. The PCR sample was gel purified, combined with the off-target site pool at the appropriate ratio, and then was sequenced at 4pM concentration.

Comparisons of the off-target lesion frequency for different ZFN treatments

To compare the difference in lesion frequency in embryos from any two ZFN treatments (oZFN 10pg normal vs oZFN 10pg deformed; oZFNs vs nZFNs and oZFNs^{DDRR} vs oZFNs^{ELKK}) the number of reads with wild type sequence for all analyzed off-target sites was combined as were all of the indel reads. This provided an overall lesion frequency that could be compared between different ZFN treatments using the Chi-square test.

Comparing the distribution of active and inactive off-target sites for the number of matches to the target sites

Each heterodimeric off-target site containing a 5-or 6-bp spacing was scored based on the number of matches to the target site (considering only ZFPL and ZFPR half-sites) and, based on the criteria given above, off-target sites were divided into active and inactive off-target sites. Kendall correlation tests were performed to determine the significance of the correlation between the number of active sites and the number of inactive sites, and the significance of the correlation between the number of matches to the target site and the ratio of active sites. In addition, a two-tailed Fisher Exact Test was performed to determine whether the relative number of active off-target sites from two different groups of ZFN-treated embryos was significantly different.

Comparing the distribution of active and inactive off-target sites for the number of conserved Guanines

Each heterodimeric off-target site containing a 5-or 6-bp spacing was scored based on the number of positions that had a guanine found in the target sequence (GXXGGXGXG and XXGGXGGGX; 10Gs) and, based on the criteria given above, off-target sites were

divided into active and inactive off-target sites.

Frequency Plots for ZFPL and ZFPR binding sites from Inactive, Active and All off-target sites

The ZFPL and ZFPR half-sites were extracted from the relevant (active, inactive and all) groups of heterodimeric off-target site containing a 5-or 6-bp spacing and aligned using MEME[216]. Frequency-logos for each group of half-site sequences were generated using Weblogo[208].

Germline transmission of off-target lesions

To assay germline transmission of off-target lesions, we crossed the *kdrl* founder zebrafish (889.7) obtained in Meng *et al.* with wild type zebrafish, which yielded 33 surviving embryos. These embryos were genotyped for the kdrl ZFN target site using the NspI digestion assay[94]. Out of 33 embryos, 17 were found to be heterozygous for the *kdrl* mutation whereas 16 did not carry the mutant allele. The genomic DNA from the 17 heterozygous and 16 homozygous embryos were pooled as two separate groups. To identify any lesions at the off-target sites, the *kdrl* ZFN active off-target sites were PCR amplified as described above from the two genomic DNA pools. The following off-target sites were analyzed: OT1, OT2, OT3, OT4, OT5, OT6, OT7, OT8, OT10, OT11, OT13, OT14, OT15, OT16, OT18, OT20. The target site was also included in the analysis. The Solexa sample was prepared and sequenced as described above. The sequences for each site were binned for analysis using the unique 9 bp "prefix" described above.

Chapter III: A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity

Contents of Chapter III have been published previously as:

Christensen RG, Gupta A, Zuo Z, Schriefer LA, Wolfe SA, Stormo GD (2011) A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. Nucleic acids research 39, e83

Ryan Christensen from Gary Stormo's lab at the Washington University developed the GRaMS algorithm and preformed all the computational analysis. I created the 6-bp binding site library and performed CV-B1H-selections for Zif268.

Introduction

The Cys2His2 zinc finger is the most frequently observed DNA binding domain family in the metazoan transcription factors (TFs)[14]. In the canonical recognition each finger specifies 3 bp through specificity determinants present on the recognition helix of the zinc fingers[45,48,58]. Many laboratories have demonstrated that fingers with novel recognition preferences can be obtained from randomized libraries through various selection methodologies[48,58,162]. Further, these synthetic zinc finger units can be assembled into tandem arrays known as ZFPs that can be fused to an effector domains to create artificial TFs, recombinases and nucleases (ZFNs). The utility of artificial ZFPs is currently restricted by the limited availability of highly specific zinc finger units that can be reliably assembled to create artificial ZFPs. Developing a method that can provide a rapid and accurate assessment of the binding site specificities of selected zinc finger units would represent an important advance for defining and engineering specificities of zinc finger units. Ultimately, such a method would provide a high quality dataset of DNA-protein interactions that can be utilized to create an accurate 'recognition code' for zinc fingers enabling rational design of artificial ZFPs as well as prediction of specificities of naturally occurring zinc finger containing TFs to elucidate gene regulatory networks.

Several methods exist for determining the specificity of DNA binding proteins[107]. SELEX (Systematic Evolution of Ligands by EXponential enrichment) allows extraction of binding sites for a protein from an unbiased pool of oligonucleotides but involves multiple rounds of enrichment of bound sites and therefore provides only the highest affinity sites for a protein[109-112]. A modified version of SELEX, HT-SELEX requires

only one round of enrichment and can be coupled with high throughput sequencing to develop energetic models that fit better than the conventional SELEX[117]. However, like SELEX this method is *in vitro* and requires protein purification or *in vitro* protein expression. Protein binding microarray (PBM) is a medium-throughput *in vitro* method for identifying binding site preferences for naturally occurring TFs or engineered ZFPs but is limited to proteins with binding sites shorter than ~10 bp [107,118,119].

In comparison to these *in vitro* methods, bacterial-one-hybrid (B1H) based selection methods provide *in vivo* determination of binding site specificities of TFs as well as artificial DNA binding proteins[19,89,90,92,93]. In this approach, the DNA binding domain (DBD) to be assayed is fused to an alpha-subunit (α-subunit) or in the recent versions the omega-subunit (ω-subunit) of the RNA polymerase[19,92]. A randomized binding site library is used which is located upstream of a weak 'lac' promoter driving the expression of yeast *HIS3* and *URA3* reporter genes. The selections are performed in the *E. coli* selection strain lacking the endogenous *hisB* (the bacterial homolog of *HIS3*) and *pyrF* (bacterial homolog of *URA3*) genes where one member of a library is compartmentalized with the α- or ω-fused DBD (Note: when using the ω-fused DBD, the endogenous *rpoZ* is also deleted). Binding of the DBD to its binding sites with high affinity and specificity will activate the reporter genes allowing bacterial cells to form colonies on selective media. The recovered binding sites can either be sequenced individually via Sanger sequencing or as a pool via high-throughput sequencing. Since the B1H-selections are performed in *E. coli*, there is no need for protein purification or i*n vitro* protein expression. Further, the bacterial genomic DNA acts as competitor DNA allowing active

binding sites to be selected for both specificity and affinity. Moreover, B1H system involves a single round of selection thus making the selection process easy and rapid.

Here we describe a new version of B1H-based selection system, the constrained variation-B1H (CV-B1H) method that uses a randomized binding site library in a fixed register allowing a rapid and easy estimation of binding site specificities of two-finger zinc fingers units. We have also developed a complementary algorithm, 'Growth Rate Modeling of Specificity' (or GRaMS) that models the relationship between binding energy and growth rate to increase the accuracy of the quantitative specificity model produced from B1H data.

Results

CV-B1H method

To determine binding site specificities of two-finger zinc finger units (2F-modules), we constructed a 6 bp-randomized library adjoining a 'GCG' DNA triplet (NNNNNNGCG) **(Figure 3-1)**. We incorporated another randomized 4 bp element six base-pairs upstream of the 6 bp-randomized region that serves as an internal control to detect any sequence biases in the library before or after selections and any jackpot effects during PCR amplification of recovered sequences. Oligonucleotides encoding the binding site library were cloned into the reporter plasmid pH3U3 upstream of a weak promoter that drives expression of the yeast *HIS3* and *URA3* reporter genes (**Figure 3-1**). The library was counter-selected by growing the library-transformed cells on 5-FOA (5-Fluoroorotic

Figure 3-1

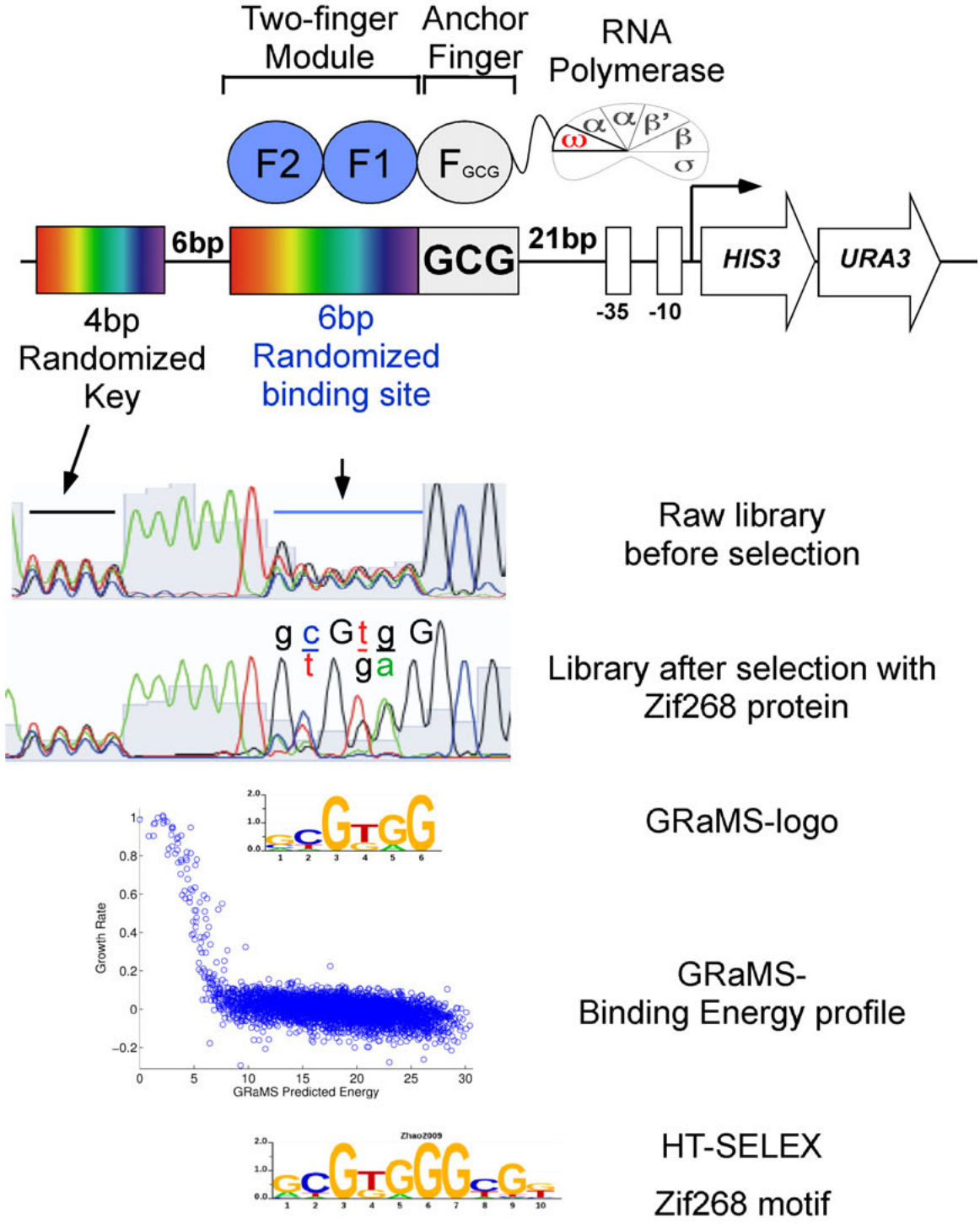

Two-finger Module
Anchor Finger
RNA Polymerase
F2
F1
F GCG
ω
α
α
β'
β
σ
6bp
GCG
21bp
HIS3
URA3
-35
-10
4bp Randomized Key
6bp Randomized binding site
Raw library before selection
Library after selection with Zif268 protein
GRaMS-logo
GRaMS- Binding Energy profile
Growth Rate
GRaMS Predicted Energy
HT-SELEX
Zif268 motif

Figure 3-1: CV-B1H method to determine DNA binding specificities of zinc fingers. The 2F-ZF unit is fused to an N-terminal finger (RSDTLAR) that binds to the 'GCG' triplet adjacent to the 6 bp randomized zinc finger binding region on the reporter plasmid. There is also a 4 bp randomized region (key region) that serves as an internal control to identify biases in the recovered DNA sequences due to jackpot effects. Following selection, the surviving colonies are pooled and the distribution of bases recovered at each position within the selected binding sites can be evaluated in a single sequencing reaction. To test this system we used the DNA binding domain of the Zif268 protein and performed CV-B1H selections to determine binding site specificity of fingers 2 and 3 of Zif268. The results from Sanger sequencing are shown both before and after selections with Zif268. The recovered binding sites are determined by Illumina sequencing and analyzed by the GRaMS (Growth Rate Modeling of Specificities) method that predicts binding energies as shown in the graph and provides a sequence logo for Zif268 fingers 2 and 3. The binding site model obtained using the and GRaMS method closely matches the motif obtained by HT-SELEX[117].

acid) containing media to remove any self-activating sequences[91,93]. To determine the binding site specificity of a 2F-module, it is fused to an N-terminal anchor zinc finger that binds to the GCG triplet adjoining the library. Apart from providing additional affinity to the 2F-module to be functional in the B1H-selection system, by binding the GCG DNA triplet the anchor finger fixes the register of the binding sites for the 2F-module thus allowing the selected binding sites to be sequenced as a pool in a single reaction. When sequencing the recovered binding sites individually, this property of the system eliminates the need to align the selected active binding sites. Binding of the 2F-module to its binding sites with high specificity and affinity allows the selection strain to grow on selective media plates lacking histidine and/or uracil. The selection stringency of the B1H selections can be tuned by incorporating variable amounts of 3-Amino triazole (3-AT) (a competitive inhibitor of the HIS3 enzyme) and IPTG (which controls 2F-module protein concentration) in selective media plates. The binding sites from the surviving colonies can be sequenced in a pool either through Sanger sequencing or as a population through Illumina sequencing.

Testing the CV-B1H method

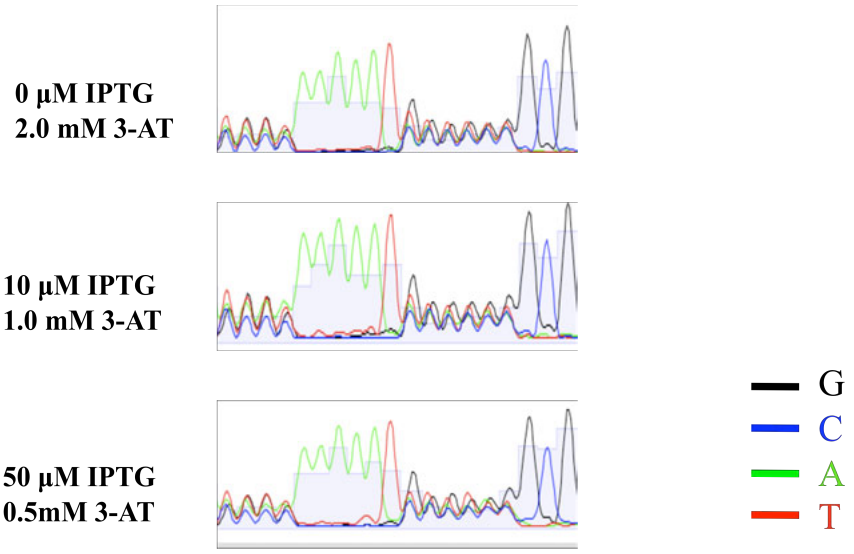
To test the utility of the CV-B1H system to determine binding site specificities we characterized Zif268, a three-finger protein with a well defined DNA binding specificity[89,93,117]. Since the N-terminal finger-1 of the Zif268 is known to bind the GCG sequence, we fused the DNA binding domain of the wild type Zif268 protein to the ω-subunit and determined the binding specificity of finger-2 and finger-3. We co-

transformed Zif268 with the library-containing reporter plasmid into bacterial cells and plated ~1 million double transformants onto selective media containing plates. The selective media lacked histidine and contained nine different combinations of 3-AT (0.5, 1.0 and 2.0 mM) and IPTG (0, 10 and 50 μM). The plates were incubated at 37°C for different time points (4, 8, 12, 18 and 24 hours). At the appropriate time point, bacterial cells were washed off the plate and plasmid DNA was isolated for a single selection condition in one tube. The binding sites were sequenced as a pool in a single sequencing reaction via Sanger sequencing. The chromatograms at 4 hr did not show much enrichment of the Zif268 binding site sequences **(Figure 3-2)**. At the 8 hr time point, selections from no-IPTG conditions did not show any enrichment of Zif268 binding sites but selections containing 10 or 50 μM IPTG displayed some enrichment as estimated by the appearance of individual peaks for certain bases. At 12 hr and 24 hr time points, the chromatograms showed clear enrichment of the Zif268 binding sites after selections **(Figure 3-2)**. Moreover, there were differences in the strength of enrichment at different stringencies imposed by 3-AT and IPTG concentrations; higher stringency selections (0 μM IPTG and 2 mM 3-AT) showed stronger enrichment of bases at certain positions than lower stringency selections (50 μM IPTG and 0.5 mM 3-AT). In sum, these results demonstrated that the CV-B1H selection system can be utilized to rapidly determine binding specificities of zinc fingers.

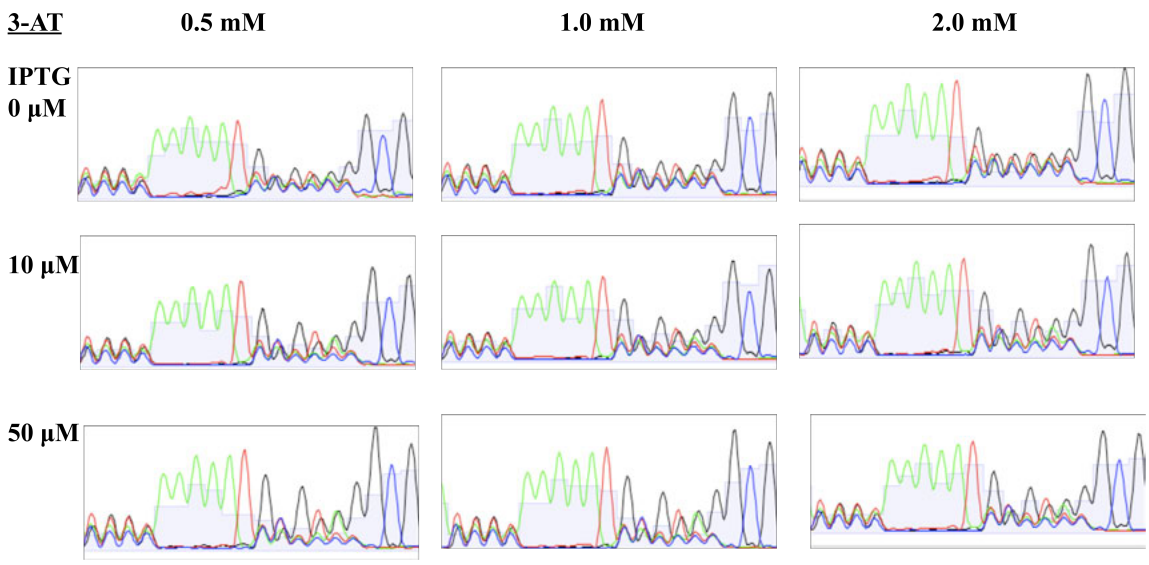
Quantitative Motifs

Figure 3-2

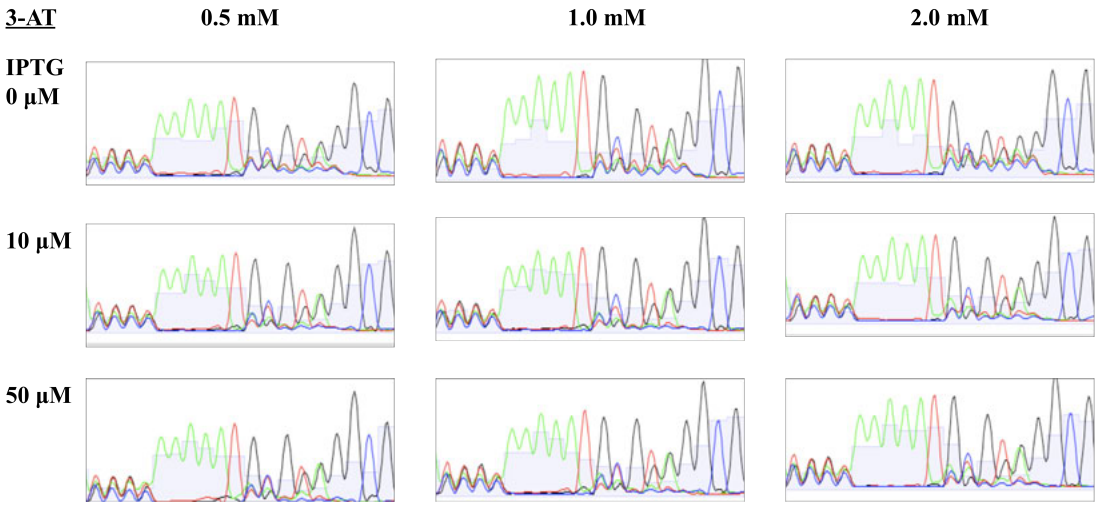
a) 4 hr



b) 8 hr



c) 12 hr



d) 24 hr

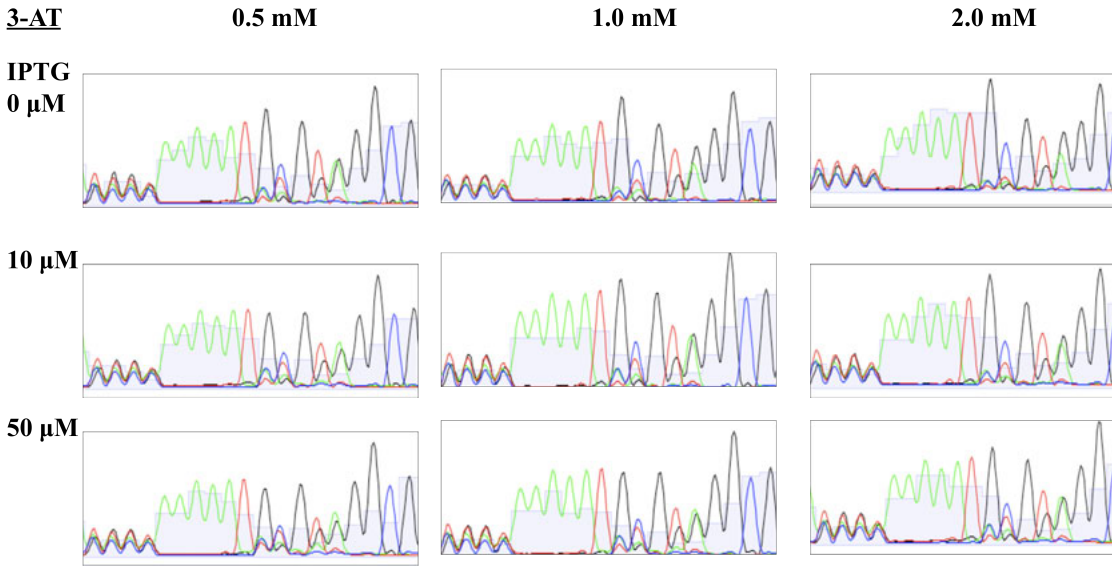


Figure 3-2: Binding site profiles for Zif268 from Sanger sequencing. After CV-B1H selections at different stringencies (3-AT and IPTG) and time points, surviving colonies were washed from the plate and the binding sites were sequenced as a pool via Sanger sequencing. The chromatograms are displayed for 4 hr (only a few conditions) **(a)**, 8 hr **(b)**, 12 hr **(c)** and 24 hr **(d)** time points. The selection condition '0.5 mM 3-AT and 50 µM IPTG' represents the least stringent condition and '2 mM 3-AT and 0 µM IPTG' represents is the highest stringency condition.

The chromatograms obtained from the Sanger sequencing post-CV-B1H selections provide qualitative information on binding preferences of a given zinc finger but quantitative assessment of binding site models require sequencing individual binding sites. We performed Illumina sequencing of the pools of binding sites where DNA from each condition was barcoded and sequenced in a single lane. Forty-five different selection conditions (3-AT, IPTG and time points) were combined, where we obtained on an average 300,000 sequences for each condition. We also sequenced the 6 bp library before selection (but after counter-selection) to construct a background model to define the enrichment of sequences from the library. Motif comparisons for data generated from these experiments were made to the subset of published Zif268 high throughput-SELEX (HT-SELEX) data[117] that contained the GCGG sequence in the last four positions on the binding site, to be consistent with the constraints in our 6 bp library selections. The BioProspector[218] and log-odds analyses on the recovered sequences from various conditions were highly variable, in particular for selections with short incubation times **(Figure 3-3)**. The constructed motifs predicted the HT-SELEX data with median R^2 values of only 0.52 and 0.54 for BioProspector and log-odds, respectively and the maximum values were only 0.71 and 0.74 **(Figure 3-3)**. These results are explained by the fact that the initial library, which has been counter selected to remove self-activating sequences, has a very low proportion of the consensus binding site and some other closely related sites. Their low initial frequencies ensure that even after the 24 hour incubation they have not become the most abundant sites, therefore leading to construction of a PWM with sub-optimal parameters. Although both the log-odds

Figure 3-3

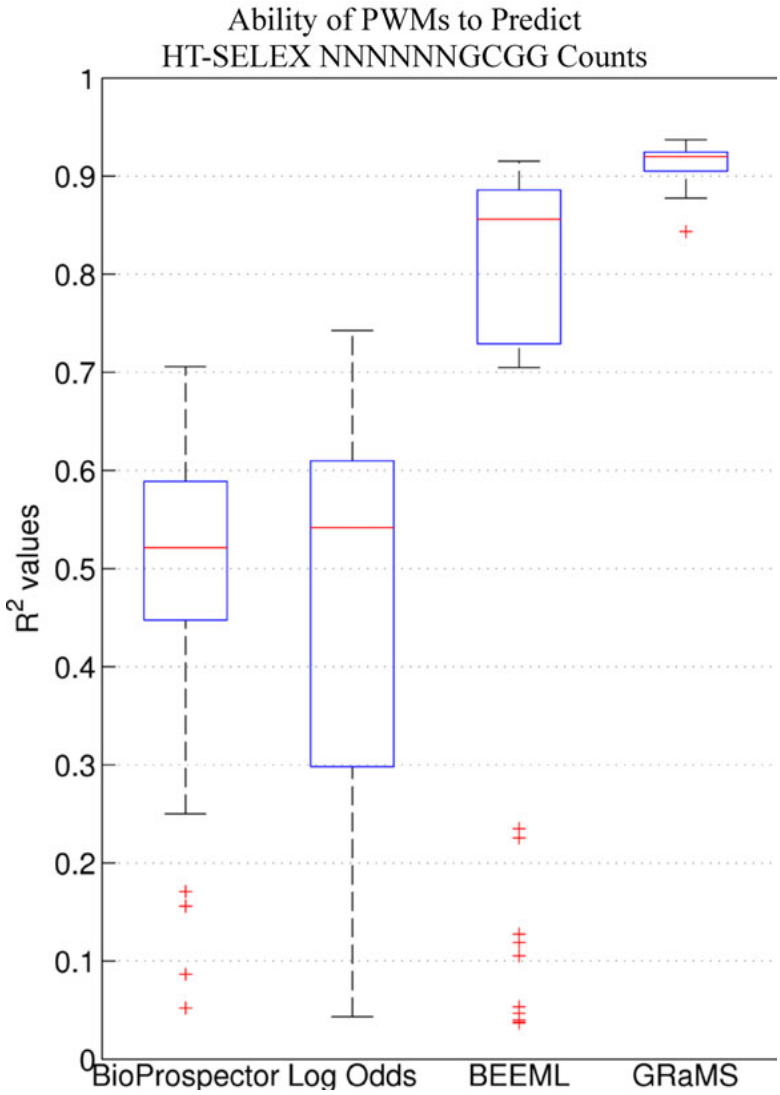


Figure 3-3: Performance of GRaMS on the CV-B1H data and comparison to other methods. Boxplot showing the ability of the 45 PWMs produced by each analysis method using each B1H data set as the training data to predict the HT-SELEX NNNNNNGCGG data[117]. For each model, R^2 was calculated to determine the correlation between the predicted and observed HT-SELEX counts.

method and BioProspector take the initial library into account through their background estimates they do not capture the explicit deficiency of specific binding sites, some of which are high affinity sites.

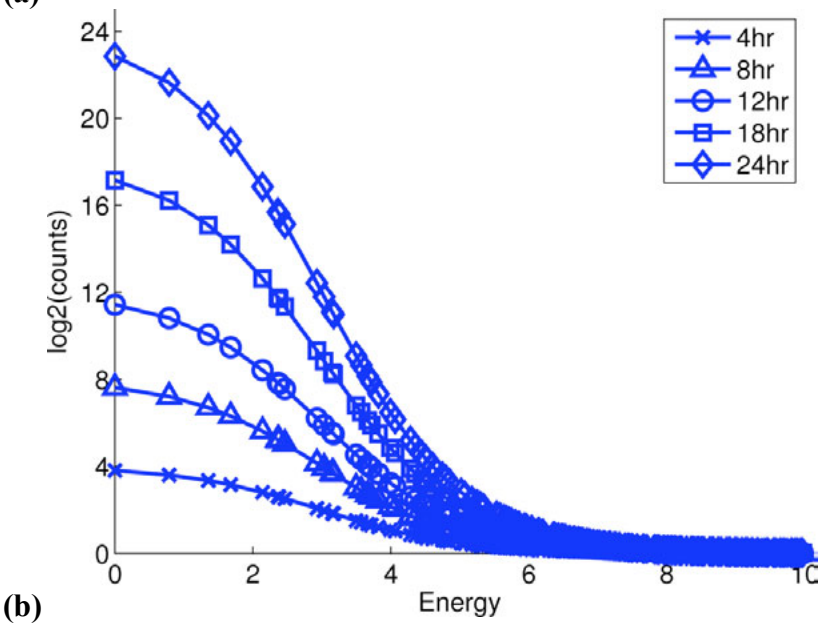
Since the BEEML program[117] takes into account each specific binding site in both the initial and selected libraries, we tested the BEEML program on the 45 data sets. Based on a biophysical model for enrichment, BEEML performs a nonlinear regression to find the optimal parameters for a PWM. While its performance is still quite poor on the earliest time points, its median R^2 is 0.86 and its best is 0.92, both significantly better than the other methods **(Figure 3-3)**.

GRaMS analysis

To improve the accuracy of the PWMs obtained from analysis of the B1H-based data we developed the 'Growth Rate Modeling of Specificities' (GRaMS) algorithm that would capture the differences in relative growth rate of the surviving colonies resulting from the differences in relative energy of binding sites. The GRaMS model is described in the 'Methods' section. For the ideal simulated data, the relative ratios of different binding sites as a function of their energetics at various time points fall on different lines, but when converted to growth rates they all converge to a single line that shows the relationship between growth rate and binding energy **(Figure 3-4)**. GRaMS was performed on all 45 selections and the binding energies were predicted **(Figure 3-5)**. While obviously noisier than the simulated data, the curves are all very similar and are consistent with our model. **Figure 3-6** shows the logos for all 45 data sets. The PWMs

Figure 3-4

(a)



(b)

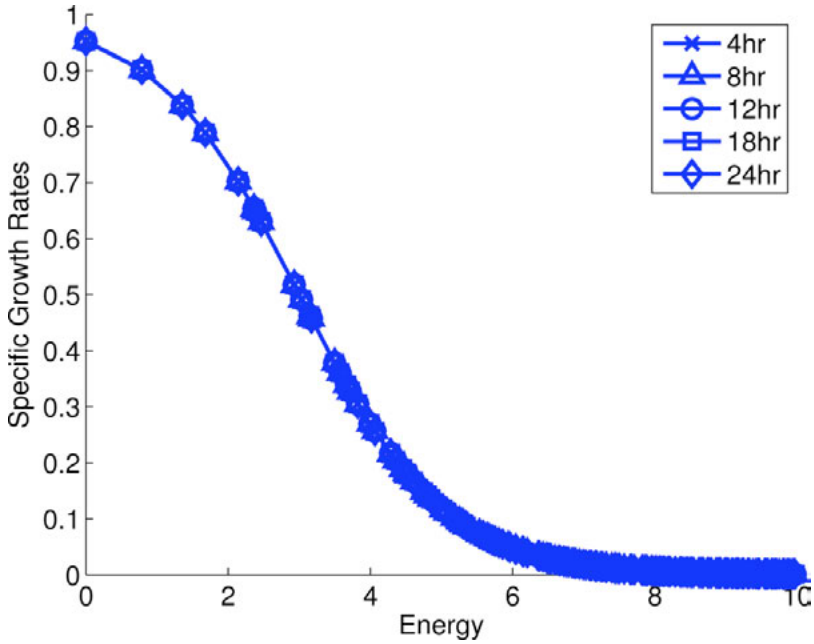


Figure 3-4. **Simulated Ideal B1H data**. (**A**) Energy versus log2(counts) curves for 5 different simulated data sets. Each curve represents a B1H experiment that was run for a different amount of time: 4, 8, 12, 18, 24 hours. (**B**) Energy versus growth rate curves for the same five data sets: 4, 8, 12, 18, 24 hours. Energies are given in units of K_bT. For these simulations, a μ value of 3 was used. All 4096 binding site alleles occurred once and each gave rise to a single simulated colony. Energies were assigned using the PWM of Zhao *et al.*[117].

Figure 3-5

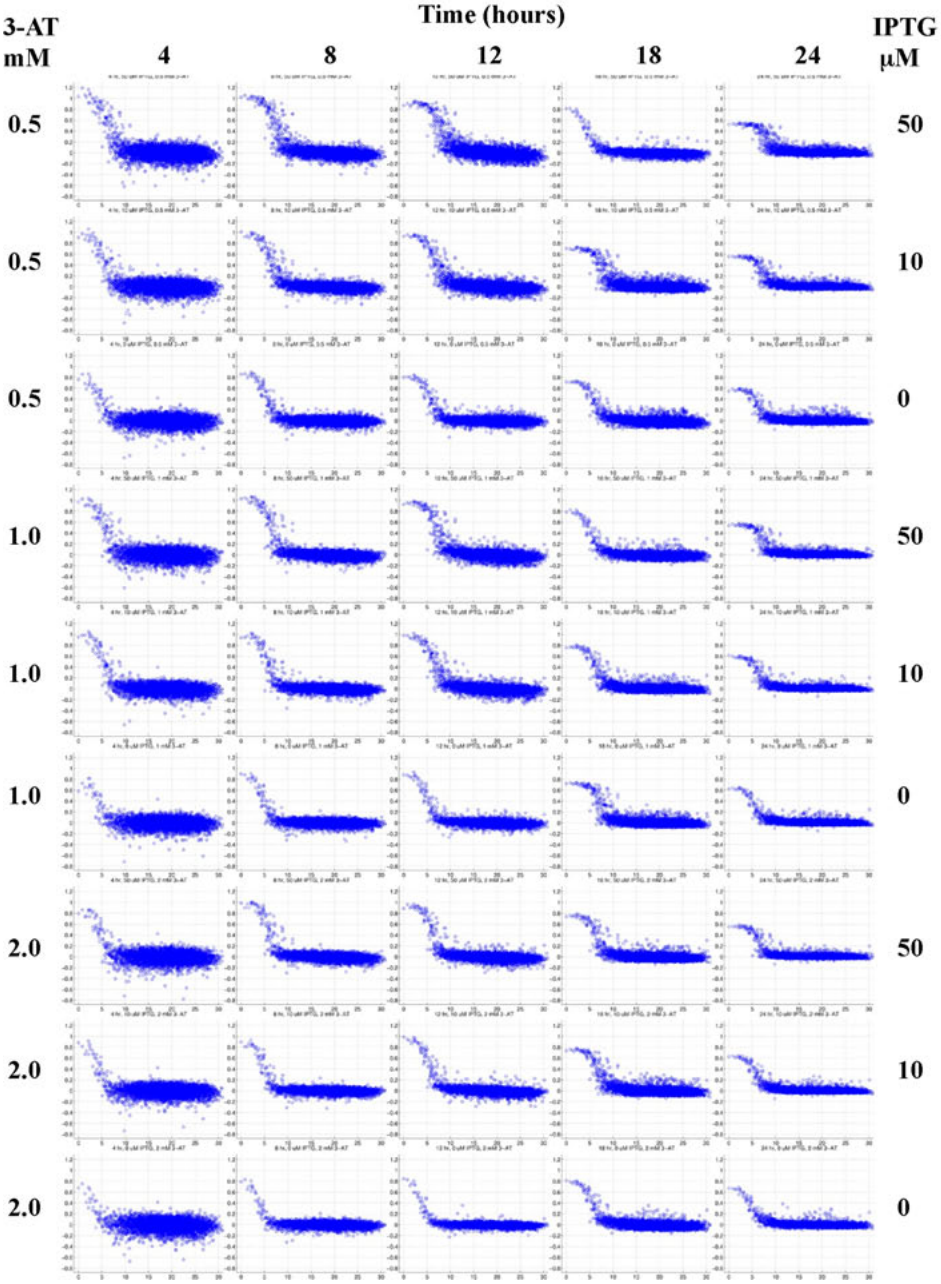


Figure 3-5: Plots of predicted energy values versus B1H growth rates (shifted so that the median is set to zero) for all 45 conditions. The PWM of Zhao *et al.*[117] was used to predict the energy of all 6mers. The growth rate is on the y-axis and the predicted energy on the x-axis. The plots are organized in a 5-column by 9-row grid. Each column corresponds to one of the 5 different time points (4, 8, 12, 18 and 24 hours), in that order, from left to right. Each row represents a combination of 3-AT and IPTG concentration mentioned on the left and right respectively.

Figure 3-6

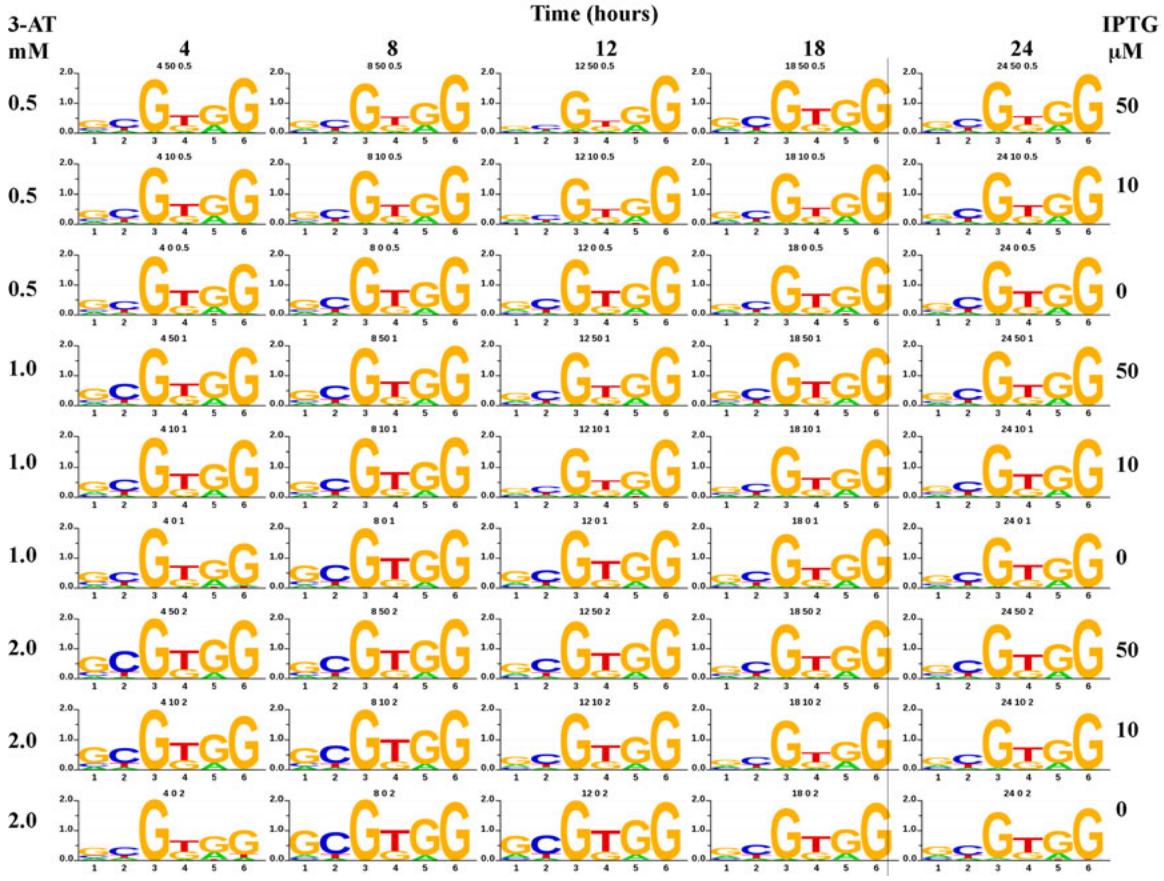


Figure 3-6: Sequence logos for all 45 PWMs produced using GRaMS on all of the HT-B1H datasets. The title of each sequence logo indicates (in order): the duration of each experiment in hours; the IPTG concentration (µM); the concentration of 3-AT (mM). All sequence logos were produced using in-house software, svgSeqLogo, written by Ryan Christensen. The y-axis of each logo is in units of bits.

obtained from GRaMS show a high correlation with the HT-SELEX data (median R^2 = 0.92). Note that the models from different conditions are very similar indicating that with GRaMS analysis the resulting models are relatively insensitive to the exact experimental protocol. Motifs constructed from the 4 h time points remain the least accurate and those from the late time points have slightly reduced accuracy probably due to the onset of colony saturation for the highest affinity binding sites **(figure 3-7)**. At the earlier times points increased stringency, using higher concentrations of 3-AT, improved the quality of the motifs somewhat but the results are relatively insensitive by the concentration of IPTG.

Discussion

We have developed a CV-B1H method that combines the advantages of B1H assay with a fixed register library, for rapidly determining DNA binding specificities of zinc finger proteins. This method is based on the 'profiling' system developed by Wolfe and colleagues[93] and offers several advantages over the original system. In the CV-B1H system we fuse the zinc fingers to the ω-subunit of the RNA polymerase instead of the α-subunit used in the profiling system. This allows us to perform the selections in a ω-deficient strain of *E. coli* (Δ*rpoZ*) which results in a uniform incorporation of the ω-fused zinc fingers in the RNA polymerase thus increasing the sensitivity of the detection of DNA-protein interactions[19,92]. Further, instead of using a 5 bp-randomized library, we created a 6 bp randomized library that allows evaluating specificities of two-finger zinc

Figure 3-7

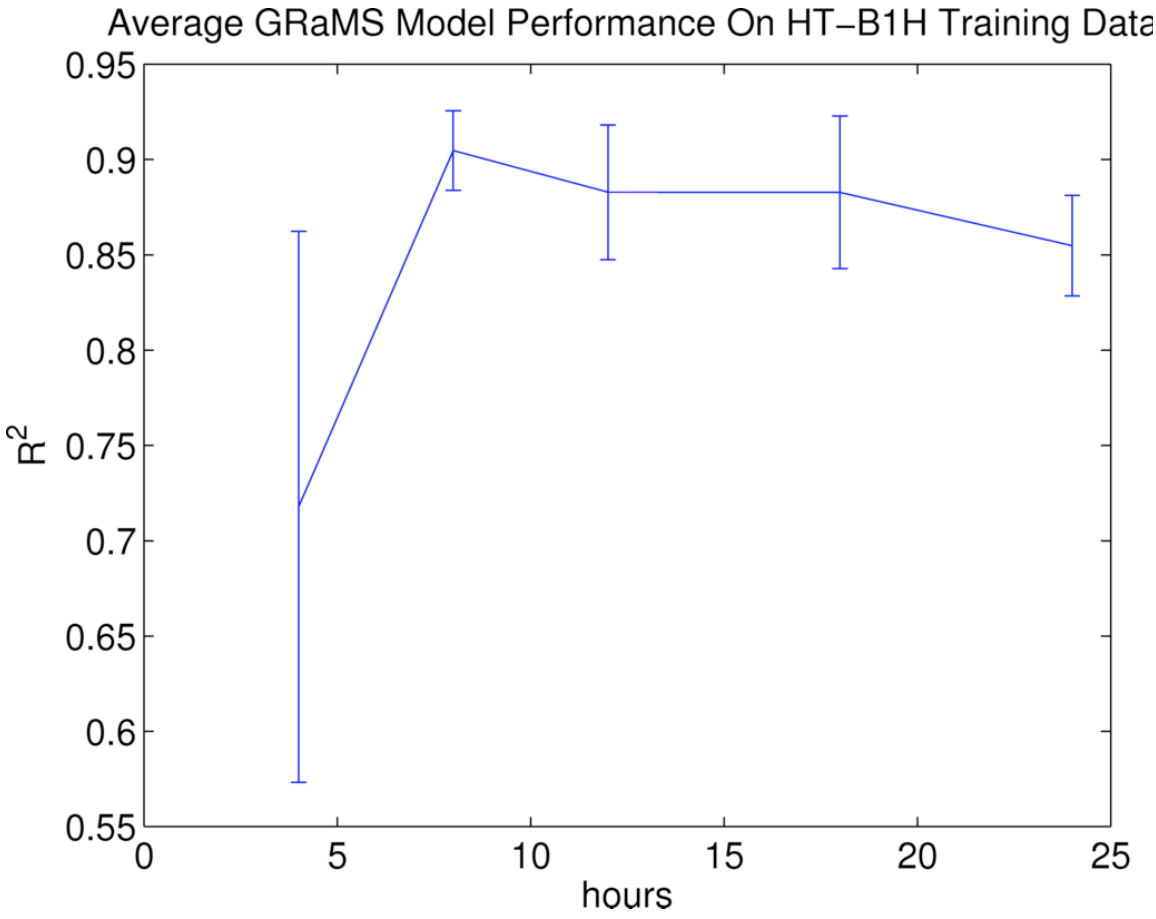


Figure 3-7: Plot showing the ability to predict the training data, as measured by R^2, of all GRaMs models, as a function of time. Points show the average R^2 value per time point. Error bars indicate the range of the nine different IPTG and 3-AT concentrations for each time point.

finger units instead of 1.5 fingers. Finally, the 6 bp library is flanked with restriction sites that allow ease of sample preparation for Illumina sequencing. Deep sequencing of the binding site pools enables us to build accurate quantitative binding energy models from a large population of recovered binding sites.

To analyze the CV-B1H data, we have developed the GRaMS algorithm that measures growth rates of cells across the distribution from high affinity to low affinity sites and uses a biophysical model for the relationship between growth rate and binding energy. GRaMS is able to obtain more accurate models from B1H data than any other approach we tested. We obtained good models under all of the variations that we tested except for the very early time point (4 h) where we believe that the non-dividing cells make a significant proportion of total cells on the selection plate and contribute to the high background in the dataset. The 10 and 50 μM IPTG conditions worked better at early time points than the 0 μM IPTG probably due to insufficient protein expression at the 0 μM IPTG condition for maximal reporter expression resulting in somewhat random growth at early time points thus increasing the noise in the data. Although, at later time points there were minor differences in the motifs obtained from lower stringency and higher stringency, the GRaMS analysis was fairly insensitive to the selection conditions. Finally, at 24 h the accuracy of GRaMS started to decrease probably due to saturation of the colony size of high-affinity binding sites, suggesting that post-CV-B1H the colonies should be harvested before 24 hours of incubation.

The log-odds and BioProspector models performed poorly on the CV-B1H data **(Figure 3-3)**. One of the reasons could be an under-representation of the high energy Zif268 binding sites in the initial library which would in turn result in their under-representation in the final binding site pools. This underrepresentation of the high energy binding sites for Zif268 in the initial library could result from a minor contamination of the library plasmid with the Zif268 expression plasmid during counter-selection that would eliminate these sequences from the initial library. However, since the growth rate of the consensus site would be higher than the non-consensus (but similar) sites, GRaMS would assign it higher energy yielding a more accurate PWM.

As shown in Chapter IV, the CV-B1H system can be employed to rapidly determine binding site specificities of the 2F-modules obtained from zinc finger selections, choose the highly specific 2F-modules for use in creating ZFNs and perform focused alterations to improve the specificities of proteins that show suboptimal specificities. Although, we utilized this method to determine the binding site specificities of 2F-modules, it can be modified to assay single zinc fingers by combining them with another anchor finger besides the 'GCG'-binding anchor finger. Ultimately, a large amount of DNA-protein recognition data can be obtained using this medium-throughput method that will facilitate building more accurate 'recognition codes' for zinc fingers.

METHODS

Zif268 B1H selections

All of the B1H binding site selections were performed as described previously using an ω-Zif268 fusion protein expressed from a UV2 promoter in the plasmid p1352[92]. Zif268 was used for these experiments because it has been thoroughly characterized by a number of other methods allowing comparison of the recognition models we obtain to its previously defined specificity.

Randomized 6bp binding site library

The binding site library (GCGGCCACTGGGCAGCTGGCCA*NNNN*AAAAAT**NNNNNN**GCGGTACCTAGGT TCTTCGAATTC) cloned between the *EcoR*I and *Not*I sites in pH3U3 contains two different randomized regions: a 6bp element (bold underlined) that is associated with the four 3' bases of the Zif268 recognition sequence (GCGG, underlined), and a 4bp randomized region (italics) that serves as an internal control to identify sequences that may be enriched in the selections or preparation for sequencing sample due to jackpot effects. We did not observe any evidence of a jackpot effect. Self-activating clones within this library were removed by 5-FOA counter-selection as previously described[92]. Approximately 10^6 co-transformed cells containing the library and the ω -Zif268-expression plasmid were plated under each selection condition on selective media plates containing 0.5, 1 or 2 mM 3-AT and 0, 10 or 50 μM IPTG, where these selections were incubated at 37 °C for 4, 8, 12, 18 or 24 hours. This was a total of 45 independent selections. At the desired time-point surviving cells were washed off the plate as a pool. Isolated plasmid DNA from the pooled cells was prepared for Illumina sequencing as for the 28bp library. Using barcodes for each experiment, sequences from all 45 experiments

were obtained from a single Illumina sequencing lane that contained over 15 million reads, leading to an average of about 300,000 binding sites per experiment. There are only 4096 different 6mer sites so this quantity of sequences is sufficient for good coverage of all possible binding sites. We also performed CV-B1H from the same initial library in liquid media with 5 mM 3-AT and 50 μM IPTG. After 4 hrs the cells were pelleted, plasmids isolated and they were prepared for Illumina sequencing as with the experiments on plates. We independently sequenced the counter selected library, which was the input to each of the binding site selection experiments, to define the initial frequency of each 6mer. More than 16 million reads were obtained and every 6mer was observed at least 472 times. This allowed us to determine the enrichment of each site after selection. The sequences from each dataset are available at http://ural.wustl.edu/htb1h_zif68 and from the GEO database (GSE26767).

Binding site modeling using existing programs

We model the binding energy of Zif268 for any sequence using a position weight matrix (PWM)[219]. We used four different motif discovery methods on the different datasets. BioProspector[218] was used on the 6bp datasets with a site size of 6bp and a fixed orientation. A 3rd-order Markov model, based on the sequences of the respective initial libraries, was used for the background model. BEEML[117] was used with the background model derived from the 6mer counts in the initial library. We also tested a simple Log-Odds method that determines the value of each PWM element from the ratio of the observed frequency of each base at each position in the aligned binding sites to the

observed frequency of each base at each position in the initial library (from the randomized region).

Binding site modeling based on growth rate analysis

We model protein-DNA binding using a biophysical model described previously[117]. Briefly, the probability that the sequence S_i is bound at equilibrium is:

$$P(S_i\ bound) = \frac{[TF \cdot S_i]}{[TF \cdot S_i] + [S_i]} = \frac{[TF]}{[TF] + K_d(S_i)} \tag{1}$$

where K_d is the dissociation constant and square brackets indicate concentrations. It is convenient to express the energy of binding, E_i, relative to the Gibbs free energy of binding to a reference sequence; we use the consensus sequence, in units of RT, with its energy defined as, $E_{ref} = 0$:

$$P(S_i\ bound) = \frac{1}{1 + e^{E_i - \mu}} \tag{2}$$

where

$$E_i = \Delta\Delta G_i^{\circ} / RT = (\Delta G_i^{\circ} - \Delta G_{ref}^{\circ}) / RT \tag{3}$$

and

$$\mu = \ln \frac{[TF]}{K_d(S_{ref})} \tag{4}$$

Binding sites with $E_i = \mu$ have a binding probability of one-half.

In order to grow and replicate, cells must express sufficient His3 enzyme to meet their histidine requirements. We define the growth rate of an allele as the number of doublings that a cell possessing it undergoes each hour during exponential growth phase. The equation

$$N_i(t) = N_i(0)2^{r_i t} \tag{5}$$

describes the exponential growth of a colony, where t is the number of hours, $N_i(t)$ is the final number of cells possessing site S_i present at time t, r_i is the growth rate for cells containing that site in doublings/hr, and $N_i(0)$ is the initial number of cells with that site at time 0.

Histidine is a rate limiting reagent, and we make the simplifying assumption that the amount of histidine is directly proportional to the occupancy of the His3 promoter by the TF (up to some saturating level) and that the growth rate, r_i, of cells possessing S_i is directly proportional to the amount of histidine produced, up to a level where it is no longer limiting. The relationship between binding energy of the TF for site S_i and the growth rate is then:

$$r_i = \log_2\left(\frac{N_i(t)}{N_i(0)}\right)/t = \frac{M}{1+e^{E_i-\mu}} \tag{6}$$

where M is the maximum growth rate for these cells under the same conditions but with histidine not being limiting. Supplemental figure 1A shows a simulated ideal experiment where the counts for each sequence depend on the binding energies as described in the biophysical model of the preceding equations. Data taken at different time points will

fall on different curves, but when converted to growth rates all of the data sets converge to a common curve describing the relationship between growth rate and binding energy (Supplemental figure 1B).

We are only able to determine the frequency of each allele from the Illumina reads. In order to convert these frequencies into numbers of cells, we need to know the initial number of cells plated, n_I, and the final number of cells on the plate, n_F, at time t. The growth rates determined by the frequencies at time t will be off by a constant

$$c = \log_2\left(\frac{n_F}{n_I}\right)/t \tag{7}$$

such that

$$r_i = \log_2\left(\frac{f_i(t)}{f_i(0)}\right)/t + c \tag{8}$$

where $f_i(t)$ is the frequency of site S_i at time t, and $f_i(0)$ is the initial frequency of S_i before selection. We refer to the quantity

$$\frac{f_i(t)}{f_i(0)} \tag{9}$$

as the enrichment of site S_i at time t. For a given experiment, every growth rate will be off by the same constant. If we assume that the minimum growth rate is 0 (cells may not divide but they do not disappear from the plate) we can determine the constant by assuming the plateau of high energy binding sites represents a growth rate of 0. For the

remainder of the paper, including all of the figures, the calculated growth rates for each site have been adjusted such that the median of the high energy plateau is defined as 0.

For a given PWM, the predicted growth rates, $\hat{r}_i$, depends on the energy model via:

$$\hat{r}_i = \frac{M}{1 + e^{\vec{S}_i \cdot \vec{W} - \mu}} \tag{10}$$

where $\vec{S}_i$ is the encoded sequence, S_i, and $\vec{W}$ is the PWM. In this analysis, M was fixed to the maximum growth rate for each data set. We use the Levenberg-Marquardt algorithm[220-222] in a program called GRaMS (Growth Rate Modeling of Specificity) to perform a least squares fit between the measured and predicted growth rates in order to find the optimum PWM.

Assessment of different protocols and analysis methods

For each experimental dataset and each analysis method we obtain a position weight matrix (PWM). We adjust the elements such that those corresponding to the reference sequence are assigned 0, and the other elements are estimates of the binding energy differences for each other base at each position in the binding site, as proposed by Berg and von Hippel[223]. We determine the accuracy of each method by measuring, using the squared Pearson Correlation Coefficient (R^2), how well it predicts the binding data from a single-round SELEX experiment[117]. In that experiment a large library of random 10mers were bound to Zif268 and the bound fraction as well as the initial library were Illumina-sequenced. For each PWM the values of μ and a non-specific binding energy, E_{ns}, are found that maximize the fit for that model so that the comparisons are strictly between

how well the PWMs capture the energy differences for each base at each position. BEEML[117] was developed specifically to model that SELEX data so we determined its R^2 value when trained on the SELEX data directly as the maximum that any other PWM could be expected to obtain. This was 0.93 and 0.96 for the 10bp PWMs and 6bp PWMs, respectively. The remaining variance is probably due to experimental noise as well as binding energy contributions not captured by the simple PWM which are known to exist but be small for Zif268[224].

CHAPTER IV

AN OPTIMIZED TWO-FINGER ARCHIVE FOR ZFN-MEDIATED GENE TARGETING

Contents of Chapter IV have been accepted for publication

Ankit Gupta, Ryan G. Christensen, Amy L. Rayla, Abirami Lakshmanan,

Gary D. Stormo, Scot A. Wolfe (2012)

An optimized two-finger archive for ZFN-mediated gene targeting, Nat. Methods, 2012

Ryan Christensen from Gary Stormo's lab at the Washington University performed the W-log Odds and the GRaMS analysis for the 2F-modules and created sequence logos. Amy Rayla performed some of the B1H-selections. Abirami Lakshmanan performed the analysis of ZFN sites in zebrafish and human genomes.

Introduction

Targeted genome editing is an essential technology for reverse genetic analysis of gene function and for the creation of disease models. Custom-designed zinc finger nucleases (ZFNs) facilitate targeted genome modification in a variety of organisms and cell types by creating site-specific double strand breaks in DNA. Repair of ZFN-induced double-strand breaks by the error-prone non-homologous end joining (NHEJ) pathway leads to efficient introduction of insertion or deletion mutations (InDels) at the site of the double-strand break often resulting in loss of gene function. Alternatively, repair of a double-strand break by homology-directed repair with an exogenously introduced donor template can promote efficient and precise introduction of custom designed alterations at or near the break site. This technology has been successfully applied in a variety of cell lines[62,63,113,115,168-171] and organisms such as drosophila[167], zebrafish[94,114,172], C. elegans[173], rats[174], mice[65,175], pigs[116,176], arabidopsis[177], maize[178] and tobacco[179,180] many of which previously lacked efficient tools for targeted genome editing[162]. Due to their demonstrated utility and minimal off-target effects[160,184,185], ZFN-based therapies are being evaluated in clinical trials. Currently, widespread utilization of ZFNs is hindered by the challenge of designing zinc finger proteins (ZFPs) with sufficient affinity and specificity for a majority of DNA sequences within a genome.

ZFNs are engineered endonucleases that are composed of two domains: a DNA binding zinc finger protein (ZFP) and a C-terminal cleavage domain from the *Fok*I endonuclease. The incorporated ZFP is a tandem array of 3-6 zinc fingers each of which recognizes a roughly 3

bp DNA element. Since the nuclease activity requires dimerization of the *Fok*I domain, two ZFPs are designed where each binds to a total of 9-18 bp of DNA and combined they bind 18-36 bp of DNA specifying a unique address even in a complex genome. The ZFN activity and precision *in vivo* depends on the DNA binding specificity and affinity of the incorporated ZFPs[104,159,160,185] demonstrating the need for the ability to create highly specific ZFPs. ZFPs that are highly specific for a desired target site can be selected from randomized finger libraries using phage or bacterial selection systems[81,85,94,98], however this process is labor intensive. By contrast, modular assembly, wherein pre-characterized single zinc finger (1F) modules are joined into ZFPs, rapidly yields ZFNs[53,82,83,161,202]. Although 1F-modules have been described for almost all 64 DNA triplets[75,82,83], modular assembly has been successful only for a limited number of binding sites mostly consisting of 'GNN' type DNA triplets or 'N-G' type of 2 bp junctions between the finger subsites[53,100,103,104] **(Figure 4-1)**. The low success rates are presumably due to unfavorable interactions between the amino-acid residues at the finger-finger interface resulting in 'context-dependent' specificity of zinc fingers[104]. Apart from the labor-intensive selection of compatible fingers in their respective contexts[81,94,98], efforts to minimize unfavorable finger-finger interactions have focused on randomizing the interface residues and selection of two-finger (2F) modules that have compatible residues at the interface[84,85]. These 2F-modules have been used by Sangamo BioSciences to build highly specific ZFPs and active ZFNs[114,162], however, their archive is proprietary limiting its use to ZFNs purchased through Sigma-Aldrich. Recently, the Zinc Finger Consortium (ZFC) described a Context Dependent Assembly (CoDA) approach whereby 2F-modules selected from

Figure 4-1

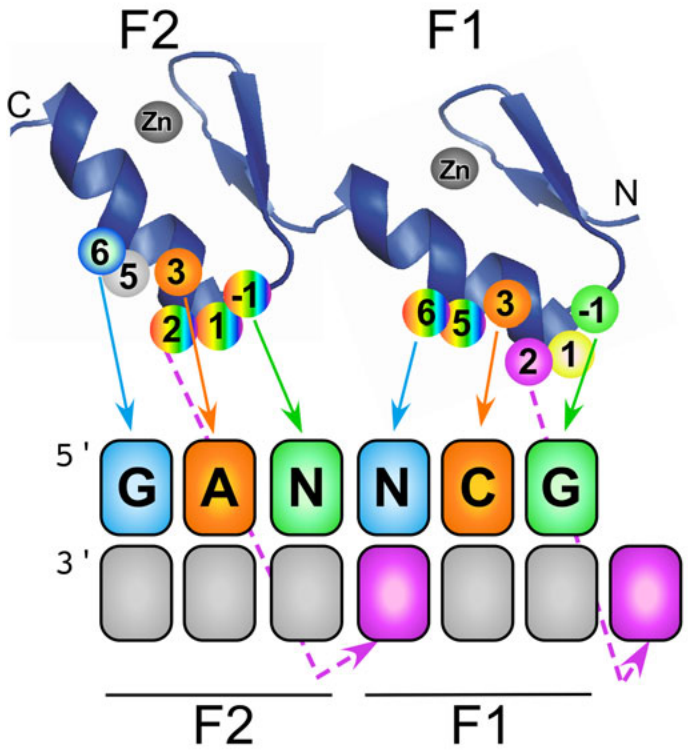


Figure 4-1: Schematic representation of the two-finger-ZFP library. Two orthogonal 2F-libraries were constructed each containing randomized amino acids at the finger-finger interface positions 5 and 6 of F1 (VNS randomization scheme) and positions -1, 1 and 2 of F2 (NNW randomization scheme). Amino acids at the remaining positions were fixed as following: F1: -1 = R, 1 = S, 2 = D, 3 = T, 4 = L; F2: 3 = N (Asn+3F2-library) or H (His+3F2-library), 4 = L, 5 = T, 6 = R. The F1 residues (R, S and D) at positions -1, 1 and 2 together represent the N-terminal cap. Fixed amino acids were chosen so as to bind GAN-NCG 6bp sites where N-N represents all 16 possible 2bp junctions.

OPEN pools are assembled into three-finger ZFNs[100]. CoDA-derived ZFNs constructed from prescreened ZFAs displayed higher success rates (~50%) than modularly assembled ZFNs, but the assayed ZFNs were almost entirely constructed from 2F-modules that recognize 'GNN-GNN' 6bp sites with 'N-G' type junctions that are not the limiting factor for advancing ZFN design (**Figure 4-2**).

In this study we have applied the bacterial-one-hybrid (B1H) based selections to create an archive of 87 two-finger modules that recognize 162 6-bp target sites. These 2F-modules, bind all possible 2 bp junctions including the 'non-N-G' junctions with high specificity and provide units for creating ZFNs. To assess their functional utility, we combined these 2F-modules with each other or with available 1F-modules[53] to create 3- or 4-finger ZFNs that demonstrated high success rate of gene editing in zebrafish.

Results

Selecting 2F-modules using B1H-based selections

We focused on selecting 2F-modules that recognize 2 bp interfaces in the 'GAN-NCG' context where 'N-N' represents the 2bp junction **(Figure 4-1)**. We constructed two two-finger libraries with randomized amino acids at the interface recognition positions; one with Asparagine (Asn+3F2 library) and the other with Histidine (His+3F2 library) at the position 3 of F2 for the recognition of Adenine at the second base position **(Figure 4-1)**. Both zinc finger libraries were fused with the engrailed homeodomain instead of a

Figure 4-2

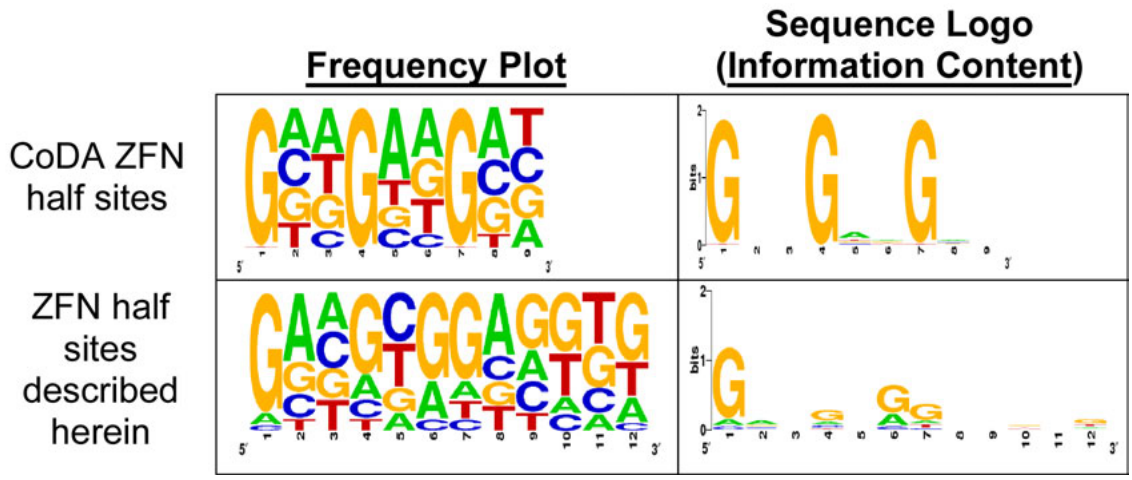


Figure 4-2: **Comparison of target site composition for CoDA-ZFNs against ZFNs described herein.** The half sites for ZFNs constructed in Sander *et. al.* using the CoDA strategy[100] and ZFNs constructed in this paper were compiled by aligning their 5' ends. Frequency plots and Sequence Logos displaying information content on a 2-bit scale were generated for each set of sites using Weblogo[208].

conventionally employed third anchor zinc finger **(Figure 4-3)**[19,84,85,100]. The fused homeodomain provides sufficient affinity to the two-fingers and avoids any interfering interactions from the anchor fingers. The selections were performed via bacterial-one-hybrid (B1H) based assay at multiple stringencies, modulated by 3-Amino triazole (3-AT), IPTG and uracil, to assess their influence on the distribution of selected residues. In general, higher stringency selections yielded fewer surviving colonies implying enrichment for clones that have higher specificity and affinity for the binding site. For 30 of 32 selections analysis of recovered clones revealed that a partial or full consensus sequence was obtained even at lower stringency, and seven of these selections displayed a tighter consensus at higher stringency **(Figure 4-6 and Table A-1)**. For two interfaces 'C-T' and 'T-T', selections with the His+3F2 library did not yield any consensus. Moreover, for 10 of 16 2bp interfaces, selections with both Asn+3F2 and His+3F2 libraries yielded similar clones suggesting that residues specifying these interfaces might be less dependent on the +3 residue of F2. Finally, achieving high stringency in the 'G-G' interface selections with the Asn+3F2 library required mutation of the anchoring homeodomain binding site, presumably to require greater specificity from the attached 2F-module. Future implementations of this selection scaffold can use this approach to further tune the selection of desired ZFP activities[45,225].

Examining the binding site specificities of 2F-modules

Previously it has been shown that the selected ZFPs might prefer a different binding site that the desired one [82,83,86]. To confirm that the recovered 2F-modules are specific for the

Figure 4-3

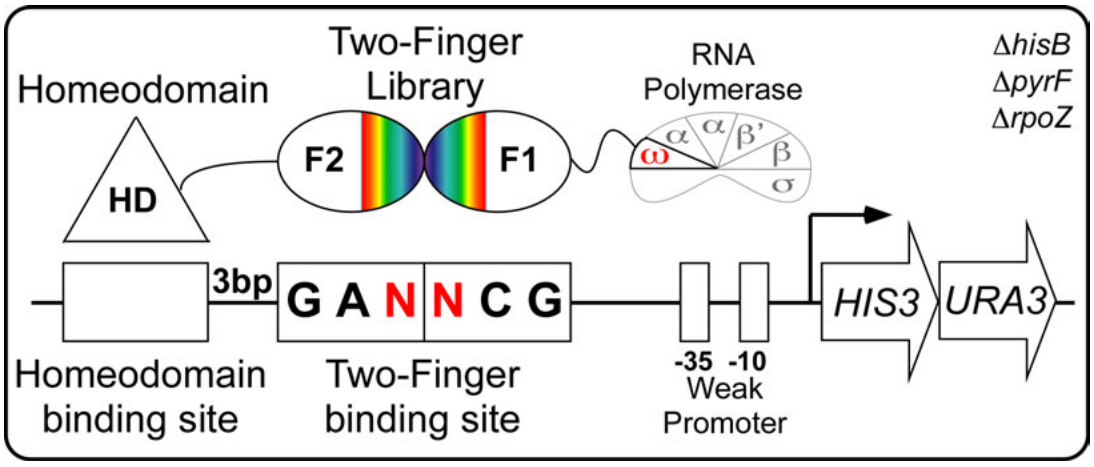


Figure 4-3: Schematic representation of bacterial-one-hybrid (B1H) based selections for 2F-modules. The B1H based zinc finger selection system employed here is a modified version of the one previously described[123]. The 2F-zinc finger libraries were fused to the DNA-binding domain of the Engrailed homeodomain on the C-terminus and the w-subunit of the RNA polymerase on the N-terminus. The fixed 6bp zinc-finger binding site is present on the His3/Ura3 reporter plasmid 3bp downstream of the homeodomain binding site and 9bp upstream of the -35 box.

desired interface, we determined the binding site specificities of 87 selected two-finger modules using the 'constrained variability-B1H (CV-B1H)' method[123] **(Chapter III, Figures 4-4 and 4-5)**. For 19 of 32 junctions, including 11 of 24 'non N-G' junctions, we identified 2F-modules that preferred the desired binding site **(Figure 4-6)**. As anticipated, most modules isolated from the His+3F2-library recognized both G and A at the second position of their 6bp binding sites **(Figure 4-6)**. Similarly, Threonine at the position 3 of finger-1 showed a preference both for C and T at the second position of the finger-1 target site. For the seven 2F-selections that displayed a more constrained consensus at higher stringency, clones from the higher stringency displayed improved sequence selectivity suggesting that increasing the stringency of selections yields modules with higher specificity (**Figure 4-7**). We also observed covariation of the residues at the interface for some of these selections indicating interactions across the finger-finger interface and selection of compatible groups of residues (**Figure 4-7**). Interestingly, clones with similar interface residues obtained from the Asn+3F2- and His+3F2-library selections showed similar sequence preferences for 9 of 10 2-bp junctions corroborating that the DNA recognition at the 2 bp junction by these residues might be less prone to influence by the type of residue at position 3 of finger-2 (**Figure 4-6**).

Improving the specificity of 2F-modules through rational design

A few selections yielded 2F-modules that, although compatible with the desired target site, preferentially recognized a different DNA sequence, a phenomenon that has been

Figure 4-4

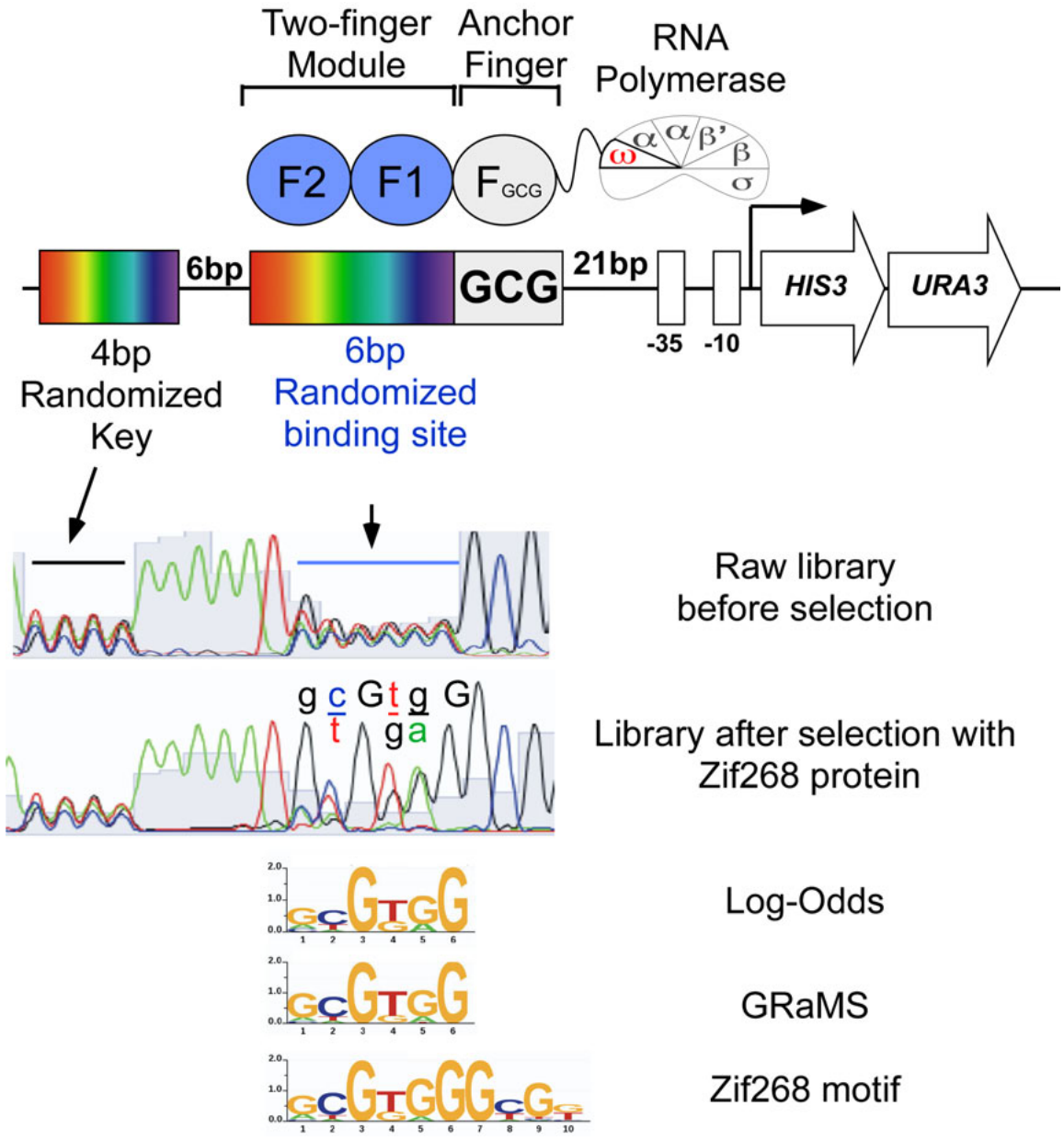

Two-finger Module
Anchor Finger
RNA Polymerase
F2
F1
F GCG
ω
α
α
β'
β
σ
6bp
GCG
21bp
HIS3
URA3
-35
-10
4bp Randomized Key
6bp Randomized binding site
Raw library before selection
Library after selection with Zif268 protein
Log-Odds
GRaMS
Zif268 motif

Figure 4-4: **Identification of DNA binding specificity for 2F-modules using the CV-B1H method**[123]. The 2F-module is fused to an N-terminal finger (RSDTLAR) that binds to the 'GCG' triplet adjacent to the 6bp randomized zinc finger binding region on the reporter plasmid. There is also a 4bp randomized region (key region) that serves as an internal control to identify biases in the recovered DNA sequences due to jackpot effects. Following selection, the surviving colonies are pooled and the distribution of bases recovered at each position within the selected binding sites can be evaluated in a single sequencing reaction as shown here for fingers 2 and 3 of Zif268. The recovered binding sites are determined by Illumina sequencing and then a binding site motif is calculated from these sequences using either log-odds-like or GRaMS (Growth Rate Modeling of Specificities) method[123]. For Zif268 F2 and F3, the binding site model obtained using the log-odds-like and GRaMS method closely matches the motif obtained by HT-SELEX[117].

Figure 4-5

(a)


W Log-Odds
GRaMS
W Log-Odds
GRaMS
F1: RSDTLTQ
F2: QRGNLTR
2FM-1*
A - A
F1: RSDTLAQ
F2: QSGNLTR
2FM-1001
A - A
F1: RSDTLKA
F2: QRGNLTR
2FM-1002
A - A
F1: RSDTLVE
F2: QRGNLTR
2FM-2*
A - C
F1: RSDTLIT
F2: QGGNLTR
2FM-1003
A - C
F1: RSDTLAD
F2: QHGNLTR
2FM-1004
A - C
F1: RSDTLVA
F2: QRGNLTR
2FM-1005
A - C
F1: RSDTLVR
F2: QEGNLTR
2FM-1006
A - G
F1: RSDTLQR
F2: QKSNLTR
2FM-3*
A - G
F1: RSDTLRL
F2: QRGNLTR
2FM-1008
A - T
F1: RSDTLKV
F2: QYGNLTR
2FM-1009
A - T
F1: RSDTLKI
F2: VRSNLTR
2FM-1010
A - T
F1: RSDTLKE
F2: VSRNLTR
2FM-1011
C - A
F1: RSDTLQE
F2: TARNLTR
2FM-1012*
C - A
F1: RSDTLVE
F2: CKPNLTR
2FM-53*
C - A
F1: RSDTLKQ
F2: DKRNLTR
2FM-1013*
C - A
F1: RSDTLKE
F2: ARRNLTR
2FM-1014*
C - A
F1: RSDTLGR
F2: ARRNLTR
2FM-1015
C - A
F1: RSDTLKQ
F2: CKPNLTR
2FM-1016
C - A
F1: RSDTLAE
F2: CARNLTR
2FM-1017*
C - A
F1: RSDTLKR
F2: CRGNLTR
2FM-1018
C - G
F1: RSDTLAR
F2: EGGNLTR
2FM-1019
C - G
F1: RSDTLER
F2: DRGNLTR
2FM-7*
C - G
F1: RSDTLVR
F2: ERGNLTR
2FM-1020
C - G
F1: RSDTLER
F2: ESGNLTR
2FM-1021
C - G
F1: RSDTLKL
F2: CGSNLTR
2FM-1022*
C - T
F1: RSDTLKG
F2: DRCNLTR
2FM-8*
C - T
F1: RSDTLKE
F2: GRSNLTR
2FM-1023
C - T
F1: RSDTLKE
F2: SKSNLTR
2FM-1024
C - T
F1: RSDTLAV
F2: DRSNLTR
2FM-1025*
C - T
F1: RSDTLKE
F2: CSRNLTR
2FM-1026
G - A
F1: RSDTLVE
F2: VHRNLTR
2FM-1027*
G - A

Figure 4-5 (a) contd.

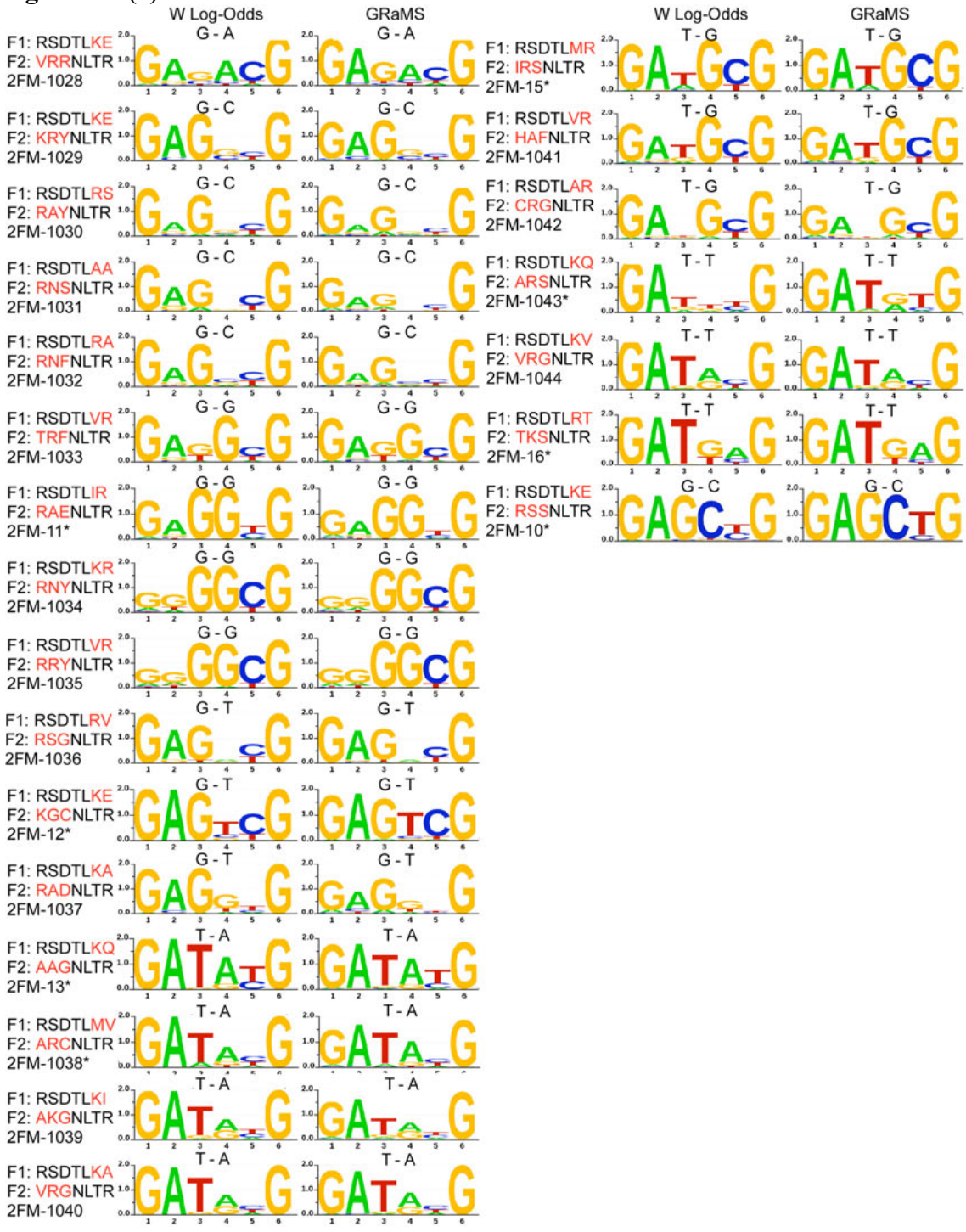


Figure 4-5 (b)

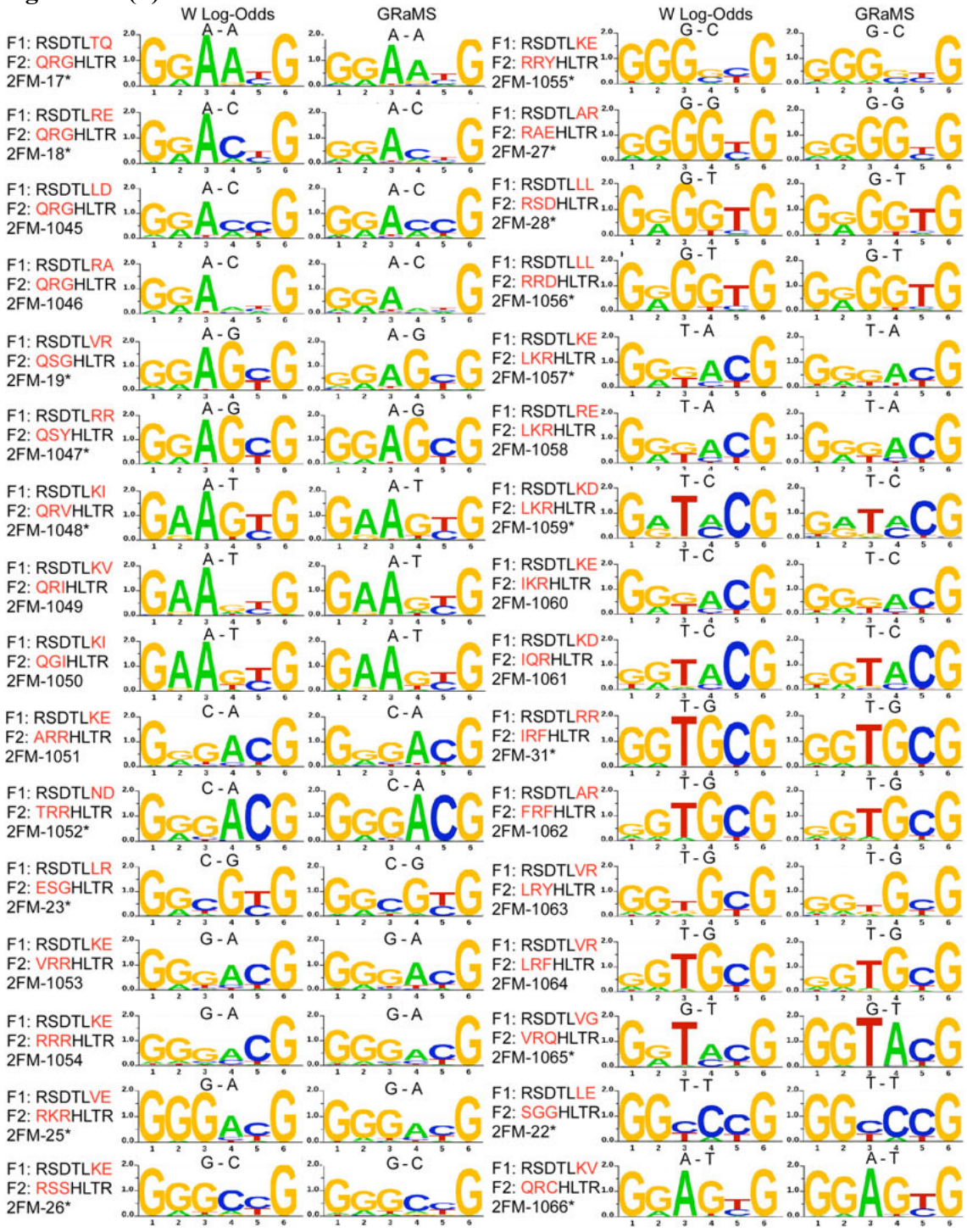

W Log-Odds
GRaMS
W Log-Odds
GRaMS
F1: RSDTLTQ
F2: QRGHLTR
2FM-17*
A - A
F1: RSDTLRE
F2: QRGHLTR
2FM-18*
A - C
F1: RSDTLLD
F2: QRGHLTR
2FM-1045
A - C
F1: RSDTLRA
F2: QRGHLTR
2FM-1046
A - C
F1: RSDTLVR
F2: QSGHLTR
2FM-19*
A - G
F1: RSDTLRR
F2: QSYHLTR
2FM-1047*
A - G
F1: RSDTLKI
F2: QRVHLTR
2FM-1048*
A - T
F1: RSDTLKV
F2: QRIHLTR
2FM-1049
A - T
F1: RSDTLKI
F2: QGIHLTR
2FM-1050
A - T
F1: RSDTLKE
F2: ARRHLTR
2FM-1051
C - A
F1: RSDTLND
F2: TRRHLTR
2FM-1052*
C - A
F1: RSDTLLR
F2: ESGHLTR
2FM-23*
C - G
F1: RSDTLKE
F2: VRRHLTR
2FM-1053
G - A
F1: RSDTLKE
F2: RRRHLTR
2FM-1054
G - A
F1: RSDTLVE
F2: RKRHLTR
2FM-25*
G - A
F1: RSDTLKE
F2: RSSHLTR
2FM-26*
G - C
F1: RSDTLKE
F2: RRYHLTR
2FM-1055*
G - C
F1: RSDTLAR
F2: RAEHLTR
2FM-27*
G - G
F1: RSDTLLL
F2: RSDHLTR
2FM-28*
G - T
F1: RSDTLLL
F2: RRDHLTR
2FM-1056*
G - T
F1: RSDTLKE
F2: LKRHLTR
2FM-1057*
T - A
F1: RSDTLRE
F2: LKRHLTR
2FM-1058
T - A
F1: RSDTLKD
F2: LKRHLTR
2FM-1059*
T - C
F1: RSDTLKE
F2: IKRHLTR
2FM-1060
T - C
F1: RSDTLKD
F2: IQRHLTR
2FM-1061
T - C
F1: RSDTLRR
F2: IRFHLTR
2FM-31*
T - G
F1: RSDTLAR
F2: FRFHLTR
2FM-1062
T - G
F1: RSDTLVR
F2: LRYHLTR
2FM-1063
T - G
F1: RSDTLVR
F2: LRFHLTR
2FM-1064
T - G
F1: RSDTLVG
F2: VRQHLTR
2FM-1065*
G - T
F1: RSDTLLE
F2: SGGHLTR
2FM-22*
T - T
F1: RSDTLKV
F2: QRCHLTR
2FM-1066*
A - T

Figure 4-5: **Binding site specificities of the B1H-selected 2F-modules.** The binding site specificities as determined by CV-B1H followed by log-odds-like and GRaMS modeling are shown for the B1H-selected 2F-modules obtained from the Asn+3F2 library (**a**) and His+3F2 library (**b**). The recognition helix sequences (positions -1, 1, 2, 3, 4, 5 and 6) for the F1 and F2 are shown. The selected interface residues are shown in red. The target 2bp junction is listed above each motif.

Figure 4-6

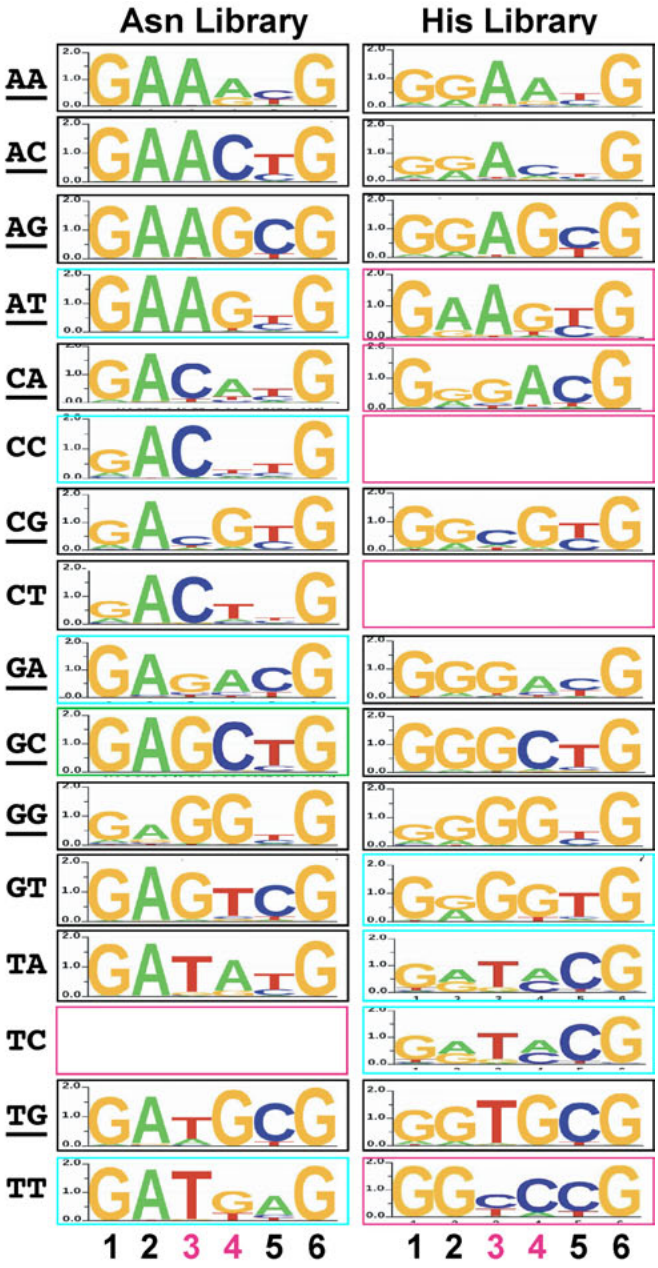


Figure 4-6: **Montage showing the binding site specificities of the best 2F-modules selected from the Asn+3F2 and the His+3F2 library for each 2bp junction.** The 2F-modules are designated as having 'preferential specificity' (black box), 'compatible specificity' (cyan box) or 'poor specificity' (magenta box) for the desired target sequence. The interfaces where ZFP selections with Asn+3F2 and His+3F2 libraries yielded clones with similar sequences are underlined.

Figure 4-7

A

GACTCG - 5 mM 3-AT

RSDTLKG	DRSNLTR
RSDTLAV	**DRSNLTR**
RSDTLVR	DPCNLTR
RSDTLRD	CRSNLTR
RSDTLSL	CRANLTR
RSDTLKL	GGSNLTR
RSDTLKM	NASNLTR
RSDTLKE	GRSNLTR
RSDTLKE	SKSNLTR

B

GACTCG - 10 mM 3-AT

RSDTLKL	**CGSNLTR**
RSDTLRL	CSSNLTR
RSDTLVL	CKSNLTR
RSDTLKL	CASNLTR
RSDTLQL	CRSNLTR
RSDTLAL	CRCNLTR
RSDTLKG	**DRCNLTR**
RSDTLAG	DRSNLTR

C

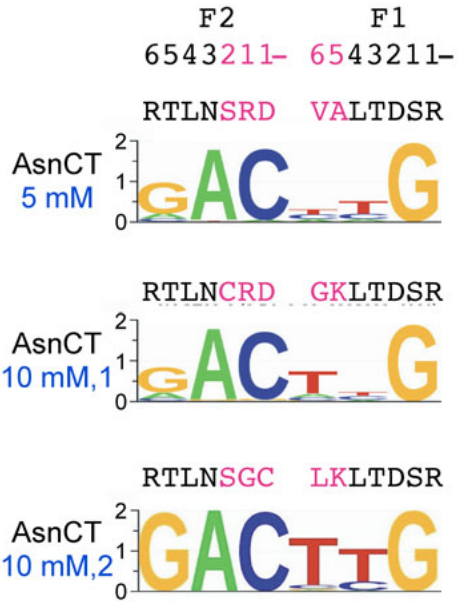


Figure 4-7: Enrichment of higher specificity 2F-modules at higher selection stringency. Displayed is the list of recognition helix sequences and frequency logos for the 2F-modules obtained from selections for the GA'CT'CG 6 mer with the Asn+3F2 library where the selections were performed at 5 mM 3-AT (a) and 10 mM 3-AT (b). Residue at positions (in red) 5 and 6 of finger 1 and -1, 1 and 2 for finger 2 were randomized. Clones that were selected for binding site specificity assessment are shown in bold. c) Comparison of specificities of 2F-modules selected at low stringency (5mM) and high stringency (10mM) for the 'CT' junction sequence.

observed in other zinc finger selections [82,83,86]. To improve the specificity of these 2F-modules, we employed rational design utilizing principles of DNA recognition derived from our B1H-selections. This successfully expanded the number of 2F-modules that preferentially bind the desired junction sequence to 24 with an additional 6 junctions that can be recognized by 2F-modules with 'compatible' specificity (**Figures 4-8, 4-9, 4-10 and Table A-2)**. Altogether, these 2F-modules can recognize a set of 60 'GRN-NYG' 6 bp sites owing to the specification of both A and G by Histidine at position 3 of finger-2 and C and T by Threonine at position 3 of finger-1.

Employing site-directed mutagenesis to expand the archive

Next, we hypothesized that if the combinations of interface residues specifying the 2 bp junctions are independent, we can increase the breadth of our 2F-modules simply by changing the specificity determinants fixed in our zinc finger libraries. First, we replaced the interface neighboring residues at position 3 of finger-2 (Asn and His) and finger-1 (Thr) with different residues for five different 2F-modules recognizing the 'N-A' junctions to obtain desired specificities at the positions neighboring the 2 bp junction. In many instances, desired alterations in specificity, as determined by the CV-B1H method, could be obtained through substitutions at specificity determinant positions (*e.g.* **Figure 4-11**) but majority of substitutions resulted in the alteration of specificity at the 2 bp junction suggesting non-independence of interface recognition **(Figure 4-12)**. In general, position 3 of finger-2 was more accommodating to substitutions that preserved the junction specificity than position 3 of finger-1 **(Figure 4-12)**. Using this approach we

Figure 4-8

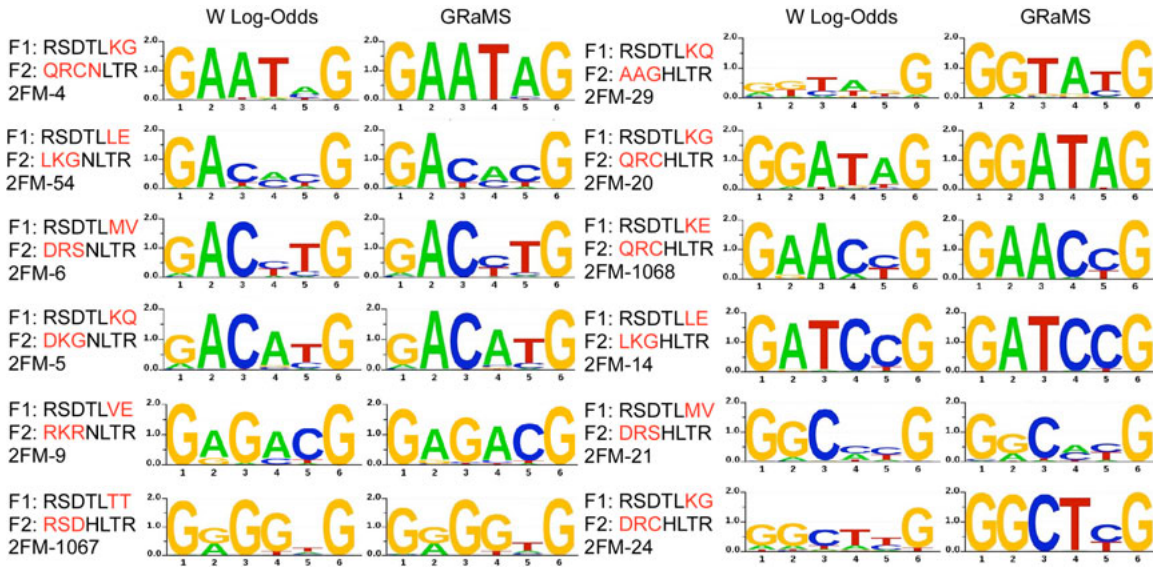


Figure 4-8: DNA-binding specificities for rationally designed 2F modules. The binding site specificities as determined by CV-B1H followed by log-odds-like and GRaMS modeling are shown for the rationally designed 2F-modules. The selected interface residues are shown in red. The target 2bp junction is listed above each motif.

Figure 4-9

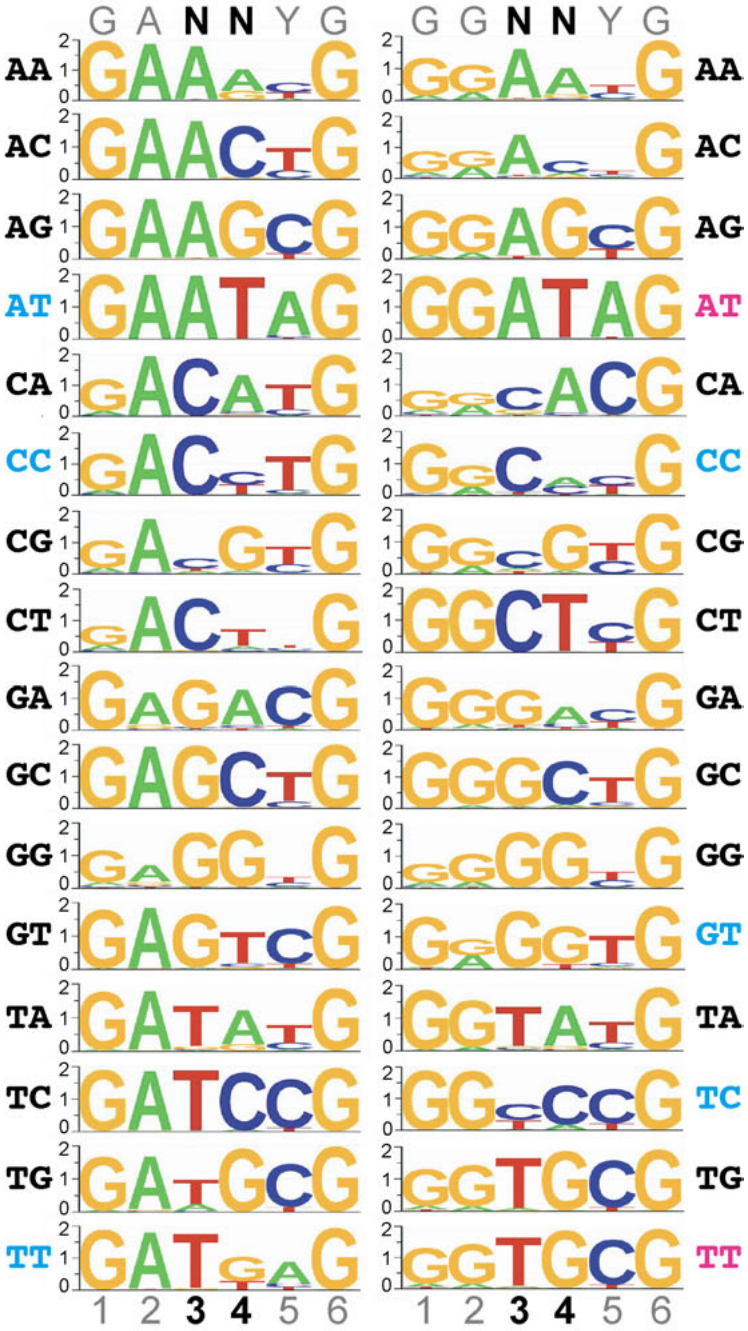


Figure 4-9: DNA binding specificities of selected and designed 2F-modules recognizing GRN-NYG sequences. Displayed is the montage of DNA-binding specificities for the most favorable 2F-modules with specificity for GAN-NYG and GGN-NYG sequences where N-N represents the sixteen 2bp junctions. The 2F-modules are designated as having 'preferential specificity' (black dinucleotide), 'compatible specificity' (blue dinucleotide) or 'poor specificity' (magenta dinucleotide) for the desired target sequence. Further details on these clones are found in **Figure 4-10**.

Figure 4-10

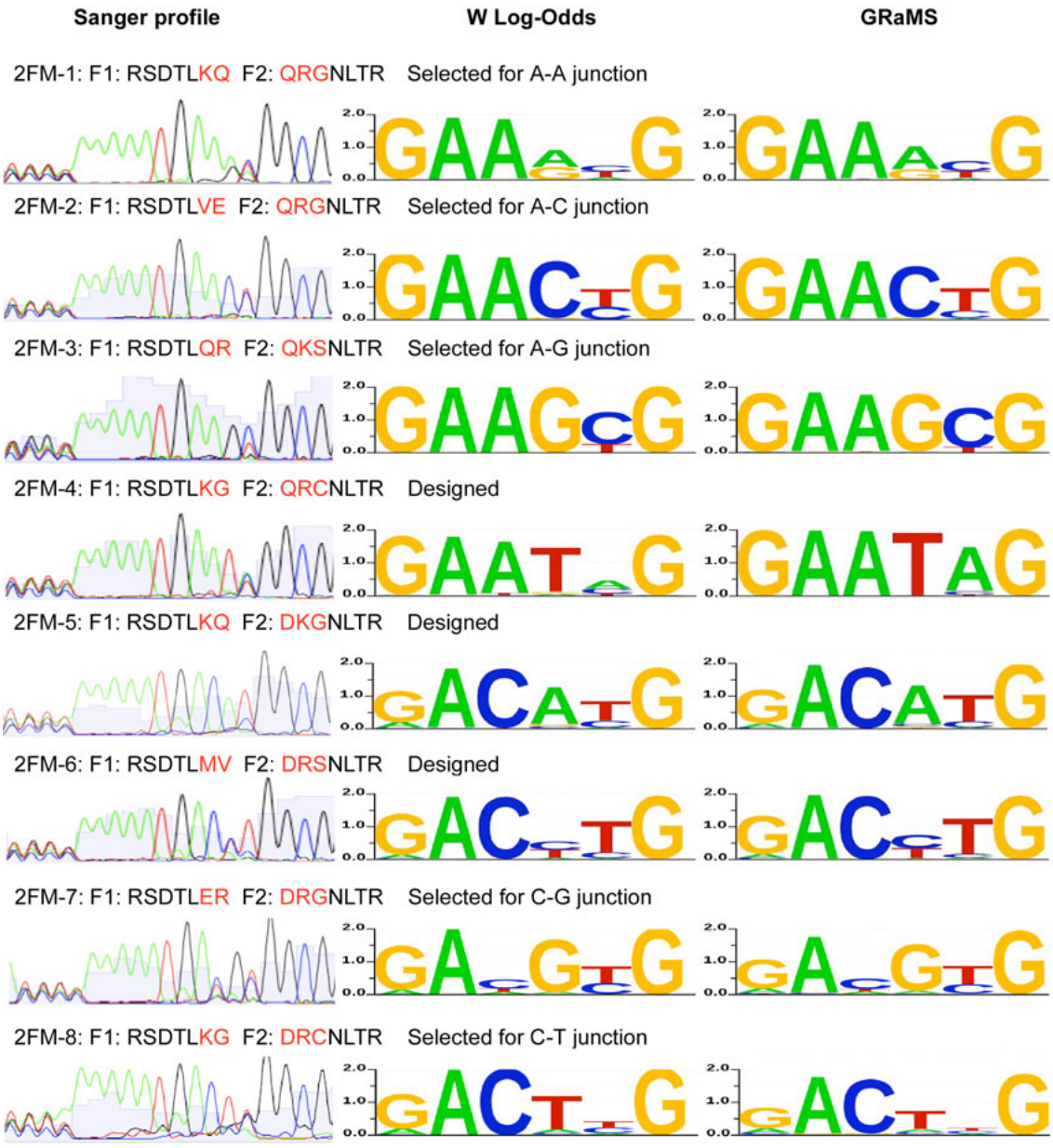

Sanger profile
W Log-Odds
GRaMS
2FM-1: F1: RSDTLKQ F2: QRGNLTR Selected for A-A junction
2FM-2: F1: RSDTLVE F2: QRGNLTR Selected for A-C junction
2FM-3: F1: RSDTLQR F2: QKSNLTR Selected for A-G junction
2FM-4: F1: RSDTLKG F2: QRCNLTR Designed
2FM-5: F1: RSDTLKQ F2: DKGNLTR Designed
2FM-6: F1: RSDTLMV F2: DRSNLTR Designed
2FM-7: F1: RSDTLER F2: DRGNLTR Selected for C-G junction
2FM-8: F1: RSDTLKG F2: DRCNLTR Selected for C-T junction

Figure 4-10 contd.

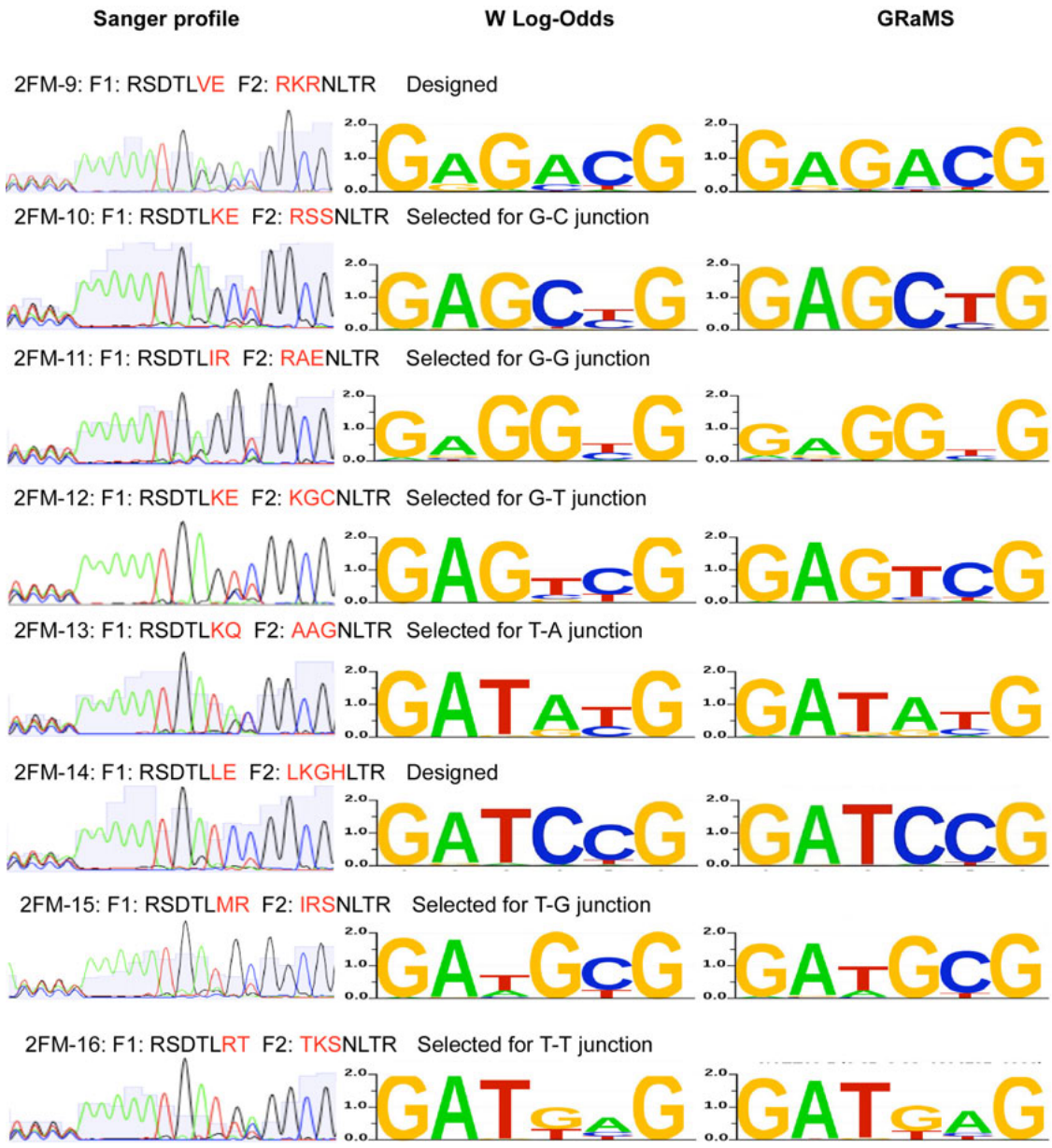


Figure 4-10 contd.

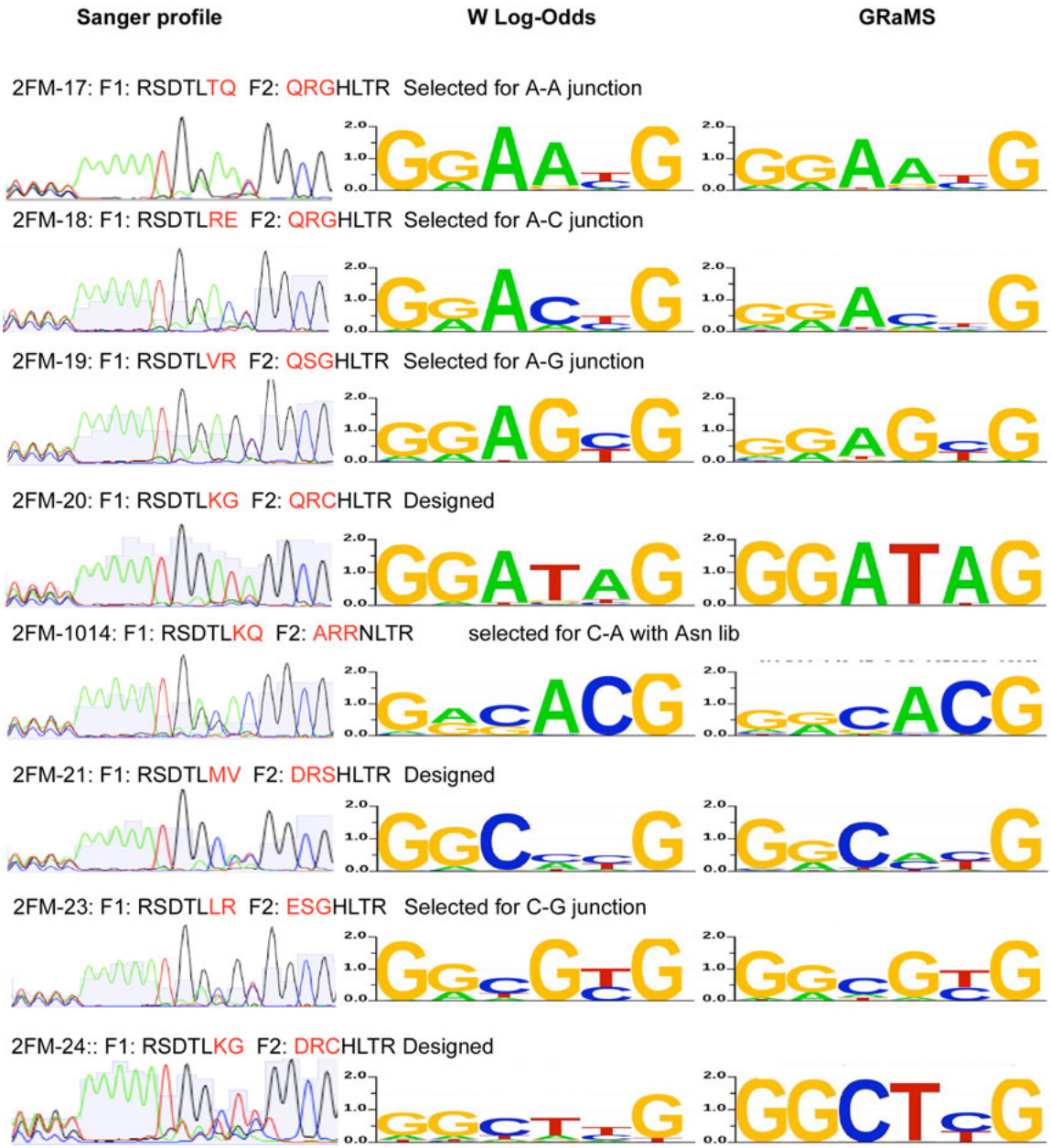


Figure 4-10 contd.

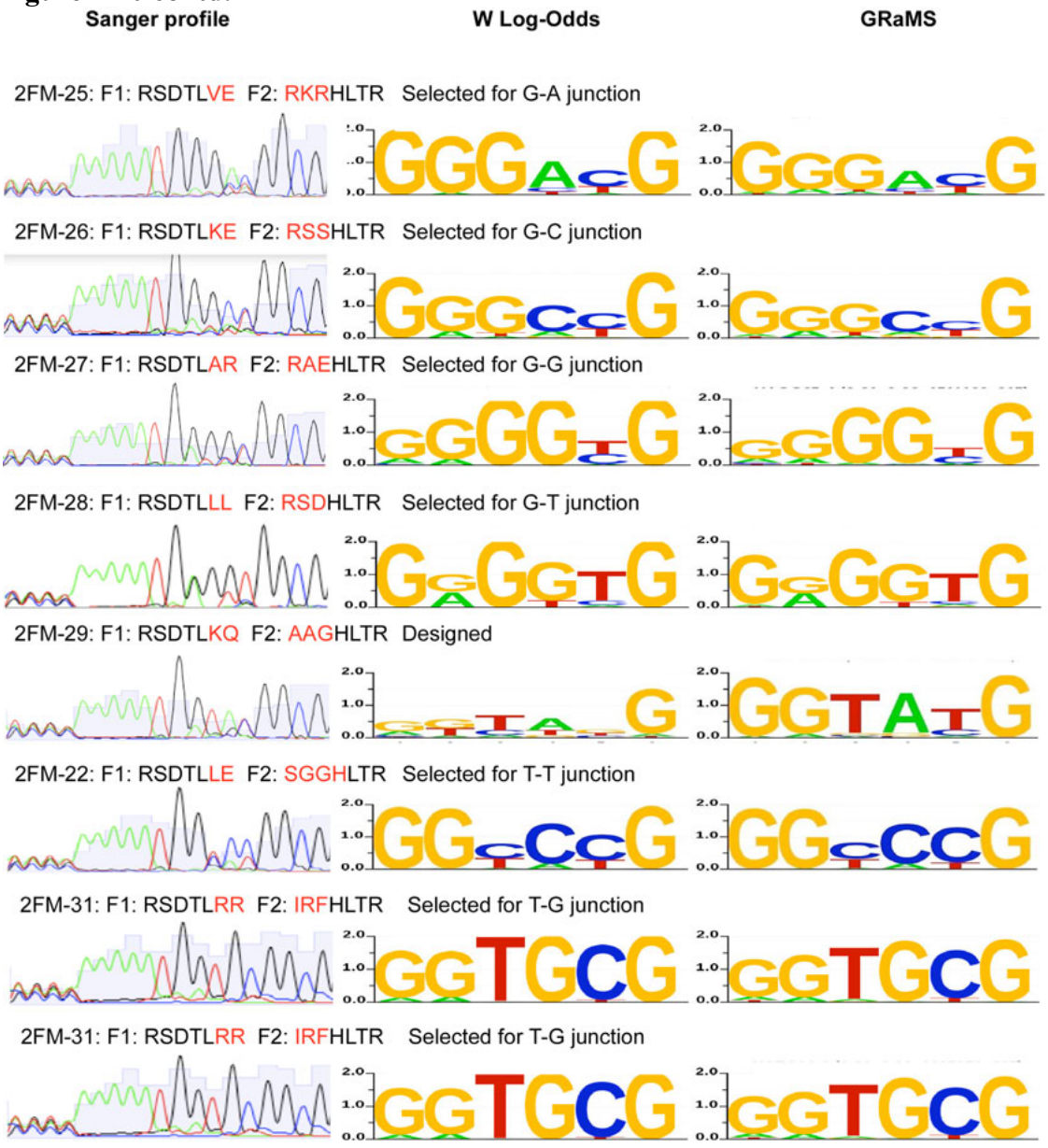


Figure 4-10: DNA-binding site specificities for 2F modules that bind GAN-NYG and GGN-NYG sequences. The 2F modules obtained via B1H-selections or rational design that bind each of 16 GAN-NYG and GGN-NYG sequences with highest specificity are displayed. The recognition helix sequences (positions -1, 1, 2, 3, 4, 5 and 6) for the F1 and F2 are shown. The selected interface residues are shown in red. Binding site specificities were determined using the CV-B1H method. The chromatograms representing the binding site profiles were obtained via Sanger sequencing of the pools of selected binding sites. Binding site logos were obtained via log-odds-like and GRaMS modeling post Illumina sequencing.

Figure 4-11

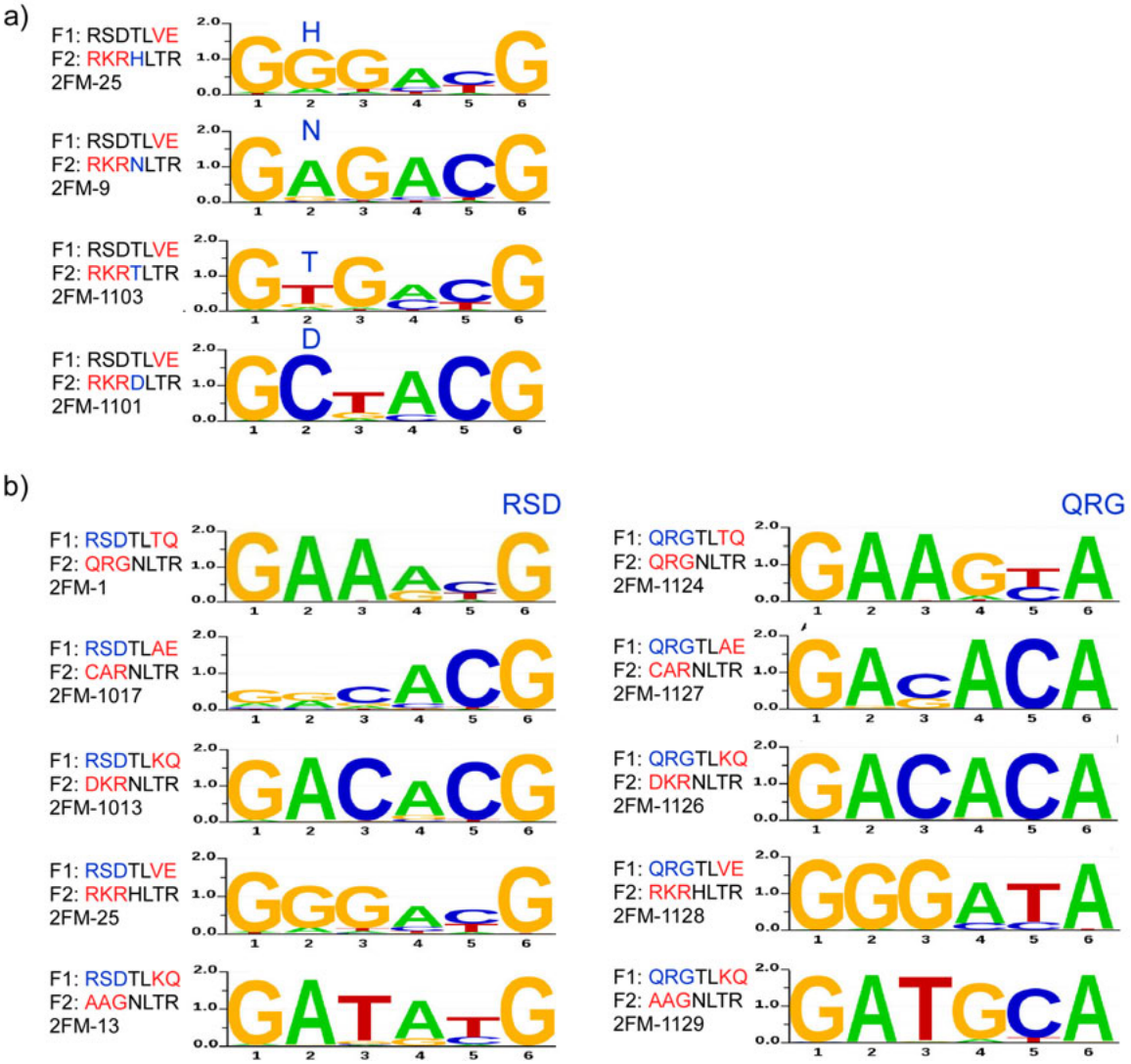


Figure 4-11: Expanding the archive of targetable sequences through rational design. The specificity determinants that were constant in the original libraries were replaced by other residue to expand the repertoire of targetable sequences. DNA binding specificity of new 2F-modules was determined using CV-B1H method and the logos were obtained using GRaMS modeling. (**a**) Examples of the influence of substitution of determinants at position 3 of finger-2 (shown in blue) on the specificity of the 2FM-25 2F-module. In three instances this results in a desired change in the specificity only at base 2, however in 2FM-1101 the introduction of Asp results in a change in the preference of base position 3, akin to the effects observed for the D20A mutation in Zif268[225]. (**b**) Substituting the N-terminal cap residues in finger-1 (RSD at positions -1, 1 and 2) with a QRG cap results in a concomitant change in base preference from G to A at the 6th base position without severely compromising the specificity for the junction sequence.

Figure 4-12

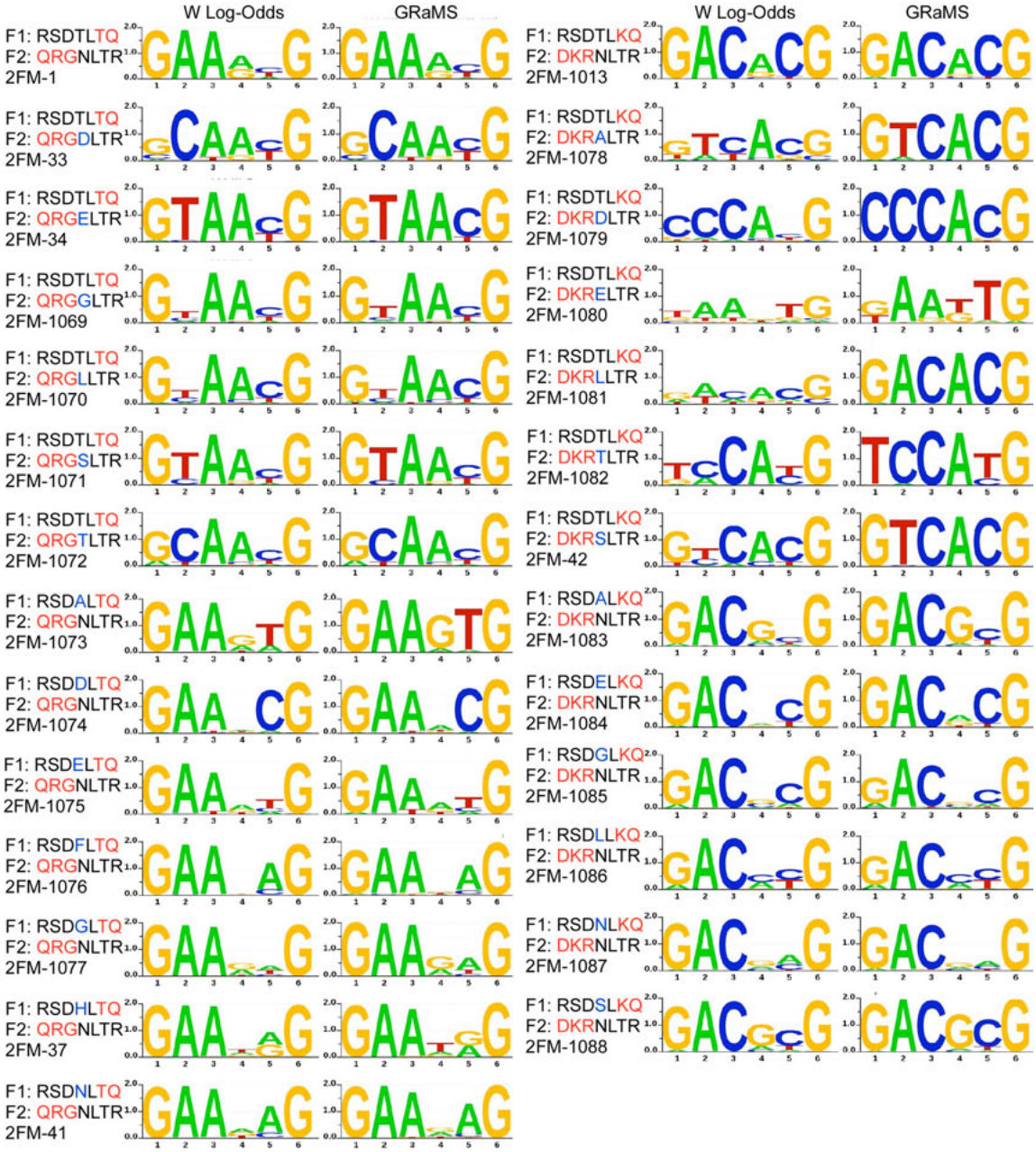

W Log-Odds
GRaMS
F1: RSDTLTQ
F2: QRGNLTR
2FM-1
F1: RSDTLTQ
F2: QRGDLTR
2FM-33
F1: RSDTLTQ
F2: QRGELTR
2FM-34
F1: RSDTLTQ
F2: QRGGLTR
2FM-1069
F1: RSDTLTQ
F2: QRGLLTR
2FM-1070
F1: RSDTLTQ
F2: QRGSLTR
2FM-1071
F1: RSDTLTQ
F2: QRGTLTR
2FM-1072
F1: RSDALTQ
F2: QRGNLTR
2FM-1073
F1: RSDDLTQ
F2: QRGNLTR
2FM-1074
F1: RSDELTQ
F2: QRGNLTR
2FM-1075
F1: RSDFLTQ
F2: QRGNLTR
2FM-1076
F1: RSDGLTQ
F2: QRGNLTR
2FM-1077
F1: RSDHLTQ
F2: QRGNLTR
2FM-37
F1: RSDNLTQ
F2: QRGNLTR
2FM-41
W Log-Odds
GRaMS
F1: RSDTLKQ
F2: DKRNLTR
2FM-1013
F1: RSDTLKQ
F2: DKRALTR
2FM-1078
F1: RSDTLKQ
F2: DKRDLTR
2FM-1079
F1: RSDTLKQ
F2: DKRELTR
2FM-1080
F1: RSDTLKQ
F2: DKRLLTR
2FM-1081
F1: RSDTLKQ
F2: DKRTLTR
2FM-1082
F1: RSDTLKQ
F2: DKRSLTR
2FM-42
F1: RSDALKQ
F2: DKRNLTR
2FM-1083
F1: RSDELKQ
F2: DKRNLTR
2FM-1084
F1: RSDGLKQ
F2: DKRNLTR
2FM-1085
F1: RSDLLKQ
F2: DKRNLTR
2FM-1086
F1: RSDNLKQ
F2: DKRNLTR
2FM-1087
F1: RSDSLKQ
F2: DKRNLTR
2FM-1088

Figure 4-12 contd.

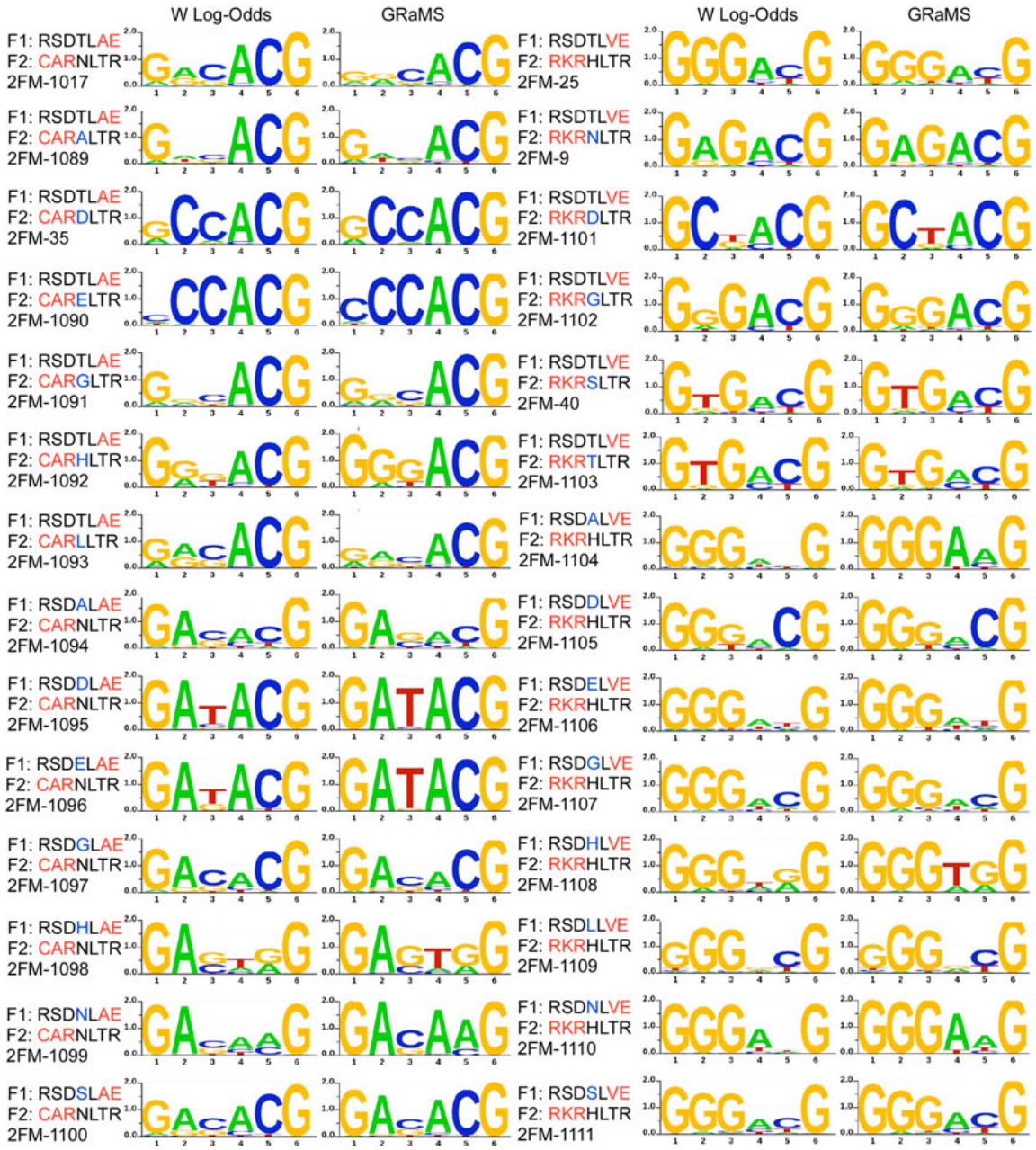


Figure 4-12 contd.

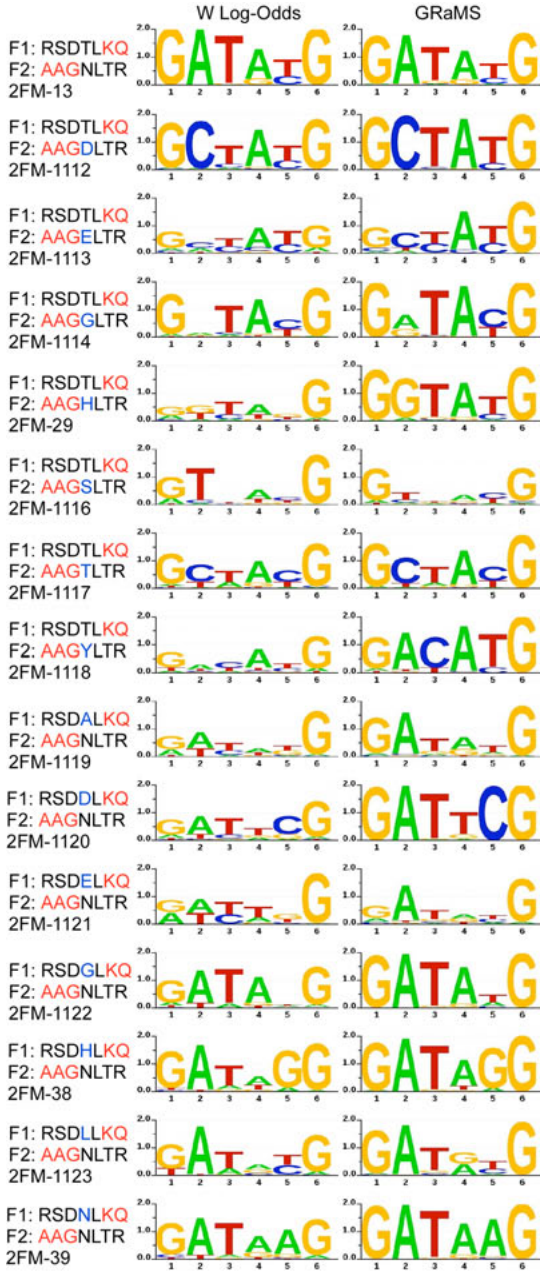


Figure 4-12: Specificities of all 2F-modules created by determinant substitution at position 3. 2F-modules with altered specificity were created by changing the residue (shown in blue) at position 3 of F1 or F2 to accommodate different bases at the 2nd or 5th position, respectively of the target site. DNA-binding specificities of the 2F modules were determined using CV-B1H method. The recognition helix sequence and specificity of each parent 2F module is displayed for comparison.

identified 2F-modules with substitutions at the position 3 residues on either finger-2 or finger-1 that recognize an additional fourteen 6-bp binding sites.

We also examined the impact of replacing distal residues (-1, 1 and 2) of finger-1 with motifs (N-terminal caps) that have been previously reported to bind different bases at the 3' position of the finger-1 triplet. CV-B1H Specificity analysis revealed that the QRG cap was the most reliable in both preferentially binding to Adenine at the 3' position and preserving the specificity at the 2 bp junction **(Figure 4-11 and Figure 4-13)**. Other substitutions either did not reliably specify the desired base and/or altered the specificity at the 2 bp junction. Employing these focused substitutions following B1H-selections and rational-design we identified 87 2F-modules that recognize a total of 162 6bp sites (**Table A-2**).

Comparison to the CoDA 2F-modules published by the ZFC

While some of our 2F-modules contain previously observed residues at the finger-finger interface[84,100], many contain novel combinations of residues. Of note, some CoDA 2F-modules recognizing 'N-A' junctions overlap in target preference with our 2F modules, yet display interface residues that might prefer alternate junction sequences[100]. We assessed five of these CoDA 2F-modules to investigate their sequence preferences using B1H binding site selections and activity assays. In this context the CoDA modules prefer 'N-G' instead of 'N-A' junctions highlighting the advantage of explicit optimization of the finger-finger interface for the generation of highly specific ZFAs (**Figure 4-14**).

Figure 4-13

Figure 4-13: Specificities of all 2F-modules created by changing the N-terminal cap. 2F-modules with altered specificity at the 6th base position were created by substitution at positions -1, 1 and 2 of finger-1 (N-terminal cap shown in blue). DNA-binding specificity of each module was determined using CV-B1H method. The recognition helix sequence and specificity of each parent 2F module is displayed for comparison.

Figure 4-14

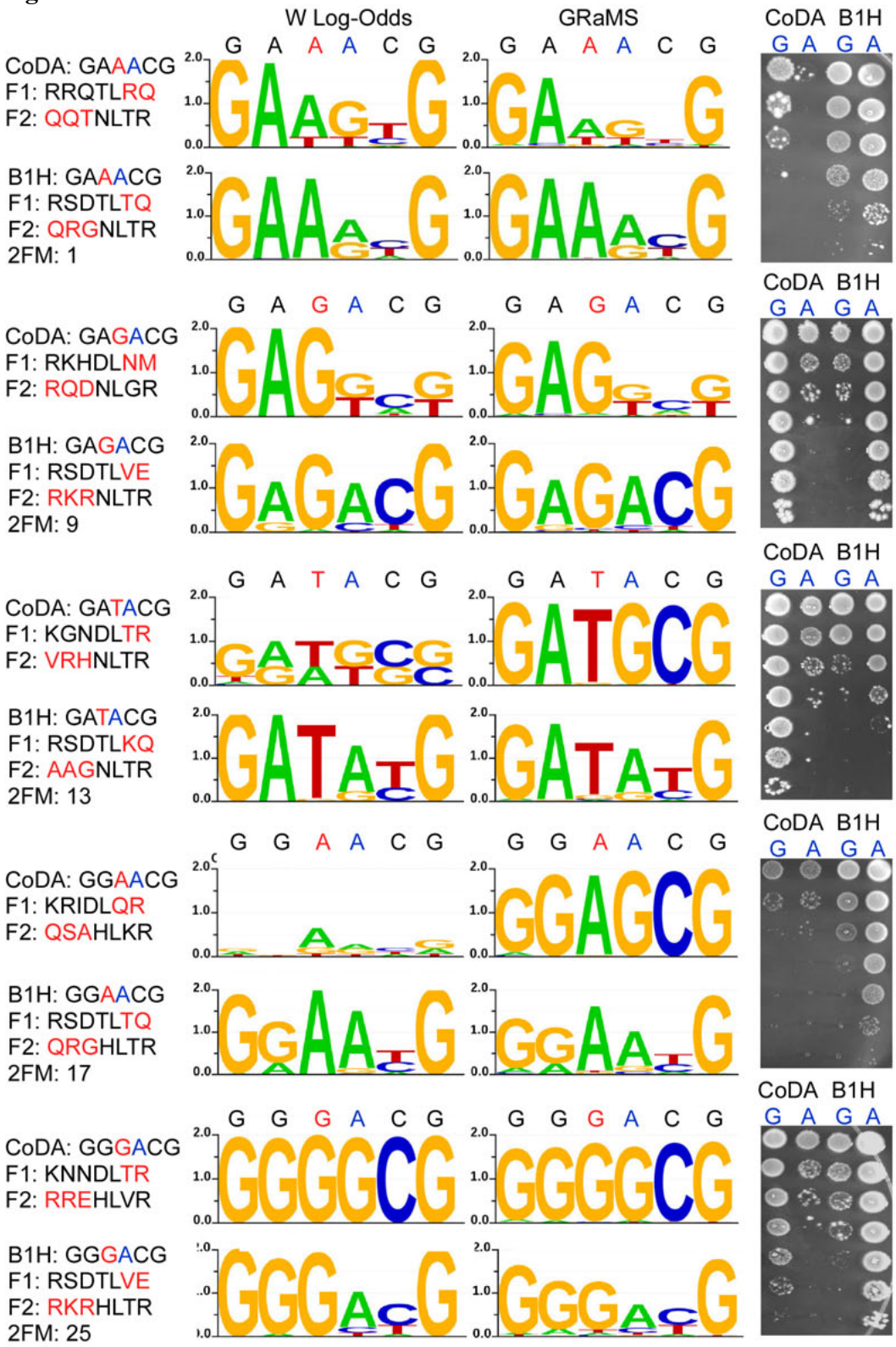

W Log-Odds
GRaMS
CoDA B1H
CoDA: GAAACG
F1: RRQTLRQ
F2: QQTNLTR
B1H: GAAACG
F1: RSDTLTQ
F2: QRGNLTR
2FM: 1
CoDA: GAGACG
F1: RKHDLNM
F2: RQDNLGR
B1H: GAGACG
F1: RSDTLVE
F2: RKRNLTR
2FM: 9
CoDA: GATACG
F1: KGNDLTR
F2: VRHNLTR
B1H: GATACG
F1: RSDTLKQ
F2: AAGNLTR
2FM: 13
CoDA: GGAACG
F1: KRIDLQR
F2: QSAHLKR
B1H: GGAACG
F1: RSDTLTQ
F2: QRGHLTR
2FM: 17
CoDA: GGGACG
F1: KNNDLTR
F2: RREHLVR
B1H: GGGACG
F1: RSDTLVE
F2: RKRHLTR
2FM: 25

Figure 4-14: Comparison of specificities of CoDA-2F modules[100] and our 2F modules. Five CoDA 2F modules were fused to the 'GCG'-binding finger-1 followed by their binding site analysis via the CV-B1H assay. The binding site logos obtained through GRaMS analysis are displayed for both the CoDA modules and the equivalent B1H-selected modules, where the recognition helices are shown for comparison. A B1H-activity assay was performed for the CoDA and B1H-selected 2F modules (in combination with the 'GCG'-binding F1) against fixed binding sites with either Adenine or Guanine at the 4th position (GAXYCG, where Y is either A or G) to determine the relative activity of the 2F module on each sequence variant. Each row in the assay represents 10-fold dilution of bacterial cells on plates containing the His3 inhibitor 3-AT to provide a stringent challenge to ZFA-driven reporter activity.

Assembling ZFNs using 2F-modules to target genes in zebrafish

To demonstrate the utility of these 2F-modules for gene disruption they were combined with each other or with published single finger (1F)-modules[53] to create ZFNs (3 pairs of 3-finger ZFNs and 8 pairs of 4-finger ZFNs) targeting 11 sites in the zebrafish genome, where each site contains at least one 'non-N-G' junction (**Table 4-1).** ZFNs targeting *dclk2* and *zgc:77041* genes were constructed exclusively from 2F-modules where the two 2F-modules were linked either with a conventionally used canonical 'TGQKP' linker or a non-canonical 'TG*S*QKP' linker that presumably shows better discrimination between consensus and mutant target sites and has been employed by Sangamo BioSciences to connect 2F-modules in their ZFNs[59-65,162]. The DNA-binding specificities of these assembled zinc finger proteins, as determined using B1H-based selections, revealed that the recognition preferences of most of the incorporated 2F-modules were consistent with the CV-B1H analysis (**Figures 4-15 and 4-17a**).

Activity of these ZFNs was initially tested using a published yeast based reporter assay[114] and compared to the activity of a positive control ZFN pair that shows high activity in both yeast assay and zebrafish. The yeast assay demonstrated that 9 of 11 ZFNs were active and the majority displayed activities comparable to the positive control ZFN **(Figure 4-16)**. Although, *lepr* ZFNs showed low activity at increased ZFN concentration, *rock1* ZFNs remained inactive even at higher ZFN concentration. Moreover, inserting the 'TG*S*QKP' non-canonical linker in one or both ZFN monomers favorably impacted both the activity and the toxicity of *dclk2* and *zgc:77041* ZFNs

Table 4-1: **List of all ZFNs and their target sites.**

Gene	Target site	ZFNL binding site	ZFNR binding site	Spacer (bp)	non-GNN fingers	non-N-G junct-ions	ZFNL-F0	ZFNL-F1	ZFNL-F2	ZFNL-F3	ZFNR-F0	ZFNR-F1	ZFNR-F2	ZFNR-F3
dab2ip	GTCCGAGTCcctgtaGACATGGAC	**GACTCG**GAC	**GACATG**GAC	6	2	2		LKGNLTR	**RSDTLKG**	**DRCNLTR**		LKGNLTR	**RSDTLKQ**	**DKGNLTR**
hey2	AACCATACCgaccgtGGGGAACTG	**GGTATG**GTT	GGG**GAACTG**	6	2	2		TSGSLSR	**RSDTLKQ**	**AAGHLTR**		**RSDTLVE**	**QRGNLTR**	RSDHLTR
rock1	CGTGCCAGCtgctccGACTCGGCC	GCT**GGCACG**	**GACTCG**GCC	6	2	2		**RSDTLQE**	**TARNLTR**	HRQSLTR		DRSDLSR	**RSDTLKG**	**DRCNLTR**
zgc77041	CAACTCCAGTTCattttgGACATGGGTACG	**GAACTGGAGTTG**	**GACATGGGTACG**	6	4	4	**RSDTLKE**	**KGCNLTR**	**RSDTLVE**	**QRGNLTR**	**RSDTLKD**	**LKRHLTR**	**RSDTLKQ**	**DKGNLTR**
dclk2	CCATTCCGGCTCtcgggGAAACGGGATCG	**GAGCCGGAATGG**	**GAAACGGGATCG**	5	4	4	**RSDHLTQ**	**QRGNLTR**	**RSDTLKE**	**RSSNLTR**	**RSDTLKG**	**QRCHLTR**	**RSDTLTQ**	**QRGNLTR**
mc4r	TTCATCTGCAGCttggctGTAGCAGACTTG	GCTGCAGATGAA	GTAGCA**GACTTG**	6	1	1	QKCNLVR	HRNNLTR	QSGDLTR	HRQSLTR	**RSDTLKG**	**DRCNLTR**	QSGDLTR	QSGALTR
lrp8	GCCTGCCCCGACagcatgGAACTGGGCCCG	GTCGGGGCAGGC	**GAACTGGGCCCG**	6	2	2	EKSHLTR	QSGDLTR	RSDHLTR	DRSALAR	**RSDTLMV**	**DRSHLTR**	**RSDTLVE**	**QRGNLTR**
mc3r	CAGCTCCACGTCagcgtaGCTGCATGAAGG	**GACGTGGAGCTG**	GCTGCATGAAGG	6	3	3	**RSDTLKE**	**RSSNLTR**	**RSDTLER**	**ESGNLTR**	RSDHLTQ	QSSHLTQ	QSSHLTQ	HRQSLTR
apoeb	AACCTACGCACCtctctGGAGGGCCGTGT	**GGTGCG**TAGGTT	GGA**GGGCCG**TGT	5	3	3	TSGSLSR	RSDNLTQ	**RSDTLRR**	**IRFHLTR**	LRHHLVG	**RSDTLKE**	**RSSHLTR**	QRGHLTR
lepr	ATCCAATTCCTTgcttcaGCAGTGGAAGGT	AAG**GAATTG**GAT	GCAGTGGAAGGT	6	2	1	TSGNLTR	**RSDTLKG**	**QRCNLTR**	RSDNLTQ	CAHHLTR	QKCNLVR	RSDALTR	QRSTRKR
irs2	TACCATGCCCCTctgtatCAGGAAAAGGTA	AGG**GGCATG**GTA	CAG**GAAAAG**GTA	6	4	2	QSGALTR	**RSDTLKE**	**ARRNLTR**	RSDHLTQ	QSGALTR	**RSDNLTQ**	**QRGNLTR**	RSDNLSE

For each ZFN target site, the ZFNL and ZFNR sites are shown in uppercase letters whereas the spacer sequences are shown in lowercase letters. The number of non-GNN and non-N-G junctions in each target site is provided. Also the recognition helix sequences (-1, 1, 2, 3, 4, 5, 6) for each ZFN are provided with the sequences of 2F-modules highlighted in bold.

Figure 4-15

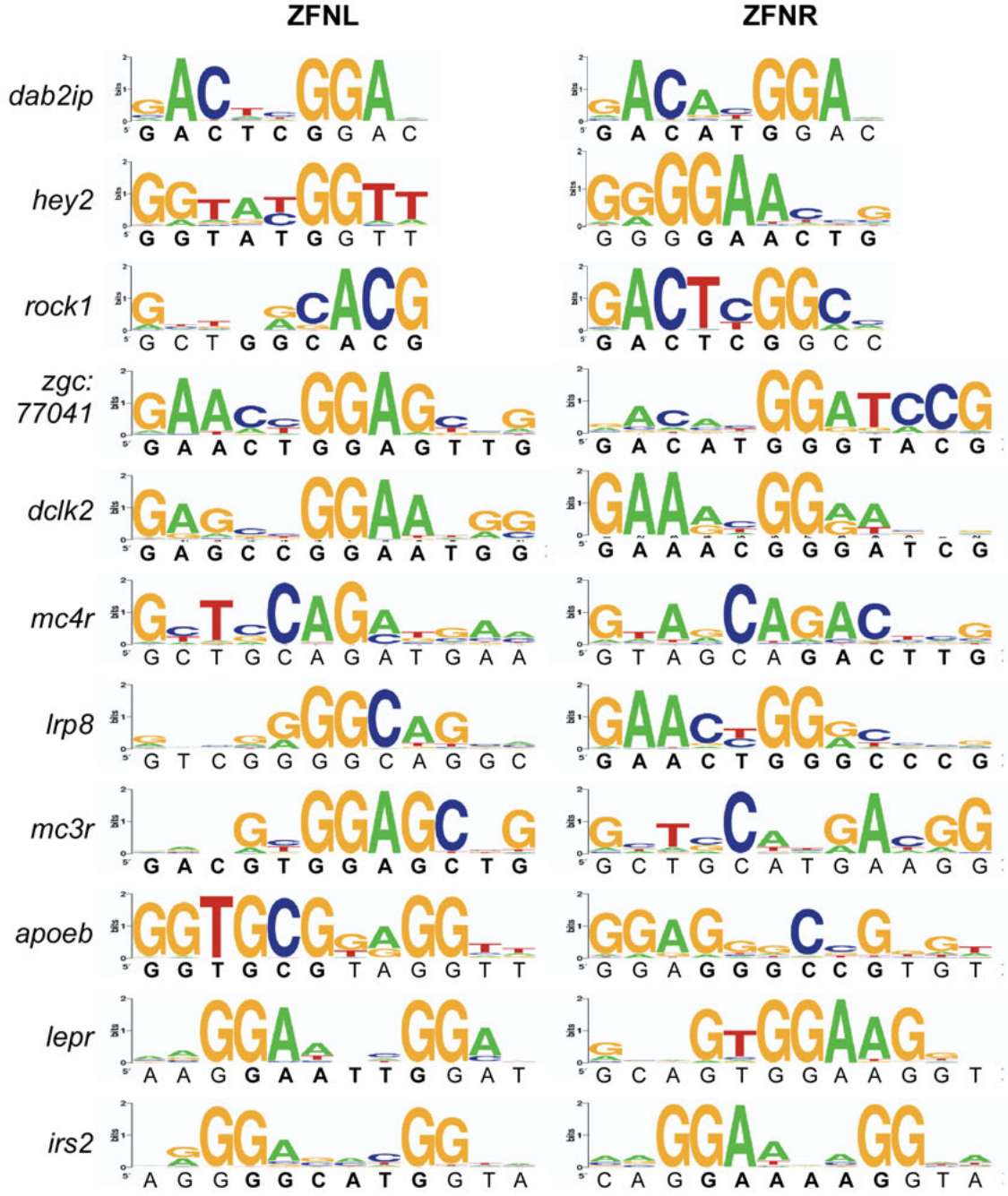


Figure 4-15: Binding site specificities of ZFAs incorporated into each ZFN pair. The binding site specificities for the ZFAs incorporated into ZFNs were determined via a B1H assay using the randomized 28bp library followed by Illumina[90]. The desired target sites are provided below each Sequence logo here the portion recognized by a 2F-module is highlighted in bold. The target gene is listed to the left of each ZFN pair.

Figure 4-16

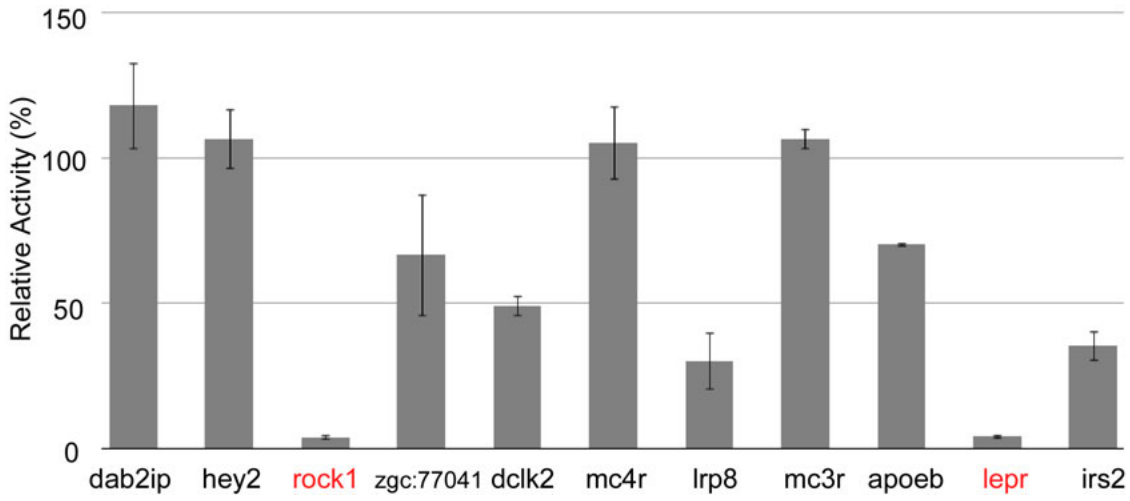


Figure 4-16: Assessment of ZFN activity using the yeast based chromosomal reporter assay. The test ZFN target site along with the target site for the positive control ZFN was integrated into the yeast genome where ZFN activity is measured by the reconstitution of a-galactosidase activity[114]. ZFN expression was induced by treating yeast cells with galactose for 30min. The activity relative to the positive control ZFN pair that yields ~10% lesion frequency in zebrafish is displayed as a mean of three experiments. Bars represent standard deviation. The rock1 and lepr ZFNs (shown in red) were inactive based on comparison to a GFP control.

Figure 4-17

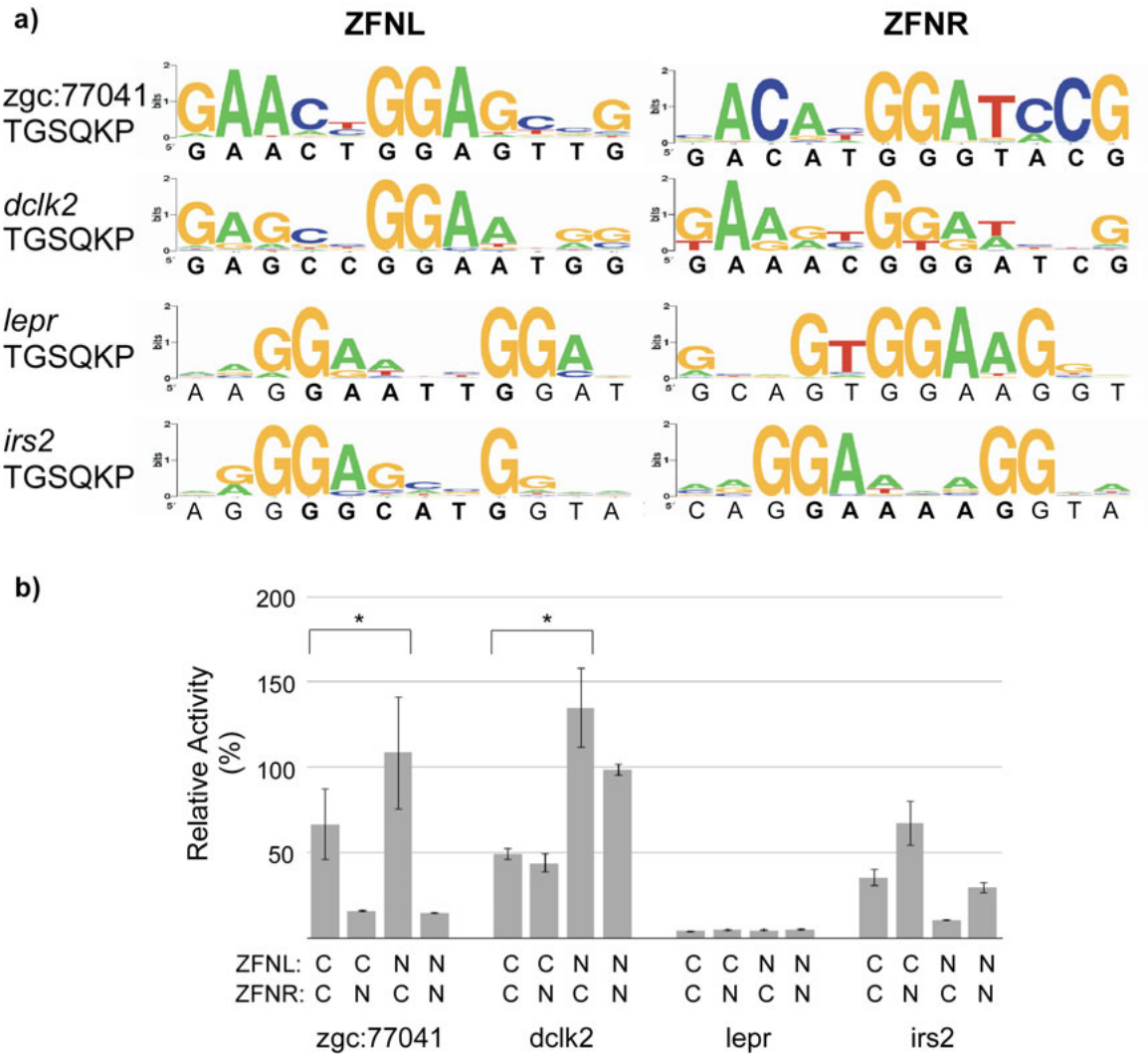


Figure 4-17: Influence of non-canonical linker on ZFN specificity and activity. (**a**) For 4 4F-ZFNs, the canonical linker ('C') between the 2nd and the 3rd finger was replaced by a non-canonical linker ('N'; TG*S*QKP). DNA binding specificities were determined for the modified ZFNs using the B1H assay as previously described. (**b**) The activity of the modified ZFNs with non-canonical linker was assessed using the yeast based reporter assay. The activities are relative to the positive control ZFN activity and is displayed as a mean of three experiments. Bars represent standard deviation and * indicates $P < 0.05$ as determined by paired student's *t*-test.

presumably by moderating the ZFN activity at non-target sequences **(Figure 4-17b)**. Consequently, we substituted the linker to 'TG*S*QKP' for *lepr* ZFNs that initially showed low activity and for *irs2* ZFNs that displayed high toxicity. In both the cases, the activity and toxicity were improved, although not as dramatically, by the linker substitution suggesting that the non-canonical linker might help improve the efficiency and precision of some ZFNs **(Figure 4-17b)**.

Next, we injected mRNAs encoding these ZFN into zebrafish embryos and then evaluated the induction of insertions and deletions (InDels) at the target site. 9 of 11 ZFNs induced InDels (>1bp) at the target sites at >0.5% frequency (**Table 4-2**). Consistent with the yeast assay, lesion frequency in zebrafish was also improved for certain ZFNs when incorporating the non-canonical 'TG*S*QKP' linker (**Table 4-3**).

Further, we also tested the influence of the DNA binding specificity of individual 2F-modules on the specificity of the assembled ZFPs and activity of the ZFNs. For the *dab2ip* ZFNR monomer, we incorporated five different 2F-modules that were either selected or rationally designed to bind to the same 'C-A' junction but display different specificities for the 'GAC-AYG' sequence **(Figure 4-18)**. The specificity of the assembled ZFPs as determined using B1H-selections on the 28 bp library[92], displayed a high correlation ($R^2 = 0.86$) with the specificity of the individual 2F-modules. Moreover, as ZFNs their activity in the yeast assay as well as in zebrafish correlated with their specificity both as individual 2F-modules and as assembled ZFPs implicating the need for highly specific ZFPs to create highly active ZFNs **(Figure 4-18)**.

Table 4-2

Gene	5p ZFP binding site	3p ZFP binding site	Spacer length (bp)	Number of Sequences with Indels	Number of wild type sequences	Lesion Frequency (%)	Most frequent Deletion	Most Frequent Insertion
dab2ip	**GACTCG**GAC	**GACATG**GAC	6	26703	334851	8.0	9bp (9198)	4bp (1471)
hey2	**GGTATG**GTT	GGG**GAACTG**	6	3438	552924	0.6	4bp (706)	4bp (1234)
rock1	GCT**GGCACG**	**GACTCG**GCC	6	191	384243	0.0	3bp (182)	None
*zgc77041**	**GAACTGGAGTTG**	**GACATGGGTACG**	6	49640	317017	15.7	9bp (7255)	2bp (8403)
*dclk2**	**GAGCCGGAATGG**	**GAAACGGGATCG**	5	2370	212738	1.1	2bp (656)	4bp (164)
mc4r	GCTGCAGATGAA	GTAGCA**GACTTG**	6	128638	998060	12.9	5bp (31856)	4bp (12193)
lrp8	GTCGGGGCAGGC	**GAACTGGGCCCG**	6	53297	732947	7.3	9bp (6780)	4bp (3534)
mc3r	**GACGTGGAGCTG**	GCTGCATGAAGG	6	24520	792371	3.1	5bp (5209)	4bp (5012)
apoeb	**GGTGCG**TAGGTT	GGA**GGGCCG**TGT	5	11180	396708	2.8	2bp (3507)	2bp (185)
*lepr**	AAG**GAATTG**GAT	GCAGTGGAAGGT	6	12264	1412846	0.9	4bp (8617)	4bp (266)
*irs2**	AGG**GGCATG**GTA	CAG**GAAAAG**GTA	6	2634	742945	0.4	6bp (969)	4bp (235)

Table 4-2: Analysis of ZFN-induced lesions in zebrafish.

ZFN target sites and the genes are shown. ZFNL and ZFNR sites are given wherein the 6bp subsites for the 2F-modules are represented in bold. Lesion frequencies and the most frequent insertion and deletion are shown where the number in parentheses shows their frequency. An asterisk indicates targets where a non-canonical linker (TG**S**QKP) between the second and the third finger was employed to increase ZFN activity, where the position of the non-canonical linker is underlined in each half-site where it is present.

Table 4-3

	Lesion Frequency (%)				
	TGQKP (C)	TGQKP (C)	TG***S***QKP (N)	TG***S***QKP (N)	ZFNL Linker
Target gene	TGQKP (C)	TG***S***QKP (N)	TGQKP (C)	TG***S***QKP (N)	ZFNR Linker
zgc77041	4.4	4.1	12.3	15.7	
dclk2	0.1	0.5	0.3	1.1	

Table 4-3: Influence of non-canonical linker on ZFN activity in zebrafish. The lesion frequencies (in %) in zebrafish are shown for different combinations of ZFNL and ZFNR with canonical (TGQKP; C) and non-canonical (TG***S***QKP; N) linkers for *zgc77041* and *dclk2* ZFNs.

Figure 4-18

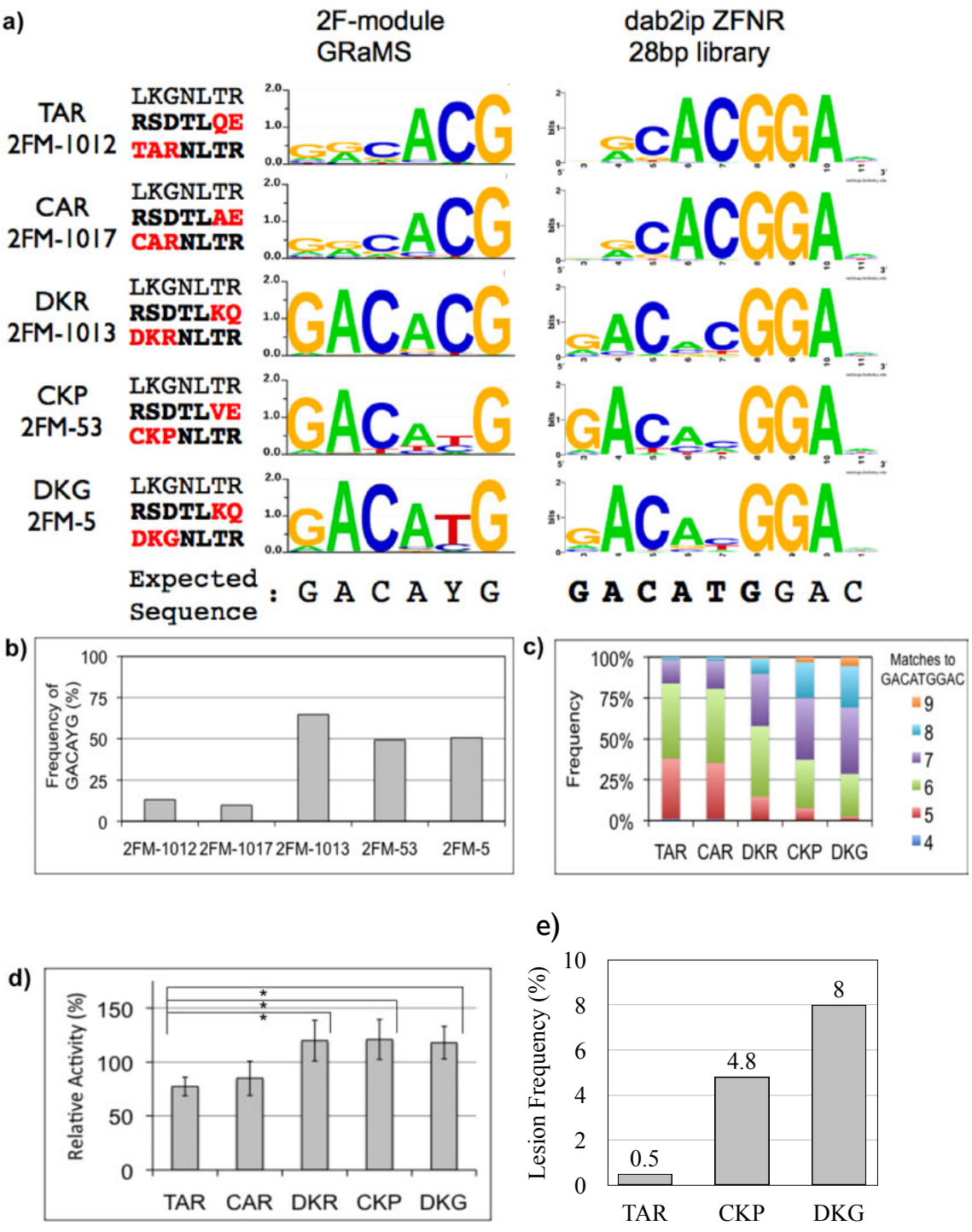

a)
2F-module GRaMS
dab2ip ZFNR 28bp library
TAR 2FM-1012
CAR 2FM-1017
DKR 2FM-1013
CKP 2FM-53
DKG 2FM-5
LKGNLTR
RSDTLQE
TARNLTR
LKGNLTR
RSDTLAE
CARNLTR
LKGNLTR
RSDTLKQ
DKRNLTR
LKGNLTR
RSDTLVE
CKPNLTR
LKGNLTR
RSDTLKQ
DKGNLTR
Expected Sequence : G A C A Y G
G A C A T G G A C
b)
Frequency of GACAYG (%)
2FM-1012 2FM-1017 2FM-1013 2FM-53 2FM-5
c)
Frequency
Matches to GACATGGAC
TAR CAR DKR CKP DKG
d)
Relative Activity (%)
TAR CAR DKR CKP DKG
e)
Lesion Frequency (%)
0.5
4.8
8
TAR CKP DKG

Figure 4-18: Influence of specificity of 2F-module on ZFN specificity and activity. (**a**) For dab2ip ZFN, five ZFNRs were constructed using different 2F-modules (recognition helices in bold and interface residues in red) that show different specificities (GRaMS logos shown on the left). These 2F-modules were attached to the same 'GAC' binding N-terminal finger to create ZFNRs and their specificities (shown on the right) were determined via B1H selections using the 28bp randomized library[90] (**b**) Quantification of specificity of 2F-modules is displayed as the frequency of expected target sequences (GACAYG) within the recovered sequences from the binding site selection. (**c**) Quantification of specificity of ZFNRs for dab2ip ZFNs. Top 1000 sequences obtained via B1H selections were aligned using MEME and then were separated based on their matches to the expected target site. Frequency of sequences in each category is displayed in the graph where 9 indicates a perfect match. The specificity of the ZFNRs correlates with the specificities of the individual incorporated 2F-modules ($R^2 = 0.86$). (**d**) Activity of different dab2ip ZFNs (constant ZFNL and different ZFNRs) was assayed using the yeast based reporter assay. Activity relative to the positive control is displayed as a mean of three experiments. Bars represent standard deviation and * indicates $P < 0.05$ as determined by paired student's *t*-test. ZFN activity correlates with the specificities of ZFNRs ($R^2 =0.95$) as well as individual 2F-modules ($R^2 =0.95$). **(e)** The mRNAs for *dab2ip* TAR, CKP and DKG ZFNRs were co-injected with the ZFNL mRNA and lesion analysis was performed.

Germline Transmission of ZFN induced lesions

Finally, we injected mRNAs for 4 of 11 ZFNs and assayed ZFN-injected adults for germline transmission of mutant alleles. In all cases we identified founder animals from a small number of screened animals that carried a mutant allele for these targets (**Table 4-4**). Many of these mutations that were observed are expected to cause a frameshift in the reading frame and may disrupt the function of the target gene.

Web interface for identification of ZFN sites within query sequences

To facilitate public use of our archive, we have developed a web interface that allows users to search for potential ZFN sites within an input sequence (http://pgfe.umassmed.edu/ZFPmodularsearchV2.html). Our website allows a user to input a single sequence or multiple sequences in FASTA format for the identification of sites that can be targeted with ZFNs constructed from our single finger[53] and two finger archives. This website is completely anonymous; no login is required to use the interface. Users can choose from multiple formats (browser, text file, word document or excel file) for the output from the initial analysis. Potential ZFN sites are ranked based on their overall score (the scoring metric is described in Methods). Additional information is provided regarding the position of the site within each input sequence, the target sequence for each ZFN monomer, the gap separating these sites, whether there is a restriction enzyme (RE) site within this spacer, and the identity of the finger modules that comprise each ZFA monomer. Each ZFA has four potential fingers (F3, F2, F1, & F0), where the fourth finger (F0 in our nomenclature) if absent is indicated by 'XXX'. Modules appearing in UPPERCASE are from the single finger module archive[53], while

Table 4-4

Gene Name	Number of ZFN injected Fish Screened	Number of Founders Identified	Size of insertions or deletions at target site (+/-bp)	ZFN Lesion Frequency in embryos
mc4r	9	2	-5, -5	**12.9**
lrp8	17	8	+5, -3, -7, -8, -10,-12, -21	**7.3**
mc3r	5	2	+4, -11	**3.1**
apoeb	11	3	-4, -37, +9	**2.8**

Table 4-4: Germline transmission of ZFN-induced lesions.

ZFN-injected embryos for four ZFNs were grown till maturity and crossed with wild type zebrafish. The progeny was screened for lesions to identify founders. The mutant alleles were cloned and sequenced to determine the mutation. Types of lesions obtained are shown.

modules appearing in lowercase are from the archive described in this study and will occur in pairs (*e.g.* grn & nyr pairs). Within the initial browser output, more detailed information on each ZFN can be output using a button at the end of each column. Again there is a choice of output formats, where for each ZFN additional pertinent information is provided: the ZFP amino acid and DNA sequences for gene synthesis, modules IDs within our archive for PCR-based construction, recognition helix sequences, and information on RE sites that overlap with the spacer region for genotyping. The DNA sequences that are provided include *Kpn*I and *Bam*HI sites at their termini for cloning into our pCS2 vectors (DD/RR or EL/KK versions) that are available from addgene. ZFAs can be either be assembled using the detailed protocol described in methods and available for download from the home page or can be synthesized which is recommended due to its affordability and ease.

Discussion

In this study we report a unique set of 87 validated 2F-modules that recognize 162 six base-pair target sites with high specificity. These 2F-modules can be readily combined together or with available single finger modules[53] to rapidly create active ZFNs that can even target sequences containing 'non-N-G'-junctions *in vivo*. Our combined archive allows targeting of ~95% of the protein-coding genes (exons, Zv9) in the zebrafish genome, with an average density of one unique ZFN site every ~140 bp, which is a ~5-fold higher density than available through the CoDA archive[100]. We have also developed

a web interface that allows users to search for potential ZFN sites that can be targeted within an input sequence (http://pgfe.umassmed.edu/ZFPmodularsearchV2.html).

Comparison to previously described Finger Archives

A number of different systems have been described for assembling Zinc Finger Arrays (ZFAs) from one-finger (1F) [53,75,82,83,88,161,226-229] or two-finger (2F) [100,230] archives. These archives display diversity in the number of fingers, the base composition of their recognition sequences and the strategies for their assembly. The quality of many of these archives have been assessed on a moderate to large scale through characterization of the constructed ZFAs[99,102,103,202] or assessment of the activity of ZFNs containing these ZFAs in cell lines or *in vivo*[53,100,103,161,230]**.** Since the utility of an archive is dependent mainly on the success rates of the ZFNs derived from them and their target site density in a genome, we compared these factors for these five archives: B1H 1/2FM (this study), B1H 1FM[53], CoDA 2FM[100], Kim 1FM[161] and Kim 1/2FM[230]; **Table 4-5).** The success rates for ZFNs derived from 1F archives are below 30% and those from two-finger archives are generally higher, 50% for CoDA 2FM and 82% for B1H 1/2FM (**Table 4-5**). To compare the targeting densities of these archives, we determined the number of potential target sites in protein-coding exons within the zebrafish (Zv9) and human (GRCh37.p5) genome. We focused our comparisons on the two-finger module archives because they generally have higher success rates (B1H 1/2FM, CoDA 2FM[100] and Kim 1/2FM[230]; (**Table 4-5).** The combination of our 2FM archive with our previously described 1FM archive (Zhu 1FM[53]) expands the targeting density of our original archive by 3-fold,

Table 4-5

	Gupta 1/2FM	**CoDA 2FM**	**Kim 1/2FM**	**Zhu 1FM**	**Kim 1FM**
Archive Reference	A	B	C	D	E
Number of **Unique** ZFN sites in zebrafish protein-coding exons (25090 unique genes Zv9.64)	608,081	110,629	8,645,342	182,698	n.d.
Fraction of zebrafish protein-coding genes containing ZFN site	95.0%	79.2%	98.8%	85.9%	n.d.
Average density of ZFN sites (# bp/site)	132	722	10	438	n.d.
Number of **Unique** ZFN sites in human protein-coding exons (20236 unique genes GRCh37.p5)	1,384,075	269,242	14,669,536	444,163	n.d.
Fraction of human protein-coding genes containing ZFN site	96.7%	92.2%	97.8	94.5%	n.d.
Average density of ZFN sites (# bp/site)	123	633	12	383	n.d.
	Tested ZFNs				
Number of ZFN pairs tested in Archive Reference	11	38	13	29	315 ^
Number "active" ZFNs	9	19	3	8	23
Percent active ZFNs	82%	50%	23%	28%	7%
Percent GNN modules in ZFNs	**64%**	**99%**	**63% ***	**86%**	**40%**
ZFNs sites with non-GNN finger (active)	11 (9)	2 (1)	(3) *	17 (2)	33 (8) ^
ZFN sites with non-N-G junctions (active)	11 (9)	1 (0)	(3) *	10 (1)	33 (8) ^

A = this manuscript
B = Sander, J. D. et al. *Nature methods* **8**, 67-69 (2011)
C = Kim, S. el al. *Nature methods* **8**, 7 (2011)
D = Zhu, C. *et al. Development* **138**, 4555-4564 (2011)
E = Kim, H. J., et al. *Genome Res* **19**, 1279-1288 (2009)
n.d. = not determined
* = only target sequences for successful ZFNs reported
^ = multiple zfn pairs were tested at each target site

Table 4-5: Metrics for comparison of different ZFN assembly systems.

while creating ZFNs with promising activity. This combined archive has a ~5-fold higher density of ZFN sites than the CoDA archive, with an average of one unique ZFN site every ~140 bp. The Kim 1/2FM archive has the highest targeting density of the three archives due to the large number of 2F-modules it contains with an average of one unique ZFN site every 10 bp, albeit with a lower overall success rate.

While the targeting density provides one important reflection on the utility of an archive, its flexibility can be inferred from the composition of target sequences evaluated in studies validating its efficacy. Although, the CoDA archive contains a combination of GNN and non-GNN finger sets (74 non N-G junction 2F-modules), the ZFNs that were evaluated by Sander and colleagues were composed almost entirely of GNN finger sets (99%). This may reflect the fact that only 3 of 10 ZFAs containing non N-G junction 2F-modules were functional in their bacterial activity-assay[100], which was used for prescreening modules employed in their ZFNs. Our ZFNs contain a more diverse set of fingers where roughly two-thirds (64%) were GNN finger sets **(Table 4-5)**. Of the 39 CoDA ZFNs that were evaluated, only one target contained a finger set recognizing a non-N-G junction between fingers, whereas all 11 of our evaluated ZFNs contained non-N-G junction, demonstrating the breadth of sequences that can be effectively targeted using our system **(Table 4-5)**. For the ZFNs evaluated in the Kim 1/2FM archive analysis, only the sequences of the three active ZFNs were reported limiting the comparisons that can be drawn between it and the other archives[230].

Choice of ZFN target sites

The ZFNs evaluated in our study were chosen to serve a number of different goals. Foremost, ZFNs were chosen to assay different numbers of fingers per ZFN and different mixtures of 2F- and 1F-modules, where all of the ZFN pairs contain at least one non-GNN finger and one non-N-G interface. While there is some bias in the composition of the fingers comprising the ZFNs that were evaluated, many of the choices were driven by the desire to inactivate specific target genes in zebrafish that if successful could potentially yield useful disease models (atherosclerosis (*apoeb*, *lrp8*), obesity (*lepr*, *mc3r*, *mc4r*), and diabetes (*irs2*)). Nonetheless, we believe that the 82% success rate achieved in this sample set will not be completely representative of ZFNs constructed from this archive. For example, this archive is a mixture of 2F-modules and 1F-modules, where about 30% of the identified ZFNs are composed of only 1F-modules. Based on our prior evaluation, we would anticipate only about one-fourth of these ZFNs to be active[53]. To aid the user in the choice of ZFNs for specific target genes we have constructed a scoring function that weights the 2F-modules based on their specificity in the B2H system. This has been integrated with our previously described 1F-module scoring function[53].

Effect of linker on ZFN activity and toxicity

In this study we demonstrated that employing a non-canonical TG*S*QKP linker to connect 2F-modules might have beneficial effects on activity and toxicity for some of the ZFNs. This non-canonical linker has been used in a large number of active ZFNs published by Sangamo BioSciences[62,113,114]. Pioneering work from the Klug lab demonstrated that incorporating this non-canonical linker between 2F-units may improve discrimination of

the target site as compared to a mutant site with a different finger-subsite, while maintaining high affinity for the target site[59]. Since we observed reduced toxicity for some of the ZFNs with the non-canonical linker as compared to the canonical linker, we believe that the ZFNs with the non-canonical linker show reduced affinity towards the non-specific binding sites thus decreasing their off-target activity and increasing their availability on the target site resulting in higher activity. A detailed analysis of the linker length and composition will help increase not only the activity of the ZFNs but also their targeting density by allowing skipping bases between otherwise adjacent zinc fingers subsites.

CV-B1H method for determining binding site specificities

As has been previously noted, selections might not always yield highly specific modules[82,83,86]. We employed the CV-B1H method to identify the 2F-modules that show high specificity for the desired binding sites. The CV-B1H method offers several advantages: since the binding site library is in a fixed register, it allows rapid and cheap semi-quantitative assessment of binding specificities of 2F-modules via Sanger sequencing of a single sample. Moreover, binding site pools for multiple (>100) 2F-modules can be indexed and sequenced into a single lane of Illumina sequencing that allows a more quantitative estimate of binding specificities for 2F-modules.

The importance of using the binding site specificity validation was emphasized by the fact that the 2F-modules obtained from the same selection condition differed in their ability to specify the desired interface and sometimes even preferred binding sites

different than the desired ones. Further, incorporating 2F-modules with different specificities in the ZFNR monomer of the *dab2ip* ZFNs influenced their activity and toxicity that correlated with the specificity of the individual 2F-modules **(Figure 4-18)**. This observation might also explain the low success rate for some of the CoDA 2F-modules that were selected to specify a non-N-G interface[100]. When incorporated into 3 finger zinc finger proteins, these 2F-modules showed poor activity in the B2H system where only 3 of 10 tested ZFPs were active above their threshold[100]. Also, when we tested some of their 2F-modules in the B1H assay, they preferred the N-G junction instead of the desired N-A junction suggesting that some of the CoDA 2F-modules that recognize non-N-G junctions might not specify the desired target sites thus highlighting the need for validating the binding site specificities of 2F-modules post selections **(Figure 4-14)**.

DNA Recognition by the residues at the finger-finger interface

The DNA recognition at the 2 bp junction of the zinc finger subsites by the interface residues of a multifinger protein is not well understood. Our zinc finger selections in two fixed contexts (Asn at position 3 of finger-2 or His at position 3 of finger-2) followed by DNA binding specificity analysis have provided some insights into the DNA recognition by the residues at the interface of two zinc fingers. For 10 of 16 '2 bp junctions', selections with both the Asn+3F2 and the His+3F2 libraries yielded similar residues at the interface suggesting similar mechanisms of DNA recognition at the interface by these residues without much interference from the neighboring residue at position 3 of finger-2.

Indeed, the binding site selections showed similar specificities for 9 of 10 of these 2 bp junctions **(Figure 4-6)**. However, for the other 6 of 16 interfaces where different residues were obtained post B1H selections, we expect greater influences from the residue at position 3 of finger-2. The influence of DNA recognition at the 2 bp junction by the residue at position 3 of finger 2 was confirmed from our focused substitution experiments where changing the residue at the position 3 of finger-2 altered the specificity at the 2 bp junction. Interactions between the residue at position 3 and the interface have been observed before; for example in the Zif268 D20A mutant where Asp at position 2 that stabilized the Arg at position -1 in the wild type protein, was mutated to Ala resulting in the rearrangement of the side chains of Glu at position 3 and Arg at position -1 allowing them to form hydrogen bonds in the mutant protein and affecting the specificity at the 3' base in the finger subsite[225]. These interactions from the residue at position 3 to the interface limit the scope of the conclusions that can be drawn from the study by Isalan *et al.* for understanding DNA recognition at the interface where along with the residues at the interface, the residue at position 3 of the C-terminal finger was also randomized[84]. These neighbor effects influence recognition along the entire finger-DNA interface, as the type of residue at the position 2 influenced that specificity of the residue at position 3. For example, the preference of His at position 3 for base G or A was influenced by the amino acid at position 2 where a branched amino acid shifts the preference to A and an aromatic residue shifts it to G **(Figure 4-19b)**.

Moreover, the B1H selections emphasized the importance of non-independent interactions for zinc fingers-DNA recognition. The frequency logo of 87 2F-modules in

Figure 4-19

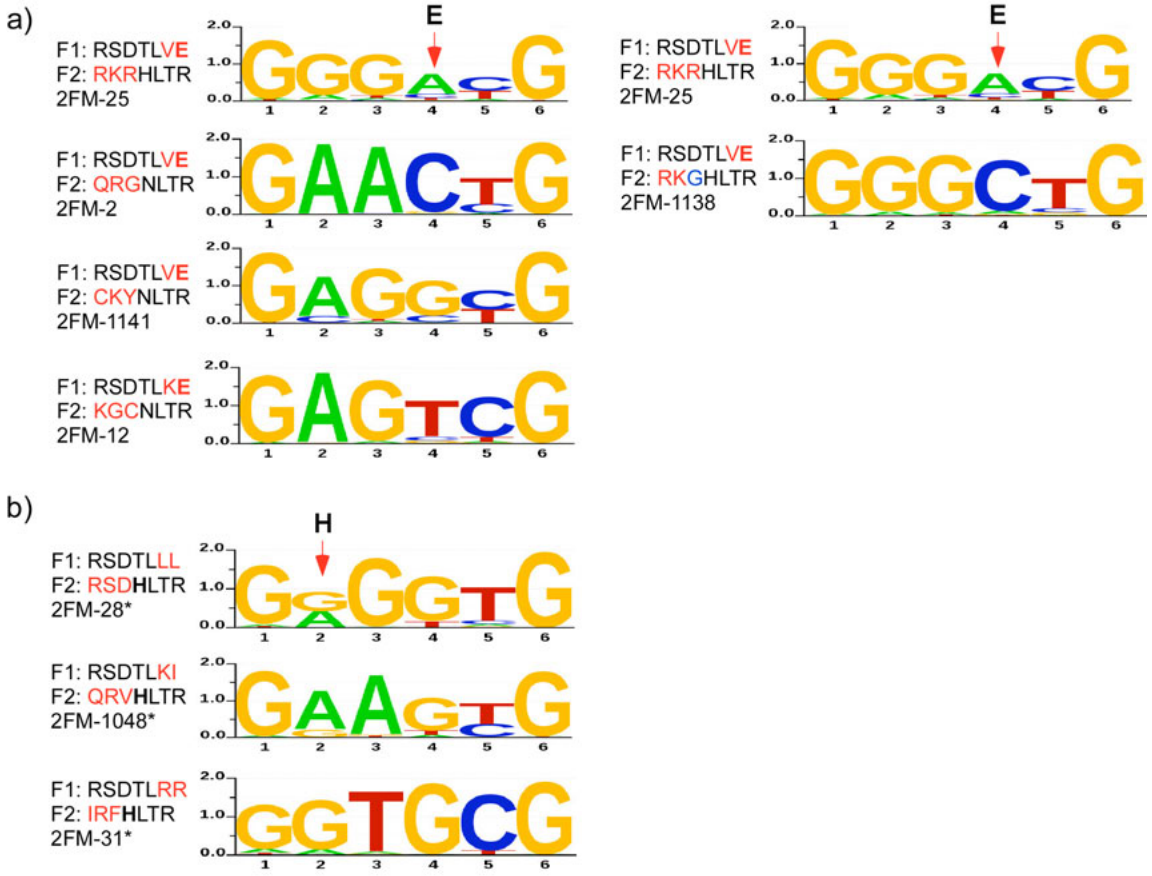


Figure 4-19: Examples of context dependent specificities within zinc finger pairs. (**a**) DNA binding specificities for 4 2F-modules each with Glu at position 6 of F1 are shown. The base preferred at the 4th position opposite the Glu is different in each of the 4 2F-modules revealing context dependent that is likely influenced by the residue at position 2 of F2. In the case of 2FM-25, substitution of R for G (blue residue rightmost logos) at position 3 of F2 results in a change in specificity at base position 4 from A to C. (**b**) The specificity of His at the position 3 of F2 is influenced by the neighboring residue at position 2 where a branched amino acid shifts the base preference of His towards Adenine and an aromatic residue at the position 2 shifts it towards Guanine displaying an influence of residue at position 2 on the specificity of residue at position 3.

our archive shows 3 most frequent residues at each of the base specifying positions -1 of finger 2 and +6 of finger 1: Arg which specifies 'G', Gln that specifies 'A' and Glu or Asp that provide some selectivity for 'C', where the specificities were obtained from the available recognition codes[48,84,126] (**Figure 4-20**). On the surface, it looks as if these residues act independently of residues at other positions to specify each of these bases. However, comparing the binding site specificities of the B1H-selected 2F-modules and their variants from the mutagenesis experiments highlights the importance of other residues especially the residue at position 2 on the specificity at the 2 bp junction. For example, the base preferred opposite Glu at position 6 of finger 1 was not always a Cytosine at the 5' position of the finger 1 DNA triplet. In fact, the identity of the base preferred opposite Glu at position 6 was greatly influenced by the residue at position 2 of the neighboring position; with an Arg at position 2, the base preferred is 'A' and with an aromatic residue at position 2 it is 'G' **(Figure 4-19a)**. Consequently, replacing the Arg at position 2 with Gly shifts the preference from Adenine to Cytosine for some of the 2F-modules but not all, suggesting that residues at other positions might also influence the preference opposite Glu at the position 6 **(Figures 4-19a and 4-21)**. Structures of zinc fingers have provided examples where the amino acid at position 2 of the C-terminal finger can influence the specificity not only at the 5' base of the subsite of the N-terminal finger but also at other positions[45,52,66,225]. These observations suggest that non-independent inter-finger interactions may be rampant in zinc fingers especially in the non-GNN recognizing set and thus the identity of one finger may affect the specificity of the neighboring fingers. Therefore, development of recognition codes with better

Figure 4-20

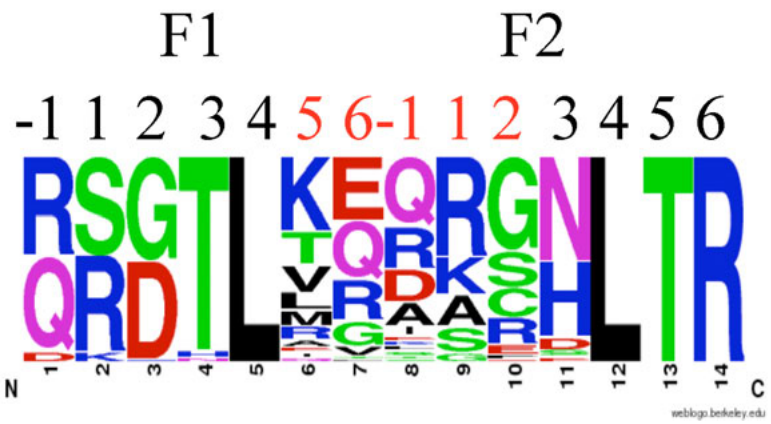


Figure 4-20: Frequency logo for 87 2F-modules. The recognition helix sequences (residues -1, 1, 2, 3, 4, 5 and 6) from the final list of 87 2F-modules were used for constructing the frequency logo. Residues (in red) at positions 5 and 6 of finger-1 and positions -1, 1, and 2 of finger-2 were randomized in the original libraries.

Figure 4-21

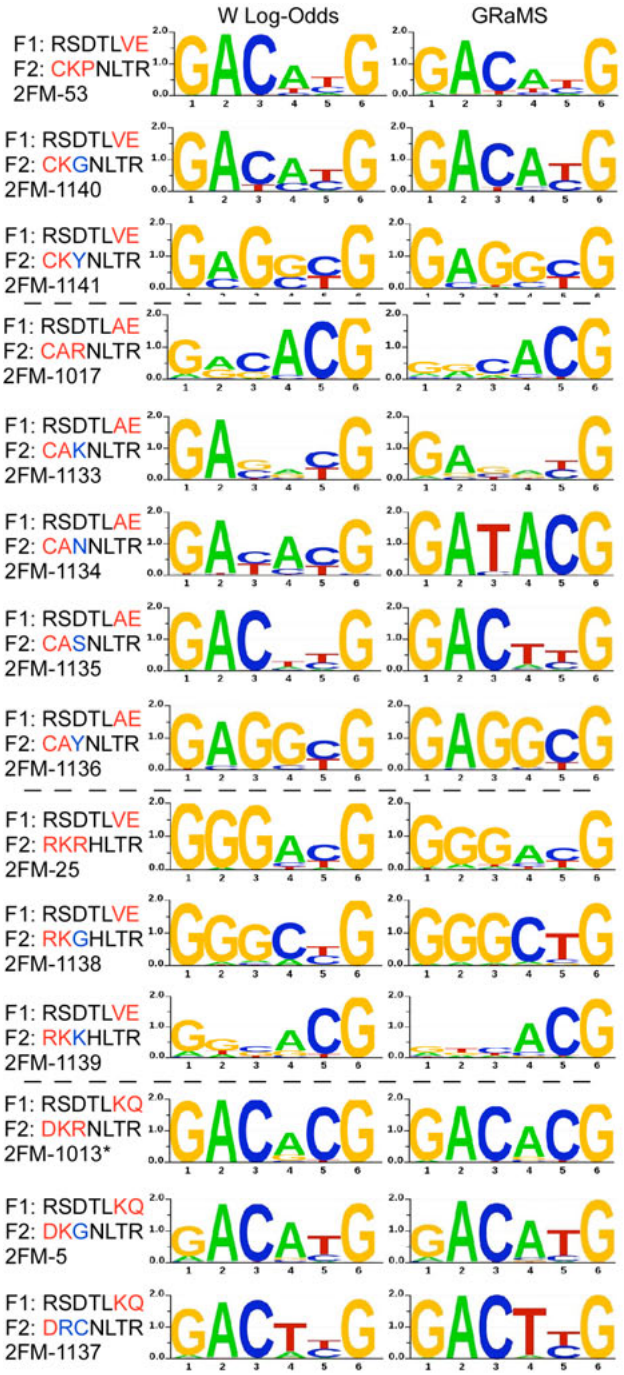


Figure 4-21: Influence of amino acid at position 2 of F2 on the specificity at the 2bp junction. For four 2F modules the amino acid at position 2 of F2 was replaced with an alternate amino acid, where all sequence changes are indicated in blue. The specificities of the original and the mutated 2F modules are displayed.

predictive capabilities would require incorporating non-independent interactions from residues of the same finger and from the residues of the neighboring fingers.

Although our archive represents the largest set of 'non-N-G'-junction recognizing 2F-modules described to date, it comprises only 132 of the possible 3072 non-N-G junction sites. Additional archives of 'non-N-G'-junction 2F-modules exist; for example, sixty-one are found in the CoDA archive, but the specificity and activity of these modules have not been characterized in detail[100]. Thus, there is a need to expand the set of quality 2F-modules covering these junctions to further increase the targeting resolution of ZFNs. However, as observed in our study, selections alone may not always be sufficient to obtain highly specific modules for a given target sequence since the isolation of precise modules can be confounded by the roles that both affinity and specificity play in activity in the selection system that is employed. Consequently, the continued development of more accurate predictive models of DNA recognition for zinc fingers will likely be needed to continue to inform design efforts. Ultimately, these efforts should lead to important advances in nuclease precision and activity not only for engineering model systems, but also for creating therapeutic reagents for the treatment of disease.

Materials and Methods:

Animal husbandry: Zebrafish were handled according to established protocols[215] and in accordance with Institutional Animal Care and Use Committee (IACUC) guidelines of the University of Massachusetts Medical School.

2F-Library construction: 2F-libraries were constructed in two stages. First, individual F1 and F2 libraries were independently constructed via cassette mutagenesis of annealed randomized oligonucleotides into pBluescript vector containing the appropriate zinc finger backbone derived from Zif268. The sequences for the randomized oligonucleotides are given below where lowercase letters denote the randomized bases.:

F1 library top oligo:

CCTGCGACCGCCGCTTCTCCAGATCTGAyACnCTnvnsvnsCATATACGTATTCACAC

F1 3' complement bottom oligo:

GCCGGTGTGAATACGTATATG

F1 5' complement bottom oligo:

AGATCTGGAGAAGCGGCGGTCG

F2 library top oligo (His+**3F2**):

CTGCATGAAGGCCTTCTCTnnwnnwnnwCAyCTnACACGTCACATCAGGACCCACAC

F2 library top oligo (Asn+**3F2**):

CTGCATGAAGGCCTTCTCTnnwnnwnnwAAyCTnACACGTCACATCAGGACCCACAC

F2 3' complement bottom oligo:

GCCGGTGTGGGTCCTGATGTGACGTGT

F1 5' complement bottom oligo:

AGAGAAGGCCTTCAT

Individual finger library diversity greatly exceeded the theoretical library size; ~$1x10^5$ transformed cells were obtained for the F1 library (>100 times theoretical size) and ~$1x10^6$ transformed cells for the F2 library (30 times theoretical size of the library).

Constructed libraries were grown at low density on 2xYT plates containing 100 μg/ml carbenicillin at 37 °C for 14 hours. Individual F1 and F2 libraries in pBluescript were harvested from pooled cells from these surviving colonies.

The 2F-library was constructed from the single finger libraries by PCR assembly, individual F1 and F2 libraries were separately amplified from the pooled pBluescript clones by PCR and then joined via overlapping PCR, where the number of amplification cycles in both steps was minimized by employing high concentrations of template DNA. This 2F-library was then cloned into the B1H expression vector 1352-omega-UV2 between unique BssHII and Acc65I restriction enzyme sites such that the w-subunit of the RNA polymerase is fused at the N-terminus of the zinc fingers and the Engrailed homoeodomain at the C-terminus. Following electroporation into bacterial cells, $1x10^8$ cells (5 times the theoretical size of the library) were plated on 10 2xYT-carbenicillin plates (150 x 15 mm) and grown at 37 °C for 14 hours. 1352-omega-UV2 plasmids containing the 2F-library were isolated from pooled surviving colonies and used for selections.

Zinc finger Binding site cloning: The 16 GANNCG zinc finger binding sites (ggccTAATTACCTGANNCGGacg) were cloned between the EcoRI and NotI sites in the pH3U3-mcs reporter vector. The Homeodomain (Engrailed) binding site TAATTA (underlined) is present 3 bp away and on the strand opposite to the zinc finger binding site to minimze any interference between the Homeodomain and the zinc fingers. For selecting 2F-modules from the Asn+3F2 library that recognize the 'G-G' interface,

sufficient stringency could not be obtained to narrow the selected clones merely through increased 3-aminotriazole (3-AT) concentration or reduced inducer (isopropyl-b-D-thiogalactoside; IPTG) levels. To reduce the activity of the ZFP-HD construct the homeodomain site was mutated to TAAAGG to increase the dependence on zinc finger binding.

2F B1H Selections: Selections for 2F-modules were performed as described previously[94]. The zinc finger library (20 ng) and the reporter vector (1 μg) containing the zinc finger target site were cotransformed via electroporation in the selection strain that lacks endogenous expression of the w-subunit of RNA polymerase (US0Δ*hisB*Δ*pyrF*Δ*rpoZ*). $2x10^7$ cotransformed cells were plated on selective NM minimal medium plates (where stringency was controlled via 3-AT and IPTG concentration) and grown at 37 °C until moderate number of colonies (typically 100s) were visible. Post-selection, 2F-modules from 6-10 surviving colonies were sequenced to identify functional amino acid sequences for further evaluation. The success of the selection was judged on the diversity of sequences obtained from these selections, with the expectation that successful selections will converge on a small number of functional residues at the critical recognition positions.

Cloning B1H-selected 2F modules into 3F F1-GCG constructs: To determine the binding specificities of 2F-modules a 'GCG' binding anchor zinc finger (recognition helix: RSDTLAR) was fused at the N-terminus of the 2F-module via overlapping PCR.

GCG-for: CCATGGTACCTCTAGACCC

GCG-rev: GGGCAAGCATACGGTTTTTCACCGGTATGA

2F-module for: GTGAAAAACCGTATGCTTGCCCTGTCGAGTC

2F-module rev:

TTACTGTGCAGAGGATCCCCTCAGGTGGGTCCTGATGTGACG

Following overlapping PCR, the 3F-ZFA was cloned into 1352-omega-UV2 vector between the Acc65I and BamHI sites for expression as an omega fusion.

CV-B1H method: To determine binding site specificities of 2F-modules, the CV-B1H assay was performed as described before[123]. Post-transformations into the selection strain, $1x10^6$ cells containing the zinc finger plasmid (1352-omega-UV2-ZFP) and the randomized binding site library plasmid (pH3U3) were plated on selective NM minimal medium plates (100 x 15 mm) containing 50 μM IPTG and 1 or 2 mM 3-AT and grown at 37 °C for 22-30 hrs. The surviving colonies were pooled and the binding site plasmid was isolated for identification of the functional DNA sequences. The binding site region was PCR amplified and Sanger sequenced to rapidly obtain binding site profiles for each 2F-module. For quantitative modeling, the binding site pools for multiple 2F-modules were barcoded and sequenced via Illumina sequencing, and then binding specificities were modeled from this data using both W log-odds and $GRaMS_c$ methods (below).

$GRaMS_c$: In the original implementation of GRaMS[123], nonlinear regression was employed to parameterize a model consisting of a PWM and a parameter, which describes the degree of saturation of each binding site due to the free concentration of the

TF. We used the same model for this study, but re-arranged the objective function. Instead of fitting to the observed growth rate of each site, we fit to the observed counts per site. We call this version of the program $GRaMS_c$.

Many of the Zif268 mutants in this study are more specific than wild type the Zif268 for their preferred sequences. We found in practice that for very specific proteins that resulted in only a handful of sites with growth rates significantly larger than the median growth rate, the most accurate recognition model was generated by re-arranging the GRaMS objective function to fit directly to the observed counts per site. Otherwise, when there were very few appreciably enriched sites (few informative data points), there was a tendency to over fit to the noise in the growth rate data. We found it also helped to adjust M, the maximum observed growth rate by a factor of 1.02. This prevented a single site from dominating the motif completely when only very few sites had growth rates appreciably larger than the median growth rate and one site clearly had a much higher growth rate than the other enriched sites. The following equations describe the adjusted model. The observed growth rate (r_i), or enrichment, of each site is given by:

$$r_i = log_2\left(\frac{f_i(t)}{f_i(0)}\right)/t \qquad (1)$$

where t indicates the duration of the selection experiment, i is an index over all 4^6 6mers, $f_i(t)$ is the frequency of site i at time t, and $f_i(0)$ is the initial frequency of site i at time 0. The growth rate of a site, S_i, is a sigmoid function of μ, the chemical potential of the TF, and the Gibbs free energy of the TF binding to the site as well as the maximum and minimum possible growth rates:

$$r_i = \frac{M-D}{1+e^{S_i W-\mu}} + D \qquad (2)$$

where W is the PWM and S_i*W yields the Gibbs free energy of binding to S_i. The variables M and D determine the upper and lower plateaus of the sigmoid curve. M is set to the maximum observed growth rate times a scalar of 1.02, and D is set to the median observed growth rate. The total number of times each site was observed is modeled by the following equation:

$$c_i = N_F f_i(0) 2^{\left(\frac{M-D}{1+e^{S_i W-\mu}}+D\right)t} \qquad (3)$$

where N_F is the total number of sequenced sites. The Levenberg-Marquardt algorithm was used to fit the parameters of the PWM and the μ parameter. Regularization was used to prevent over fitting.

W log-odds: The W log-odds ('W' stands for 'word based' log-odds) method more accurately reflects our knowledge of the initial frequency of each 6mer in the library than a simple log-odds weight matrix. Generally, the following formula is used to compute log-odds PWMs:

$$W_{bj} = -\log\left(\frac{P_{bj}}{P_b}\right) \qquad (4)$$

where W_{bj} is the log-odds matrix, P_{bj} is the probability after selection of observing base b at position j in the binding site, and P_b is the initial probability of observing base b before selection. Because the initial frequency of each 6mer binding site prior to selection was known from deep sequencing of the initial counter selected library, the enrichment of

each site after selection was calculated directly. The enrichment ratio of the i^{th} site is given by the equation

$$\frac{f_i(t)}{f_i(0)} \qquad (5)$$

where t is the final time or duration of the selection experiment, $f_i(t)$ is the final frequency at time t, $f_i(0)$ is the initial frequency at time 0, and i is an index over all 4^6 6mers. A site's enrichment ratio can be thought of as the K_a of that site. A pseudo count of one was added to all final and initial counts when calculating the initial and final frequencies. The sum of all the enrichment ratios for all 6mers containing base b at position j was used to calculate each element of the W log-odds matrix:

$$W_{bj} = -\log\left(\sum_{i=1}^{4096} \delta_{S_{ij},B_b} \frac{f_i(t)}{f_i(0)}\right) \qquad (6)$$

where S_{ij} indicates the base at position j of site i, b is an index over the four nucleotide bases, B_b returns base b and $\delta_{x,y}$ is the Kronecker delta function which returns 1 if the bases x and y are identical and zero otherwise. For example, to determine the energy contribution of an A in the first position of the binding site ($W_{1,1}$) the set of all 1024 6mers that have an A at position 1 was determined and the enrichment ratios for all of these sites were summed and the negative of the log of this value was taken.

Rating of 2F modules: For every 2F-module, the frequency of each of the 16 possible 2bp-junctions was determined in the binding sites that were recovered by Illumina sequencing. The 2F-modules for which the frequency of the desired 2bp-junction was the highest among all 16 2bp-junctions were designated as possessing 'preferential

specificity'. If the frequency of the desired 2bp-junction was the second highest and represented more than 20% of the dinucleotide population, the 2F-module was designated as having 'compatible specificity'. The remaining 2F-modules were designated as having 'poor specificity'.

Comparison of CoDA-2F modules and B1H-selected 2F-modules: The CoDA 2F-modules were created using overlapping PCR where the desired recognition helix sequences were introduced into the Zif268 finger 2 backbone. The 2F-modules were fused to the N-terminal 'GCG' binding finger and CV-B1H assay was performed followed by binding site modeling using the W log-odds and GRaMS methods as described above. B1H-based activity assay were performed as described previously[160].

Creating ZFAs: Three Finger (3F) and Four Finger (4F) ZFAs for use in ZFNs were assembled from the 2F-module archive described herein and a 1F-module archive that we recently described[53] using overlapping PCR. The primer sequences used for these different assemblies are listed in **Table A-3**. If desired these ZFAs can also be synthesized from the DNA sequence output from our website application.

For amplifying individual 1F and 2F modules, the following PCR conditions were used: 10 ng DNA template, 1 μM each of forward and reverse primer, 200 μM dNTPs and 0.5 unit of Phusion High Fidelity DNA polymerase (New England Biolabs) in 25 μl reaction volume. PCR cycles: 98 °C 3 min, [98 °C 15 sec, 50 °C 15 sec, 72 °C 30 sec] 6 repeats, [98 °C 15 sec, 56 °C 15 sec, 72 °C 30 sec] 24 repeats, 72 °C 5 min, 4 °C.

For ZFA assembly from the individual 1F and 2F module amplicons was mediated by overlapping PCR under the following conditions: 1-5ng DNA for each component, 200 μM dNTPs and 0.5 unit of Phusion High Fidelity DNA polymerase (New England Biolabs) in 25 μl reaction volume. PCR cycles: 98 °C 3 min, [98 °C 15 sec, 50 °C 15 sec, 72 °C 30 sec] 6 repeats 72 °C 5 min. Following this initial assembly step the forward and reverse primers (final concentration of 1 μM each) were added to the reaction and PCR amplification proceeded using the following cycles: 98 °C 3 min, [98 °C 15 sec, 56 °C 15 sec, 72 °C 30 sec] 25 repeats, 72 °C 5 min. Post-amplification, the 3F/4F PCR products were digested with Acc65I and BamHI enzymes and cloned into appropriate vectors.

Note: The QRG cap is introduced into the 2F module using a special QRG(X) primer set that substitutes the RSD cap with the QRG cap in F1. When Thr is present at position 3 of F1 use the QRG(T) primer, when Asn is present at position 3 of F1 use the QRG(N) primer, and when His is present at position 3 of F1 use the QRG(H) primer. For ZFNs recognizing a seven base pair gap utilize the F3RnTGPGAAGS or 2FM-F3RnTGPGAAGS instead of the F3RnLRGS or F3RnLRGS primers to incorporate the longer linker associated with increased activity. To incorporate the non-canonical linker (TGSQKP) between F1 and F2 fingers of a 4F construct, the sequences for F2-forward primer and F1-reverse primer were modified to introduce the additional Serine in the linker.

3F-ZFAs assemblies from F1, F2 and F3 1F-modules: The single fingers are amplified individually and then assembled. F1 was amplified using the F1(noF0)Fn and F1Rn

primers. F2 was amplified using the F2Fn and F2Rn primers. F3 was amplified using F3Fn and F3RnLRGS primers. The amplified DNA was gel purified using a Qiagen gel purification kit. For finger assembly, 5ng of the F1, F2 and F3 amplicons were combined and assembled as described above, where the F1(noF0)Fn and F3RnLRGS primers were added to the PCR reaction for the final amplification.

3F-ZFAs assemblies from F1 1F-module and 2F-module: F1 was amplified using the F1(noF0)Fn and F1Rn primers. 2F-module was amplified using 2FM-F2Fn and 2FM-F3RnLRGS primer. The amplified DNA was gel purified and the finger amplicons were assembled as described above, where the F1(noF0)Fn and 2FM-F3RnLRGS primers were added to the PCR reaction for the final amplification.

3F-ZFAs assemblies from a 2F-module and F3 1F-module: 2F-module was amplified using the 2FM-F1(noF0)Fn and 2FM-F2Rn primers and F3 was amplified using the F3Fn and F3RnLRGS primers. The amplified DNA was gel purified and the finger amplicons were assembled as described above, where the 2FM-F1(noF0)Fn and F3RnLRGS primers were added to the PCR reaction for the final amplification.

3F-ZFAs assemblies from a F1 1F-modules and 2F-module-QRG cap: F1 was amplified using the F1(noF0)Fn and F1Rn primers. 2F-module was first amplified with 2FM-F1Fn and 2FM-F2Rn primers, gel purified and then 1-5ng of gel purified DNA was used as template for amplification with 2FM-F2-QRG(X)Fn and 2FM-F3RnLRGS primers to substitute the RSD N-terminal cap with the QRG-N-terminal cap. The amplified 2F-module-QRG and F1 module were gel purified and the finger amplicons

were assembled as described above, where the F1(noF0)Fn and 2FM-F3RnLRGS primers were added to the PCR reaction for the final amplification.

3F-ZFAs assemblies from a 2F-module-QRG cap and F3 1F-modules: 2F-module was first amplified with 2FM-F1Fn and 2FM-F2Rn primers, gel purified and then 1-5ng of gel purified DNA was used as template for amplification with 2FM-F1(noF0)-QRG(X)Fn and 2FM-F2Rn primers. F3 was amplified using the F3(noF0)Fn and F3RnLRGS primers. The amplified 2F-module-QRG and F3 module were gel purified and the finger amplicons were assembled as described above, where the 2FM-F1(noF0)Fn and F3RnLRGS primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from F0, F1, F2 and F3 1F-modules: F0 was amplified using the F0Fn and F0Rn primers. F1 was amplified using the F1Fn and F1Rn primers. F2 was amplified using the F2Fn and F2Rn primers. F3 was amplified using F3Fn and F3RnLRGS primers. The amplified DNA was gel purified and the finger amplicons were assembled as described above, where the sF0Fn and F3RnLRGS primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from F0 and F1 1F-modules, and 2F-module: F0 was amplified using the F0Fn and F0Rn primers. F1 was amplified using the F1Fn and F1Rn primers. The 2F-module was amplified using 2FM-F2Fn and 2FM-F3RnLRGS primer. The amplified DNA was gel purified and the finger amplicons were assembled as described

above, where the F0Fn and 2FM-F3RnLRGS primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from F0, 2F-module and F3: F0 was amplified using the F0Fn and F0Rn primers. 2F-module was amplified using the 2FM-F1Fn and 2FM-F2Rn primers and F3 was amplified using the F3Fn and F3Rn primers. The amplified DNA was gel purified and the finger amplicons were assembled as described above, where the F0Fn and F3RnLRGS primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from 2F-module, F2 and F3: The 2F-module was amplified using the 2FM-F0Fn and 2FM-F1Rn primers. F2 was amplified using the F2Fn and F2Rn primers. F3 was amplified using the F3Fn and F3Rn primers. The amplified DNA was gel purified and the finger amplicons were assembled as described above, where the 2FM-F0Fn and F3RnLRGS primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from N-terminal-2F-module, C-terminal-2F-module: The N-terminal 2F-module was amplified with the 2FM-NT-in-Fn and 2FM-F1Rn primers and the C-terminal 2F-module was amplified with the 2FM-F2Fn and 2FM-F3RnLRGS primers. The amplified products were gel purified and the finger amplicons were assembled as described above, where the 2FM-NT-out-Fn and 2FM-CT-out-Rn primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from F0, F1 and 2F-module-QRG: F0 was amplified using the F0Fn and F0Rn primers. F1 was amplified using the F1Fn and F1Rn primers. 2F-module

was first amplified with 2FM-F1Fn and 2FM-F2Rn primers, gel purified and then 1-5ng of gel purified DNA was used as template for amplification with 2FM-F2-QRG(X)Fn and 2FM-F3RnLRGS primers to substitute the RSD N-terminal cap with the QRG-N-terminal cap. The amplified 2F-module-QRG, F0 and F1 modules were gel purified and the finger amplicons were assembled as described above, where the F0Fn and 2FM-F3RnLRGS primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from F0, 2F-module-QRG and F3: F0 was amplified using the F0Fn and F0Rn primers. 2F-module was first amplified with 2FM-F1Fn and 2FM-F2Rn primers, gel purified and then 1-5ng of gel purified DNA was used as template for amplification with 2FM-F1-QRG(X)Fn and 2FM-F2Rn primers. F3 was amplified using the F3Fn and F3Rn primers. The amplified the F0, 2F-module-QRG and F3 modules were gel purified and amplicons were assembled as described above, where the F0Fn and F3RnLRGS primers were added to the PCR reactions for the final amplification

4F-ZFAs assemblies from 2F-module-QRG, F2 and F3: The 2F-module was first amplified with 2FM-F1Fn and 2FM-F2Rn primers, gel purified and then 1-5ng of gel purified DNA was used as template for amplification with 2FM-F0-QRG(X)Fn and 2FM-F1Rn primers. F2 was amplified using the F2Fn and F2Rn primers. F3 was amplified using the F3Fn and F3Rn primers. The amplified DNA was gel purified and the finger amplicons were assembled as described above, where the 2FM-F0Fn and F3RnLRGS primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from N-terminal-2F-module-QRG, C-terminal-2F-module: The N-terminal 2F-module was amplified with 2FM-F1Fn and 2FM-F2Rn, gel purified and then 1-5ng of gel purified DNA was used as template for amplification with 2FM-F0-QRG(X)Fn and 2FM-F1Rn primers. This amplified 2F-module-QRG was again gel purified and PCR amplified with 2FM-NT-in-Fn and 2FM-F1Rn primers. The C-terminal 2F-module was amplified with 2FM-F2Fn and 2FM-F3RnLRGS. The amplified N-terminal and C-terminal 2F-modules were gel purified and the finger amplicons were assembled as described above, where the 2FM-NT-out-Fn and 2FM-CT-out-Rn primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from N-terminal-2F-module, C-terminal-2F-module-QRG: The N-terminal 2F-module was amplified with 2FM-NT-in-Fn and 2FM-F1Rn. The C-terminal 2F-module was amplified with 2FM-F1Fn and 2FM-F2Rn, gel purified and then 1-5ng of gel purified DNA was used as template for amplification with 2FM-F2-QRG(X)Fn and 2FM-F3RnLRGS primers. The amplified N-terminal and C-terminal 2F-modules were gel purified and the finger amplicons were assembled as described above, where the 2FM-NT-out-Fn and 2FM-CT-out-Rn primers were added to the PCR reaction for the final amplification.

4F-ZFAs assemblies from N-terminal-2F-module-QRG, C-terminal-2F-module-QRG: The N-terminal 2F-module was amplified with 2FM-F1Fn and 2FM-F2Rn, gel purified and then 1-5ng of gel purified DNA was used as template for amplification with 2FM-F0-QRG(X)Fn and 2FM-F1Rn primers. This amplified 2F-module-QRG was again

gel purified and PCR amplified with 2FM-NT-in-Fn and 2FM-F1Rn primers. The C-terminal 2F-module was amplified with 2FM-F1Fn and 2FM-F2Rn, gel purified and then 1-5ng of gel purified DNA was used as template for amplification with 2FM-F2-QRG(X)Fn and 2FM-F3RnLRGS primers. The amplified N-terminal and C-terminal 2F-modules were gel purified and the finger amplicons were assembled as described above, where the 2FM-NT-out-Fn and 2FM-CT-out-Rn primers were added to the PCR reaction for the final amplification.

ZFN website scoring function: Our new ZFN site identification tool (http://pgfe.umassmed.edu/ZFPmodularsearchV2.html) uses 2F-modules from this study and 1F-modules from our previous archive[53] to define favorable combinations of these modules for constructing active ZFNs. These ZFNs are designed to target sequences with 5, 6 or 7 bp gaps between the monomer recognition sequences, where each ZFN monomer can contain three or four fingers. ZFNs with higher scores are more likely to be active, where the current 2F-modules are scored based on their DNA-binding specificity (as determined in the B1H system) where good, fair and poor represent 4, 3 and 2 points respectively. If the modules utilize an A-cap (QRG at the N-terminus of the 2F-module) instead of the standard RSD sequence for G-recognition one point is subtracted from the score. The 1F-modules are scored as previously described[53]. ZFNs containing 2F-modules are readily identified in the output from the website by the presence of lowercase triplet sequences in the site breakdown, and by the presence of "2FM-#" in the output Module ID information.

B1H-binding site selections using the 28bp library: The selections for 3F and 4F ZFAs were performed as previously described[160]. 1-5x10^7 selection strain cells transformed with the 1352-omega-UV2 ZFA expression plasmid and the 28 bp pH3U3 library plasmid were plated on NM minimal medium selective plates lacking uracil and containing 3-AT (2.5, 5 or 10 mM) as the competitor and grown at 37 °C for 36-72 hours. The number of surviving bacterial colonies on each plate was estimated and then these colonies were pooled and the population of recovered DNA sequences was determined via Illumina sequencing. Unique sequences were ranked based on the number of recovered reads. From this list an overrepresented sequence motif was determined with MEME[216] using as input the number of unique sequences from the top of the list that correspond to the estimated number of colonies on the selection plate (typically >1000). The aligned sequences were then used to generate Sequence logos using Weblogo[208].

Yeast-based ZFN activity assay: To assess the activity of our ZFNs in an independent system we employed a Mel1-based yeast activity assay[114]. The target sites for test ZFNs and the positive control ZFN were cloned in the modified ySSA vector and then integrated into the yeast genome (BY4741 strain) at the HO locus. The ZFAs were cloned at the N-terminus of the wild type *FokI* nuclease domain in the pYHis3 and pYLeu2 vectors in between the *Acc651* and *BamH1* sites. Test ZFNs, positive control ZFN and negative control (EGFP containing pYHis3 and pYLeu2 vectors) were transformed in the yeast strain containing the ZFN target site. ZFN expression was induced by 2% galactose treatment for 30 min. The activity assay was performed ~16 hours post-induction as previously described[231]. In brief, yeast cultures were diluted to an OD_{600} of 0.4-0.6. 950

μl of diluted cells were centrifuged and the pellet was resuspended in 200 μl of 20 mM HEPES (pH 7.5)–10 mM dithiothreitol–0.002% sodium dodecyl sulfate. 10 μl of chloroform was added to cells and vortexed for 10 s. After a 5 min pre-equilibration at 30 °C, 800 μl of a 7 mM solution of PNPG (4-Nitrophenyl α-D-galacto-pyranoside; Sigma) in 61 mM citric acid–77 mM Na_2HPO_4 (pH 4) was added and incubated at 30 °C for 30 min. Following incubation, 100 μl aliquot was added to 900 μl of 0.1 M Na_2CO_3 to stop the reaction and the OD_{405} was recorded. a-galactosidase units were calculated as follows: a-gal = $(OD_{405}*1000)/(OD_{600}*t_{pnpg})$ where t_{pnpg} is the time of incubation with PNPG. Relative activity for test ZFNs was calculated as follows: $(100*\text{a-gal}_{\text{test ZFN}})/\text{a-gal}_{\text{positive control}}$.

ZFN injections and lesion analysis: For gene targeting in zebrafish, ZFAs were cloned in pCS2 vectors containing the DD/RR obligate heterodimer version of the *FokI* nuclease domain[153,154]. pCS2-ZFN constructs were linearized with NotI and mRNA was transcribed using the mMesagemMachine SP6 kit from Ambion. ZFN mRNAs were injected into the blastomere of one-cell-stage zebrafish embryos as previously described[94]. ZFN-injected embryos with normal appearance (8-30) and uninjected embryos were collected 24 hpf and incubated in 50 mM NaOH (15 μl/embryo) for 15 min at 95 C to isolate genomic DNA and then neutralized with 0.5 M Tris-HCl (4 μl/embryo). The DNA solution was centrifuged for 1 min at 13,000 rpm and supernatant was taken for lesion analysis. For initial validation of ZFN activity, the region flanking the ZFN target site was amplified using the Phire Hot Start DNA Polymerase

(Finnzymes) and RFLP analysis or *Cel I* nuclease assay (Transgenomics) was performed as described previously[94]. For Illumina sequencing, the region flanking the ZFN sites was amplified using the primers listed in **Table A-4** and then digested with the appropriate restriction enzyme (listed in **Table A-4**). The ends for the digested DNA were polished using Klenow exo$^-$ enzyme (New England Biolabs) or T4 DNA polymerase (New England Biolabs) and A-tailed using Klenow exo$^-$ enzyme (New England Biolabs). The barcoded adapters (**Table A-4**) were ligated to each DNA pool and then PCR amplified with the Illumina genomic primers 1.1 and 1.2. Following sequencing, identification of InDels was performed as described previously[160]. Briefly, two tags unique to a ZFN target site were employed, a 5' tag and a 3' tag **(Table A-4)** and the distance between the tags was used to distinguish wild type sequence from the InDel containing sequence. Lesion frequency was calculated as follows: Lesion frequency = $(100*N_{InDels})/N_{total}$ where, N_{InDels} represents number of sequences containing InDels that are >1bp in length and N_{total} represents number of total sequences.

Genomic analysis of ZFN target sites: The targeting density and overlap of ZFN sites were determined for three archives (Gupta 1/2FM, CoDA 2FM[100] & Kim 1/2FM[230]) on the unique protein-coding exons zebrafish (Zv9) and human (GRCh37.p5) Ensembl genes 64. Target sites for each finger archive were determined using custom perl scripts, where only ZFN sites that map to a single unique gene were counted in this analysis. This analysis provides information on the fraction of genes that can be targeted and the density of the sites per base pair.

Germline Transmission Analysis: ZFNs were injected at optimal doses in wild type zebrafish embryos. Injected embryos were grown to maturity and crossed with wild type zebrafish to identify carriers. PCR products spanning the target loci in F1 embryos were screened using Cel1 surveyor nuclease assay for presence of lesions[94]. The compositions of these lesions were characterized through cloning and sequencing PCR products spanning the ZFN target site for each gene (**Table A-4**).

CHAPTER V

DISCUSSION

The work in this chapter has not been published and is still ongoing. Our collaborators, the Stormo lab at Washington University, are developing the zinc-finger predictive model. The Joung lab at Massachusetts General Hospital provided the B2H-selected 2F-modules. Heather Bell, an undergraduate student at Worcester Polytechnic Institute (WPI), performed the CV-B1H-selections for B2H-selected 2F-modules. Cong Zhu performed the selection and characterization of ANNA-2F-modules. Victoria Hall is providing support with developing the zebrafish disease models.

Creating predictive models for zinc fingers

The Cys2His2 zinc finger is the most frequently utilized DNA binding domain (DBD) family in metazoan transcription factors (TFs)[14]. Since their discovery in 1985 [29], there has been a continuous effort to understand DNA recognition by zinc fingers, spurred on in part by the X-ray crystal structure of Zif268, which revealed a rather simple pattern of DNA recognition where a few amino acids on the recognition helix recognize a 3 bp DNA core element[45]. Supported by the structural studies, finger swapping experiments revealed some degree of modularity for zinc fingers[71]. Together, these results triggered a series of novel phage display experiments utilizing randomized zinc finger libraries coupled with SELEX-based DNA specificity determination that not only yielded zinc fingers with novel specificities but also improved the understanding of their underlying principles of DNA recognition[48,58,78,79,82,83,101,108,162,232]. These principles of zinc finger recognition, 'Recognition Codes', provided a Rosetta stone that enabled prediction of the DNA binding specificity for a zinc finger based on its sequence. This code could be used to predict binding sites for naturally occurring zinc finger proteins or design of artificial zinc finger proteins with novel DNA-binding specificity. Unfortunately, the recognition codes based on this experimental data have only limited predictive capacity for multiple reasons[126,233,234]. Structures of both naturally occurring ZFPs and synthetic ZFPs demonstrated existence of a more complex pattern of DNA recognition where the specificity of one finger is influenced by the neighboring fingers through inter-finger interactions (concept termed as context-dependent recognition)[52,66,69]. Consequently, recognition models incorporating non-additivity of individual contacts perform better

than those assuming independence of contacts[233,234]. Moreover, both phage display and SELEX involve multiple rounds of selections and therefore sometimes yield only the highest affinity interaction partners with little or no data on the lower affinity interactions that is critical for obtaining accurate binding energy profiles. Finally the available dataset is small and highly biased towards zinc finger that recognize 'GNN' triplets, which limits its utility for other recognition elements.

The development of more accurate recognition models requires a large, unbiased DNA-protein interaction dataset to allow incorporation of non-independent interactions. One of the ways to obtain such a dataset is by selecting zinc fingers from randomized libraries that bind a diverse set of DNA sequences. However, to select zinc fingers while taking into account the context dependence, one will have to randomize specificity determinants of multiple fingers (at least two) that would require building and searching a large library of zinc fingers ($> 10^{12}$ members), which is not feasible by currently available *in vivo* methods. In the orthogonal approach, a large dataset of DNA-protein interactions can be obtained by defining DNA binding profiles of zinc fingers that have a wide variety of specificity determinants. The CV-B1H method that we have developed provides a medium-throughout system to rapidly characterize DNA binding specificities of zinc fingers and when combined with Illumina sequencing, it provides information on high- as well as low-affinity binding sites for a given zinc finger. Also, the CV-B1H method involves a single round of selection, which is performed in the presence of competitor DNA in the form of the bacterial genome thereby giving B1H an edge over the SELEX, and phage display based methods. In our study, we have also used the B1H system to

select two-finger zinc finger units (2F-modules) that bind a wide variety of desired target sites many of which contain non-GNN finger subsites. Using the CV-B1H selections followed by the GRaMS analysis, we characterized the binding site specificities of ~200 2F-modules. These 2F-modules in general recognize GRN-NYG sequences and show a high diversity of amino acid residues at the interface recognition positions -1, 1 and 2 of finger-2 and position 5 and 6 of finger-1. However, there is only limited diversity at the other positions. Additionally, using the CV-B1H method and GRaMS analysis, we have characterized the DNA binding specificities of ~100 2F-modules that were selected in the bacterial-two-hybrid (B2H) system using the OPEN (Oligomerized Pool Engineering) method **(Table A-6 and Figure A-1)**[99]. These modules were selected to bind GNN-GNG sequences. The combined dataset of the DNA binding specificities on the B1H- and B2H-selected 2F-modules is by far the largest dataset of DNA-protein interactions for zinc finger proteins. To develop predictive models for zinc fingers, in collaboration with the Stormo lab we will use the 'Random Forest' method, which is an ensemble classification method that requires minimum information regarding the parameters it needs to calculate and is easier to implement when compared to other predictive methods[235]. Stormo and colleagues recently employed the RF method to develop predictive models for homeodomains and demonstrated that RF is superior to the k-nearest neighbor based methods (Christensen *et al.,* unpublished data). We recently developed a preliminary RF-based predictive model for zinc fingers that uses the DNA recognition residues (-1, 2, 3, and 6) of each finger of the 2F-module as input and predicts their DNA binding specificities. When used to predict specificities of a subset of

our 2F-modules that were not used to train the model, it performed well where, for the majority of the 2F-modules the predictions had a low mean squared error (MSE < 0.01). However, almost all 2F-modules used to build this model recognize GNN-NNG sites resulting in lower accuracy of the model at the edges of the predicted binding specificities when the corresponding specificity determinants are not 'G'-recognizing. Therefore, further modifications to the model will be needed to improve its predictive capabilities. Nevertheless, in its current form, the model can be utilized to predict specificities of a subset of CoDA 2F-modules that were selected to bind GNN-NNG sequences[100]. Once completely developed, we can use this model to predict specificities of 2F-modules obtained from other methods such as the B2H-selected CoDA modules[100] or phage display selected 2F-modules employed by Sangamo BioSciences[85]. We will also be able to predict the specificities of multi-finger artificial zinc finger proteins and design fingers with novel specificities. Ultimately, we would predict specificities of naturally occurring zinc finger proteins.

ZFNs: increasing their activity, precision and targeting density

Owing to their semi-modular nature and ability to recognize a wide range of DNA sequences, the Cys2His2 zinc finger has been the DNA binding domain of choice for creating artificial DNA binding domains. Numerous studies have demonstrated their use in creating artificial transcription factors, recombinases, histone modifying enzymes, and nucleases (ZFNs)[58,127,138,162,171]. Synthetic ZFNs have been the most popular application

of artificial zinc finger proteins mainly because they provide a universal tool for targeted gene editing that further offers a wide range of applications for reverse genetics as well as gene therapy. Improving the activity and precision of ZFNs has value in all applications, but is especially critical for their application as therapeutics.

The main determinants of activity and specificity for ZFNs are the incorporated zinc finger proteins. In our study, we demonstrated that even modest improvements in the specificity of ZFPs can increase the precision of the ZFNs[160]. However, current modular assembly methods often result in sub-optimal specificity of the assembled ZFPs, which presumably is due to the context-dependent recognition of zinc fingers that arise from the poorly understood interactions at the finger-finger interface. Using B1H-based selections we have identified two-finger zinc finger modules (2F-modules) that have optimized groups of residues at the interface of neighboring fingers and can recognize 162 unique 6 bp DNA sequences (Chapter IV). These 2F-modules can be combined with themselves or with published 1F-modules such that the context of their interface residues is preserved and the unfavorable context-dependent effects are avoided, therefore rapidly yielding highly specific artificial ZFPs. Moreover, when incorporated into ZFNs, these assembled ZFPs demonstrated high success rate of activity in zebrafish. Our archive of 2F-modules, in combination with the published single finger modules, can target ~95% of the zebrafish genes, with a density of 1 ZFN site every 142 bp. However, the current archive recognizes only 162 of 4096 possible 6 bp sequences and we believe that expanding the archive to recognize a wider range of sequences is necessary to increase the activity and targeting density of ZFNs. Our initial efforts to expand the archive

through rational design failed in most instances demonstrating that the residues neighboring the interface that were fixed in our zinc finger library can influence the specificity at the 2 bp junction of the zinc finger subsites (Chapter IV). Therefore, to effectively expand the 2F-module archive, new 2F-modules will have to be selected from either expanded zinc finger libraries where the residues neighboring the interface are also randomized or from libraries where the interface residues are randomized in different contexts. We have employed the latter approach where we randomized the interface residues in the context of Asparagines at position 3 of both finger-1 and finger-2 (instead of Thr at position 3 of finger-1 and Asn at position 3 of finger-2 in the original library) and performed B1H-selections for all 16 GAN-NAG sites (instead of GAN-N*C*G sites). The specificities of selected 2F-modules (ANNA-2F-modules) were determined using CV-B1H and validated, high specificity modules were included in the archive. In addition to conventional modular assembly where these new modules can be combined with other 2F-modules and single finger modules[53] as described in Chapter IV, we are exploring new strategies of combining 2F-modules such that the interface residues and their context can be preserved **(Figure 5-1)**. Since these ANNA-2F-modules recognize G*A*NN*A*G sequences, where the underlined Adenines are recognized by Asparagines at position 3 of finger-2 and finger-1, we hypothesized that we can consider interfaces as modules and create chains of interfaces keeping the neighboring Asn common **(Figure 5-1).** Initial attempts to assemble ZFPs with both the conventional modular assembly method and this new interface stitching method have shown promising results demonstrating that stitching is a viable method for assembling ZFPs and ZFNs.

Figure 5-1

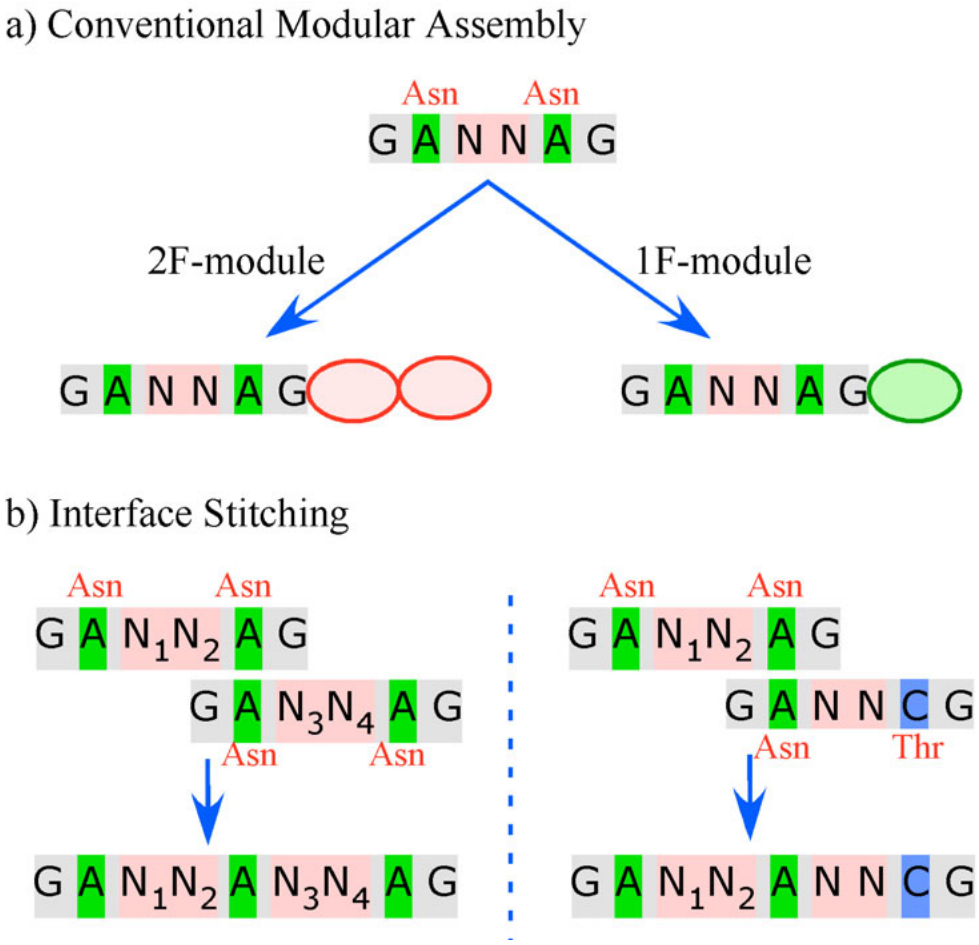


Figure 5-1: ZFN-assembly approaches for the ANNA-2F-modules. The ANNA-2F-modules are depicted by their 6 bp binding sites GANNAG, where the NN represents the di-nucleotide junction specified by the interface residues and are flanked by adenines that are recognized by the asparagines at positions 3 of fingers-1 and -2. In the modular assembly approach (a) the ANNA-2F-modules can be combined either with ANNA-2F-modules (shown as red ovals), GANNCG-binding 2F-modules (or ANNC-2F-modules) described in chapter IV (shown as red ovals) or with published single finger modules (represented by a green oval)[53]. In the interface stitching approach (b) the interface residues from the ANNA-2F-modules or from the ANNC-2F-modules are stitched using the common Asn at position 3 such that the interface residues specifying the NN di-nucleotide junction and their contexts are preserved. Although 3-finger proteins are shown in the figure, longer proteins can be constructed using multiple 2F-modules.

However, with the current archive 2F-modules, only a few sites in the zebrafish genome can be targeted and further expansion of 2F-modules will be required to increase the targeting density of ZFNs assembled using the stitching method.

Another approach to increase the activity and precision of ZFNs would involve identifying optimal linkers between two 2F-modules. Moore *et al.* demonstrated that linking 2F-modules with a non-canonical linker where an extra amino acid residue is inserted in the canonical linker (TGQKP) may allow the assembled ZFPs to discriminate better the cognate and the non-cognate binding sites when larger arrays of fingers are employed[59]. In agreement with their results, incorporating the non-canonical linker (TG*S*QKP) influenced the activity and toxicity of the ZFNs that were assembled from our two 2F-modules. Therefore, selecting for alternate linkers connecting 2F-modules might yield an improved linker that allows the 2F-modules to discriminate the non-cognate sites better while maintaining high affinity to their cognate site.

Further, we can use the selections to identify longer, more rigid linkers that allow skipping one or more bases between the two 6 bp sites for individual 2F-modules. This will not only increase the number of potential ZFN target sites in the genome but may also minimize interface interactions between the two 2F-modules. Although, longer linkers that skip one or two nucleotides between two 2F-module binding sites have been described and incorporated in ZFNs[59-65], they are generally flexible linkers and therefore allow multiple spacings between the 2F-module sites that may result in greater off-target cleavage.

Another potential way to increase the targeting range of the available 2F-modules is by fusing them with a dimerization domain such as a coiled coil domain that would allow cooperative binding of 2F-modules as well as skipping bases between the two half sites[236,237].

In summary, there is both room and need for the improvement in the zinc finger design to enhance the activity and precision of ZFNs.

Gene Targeting in zebrafish using ZFNs: new strategies

Zebrafish is as a model system for studying embryonic development. Until recently, targeted gene manipulation in zebrafish was limited to ZFN-mediated gene disruption. ZFNs have been employed in zebrafish to induce double strand breaks (DSB) and introduce non-homologous end joining (NHEJ) mediated insertions and deletions often resulting in targeted gene disruption[53,94,172,187,189,96,100,114,188]. However, targeted insertion and deletion of desired sequences via homologous recombination (HR) has been elusive in zebrafish. Attempts for targeted knock-in of desired sequences using conventional HR following ZFN treatment did not yield favorable results (McNulty *et al.*, unpublished results). Based on a recent publication where single stranded DNA oligonucleotides (ssODNs) were employed in human cell lines to insert and delete desired DNA elements, we designed ssODNs to introduce a new restriction enzyme site at the site of a ZFN-induced DSB. Although, low activity insertions could be achieved at very high doses of ssODNs, high toxicity was observed in embryos. This toxicity can be somewhat ameliorated by knocking down p53-mediated apoptosis with morpholinos. Further

standardization of conditions will be required to increase the HR rates and decrease the toxicity of ssODNs. This low level of HR-mediated insertions raises questions over the activity and presence of the HR machinery in early zebrafish embryos that opens the possibility of triggering HR rates by exogenously adding a few components of HR or repressing the NHEJ pathway.

Comparison of ZFNs with TALENs

Transcription activator-like (TAL) DNA binding domains are found in naturally occurring virulence factors, TAL effectors, encoded by the *Xanthomonas* plant pathogens [194,195,238]. These virulence factors, containing long repeats of the TAL domains, bind to promoters of the target genes in the host organism and activate their expression affecting the disease process. The TAL effectors naturally comprise multiple repeats of 34 amino acid long TAL domains arranged in tandem where each repeat binds one base-pair of DNA[194,195]. The amino acids at positions 12 and 13 (known as repeat variable di-residues or RVDs) dictate the identity of the recognized nucleotide where the RVDs specifying each of the four bases have been identified[194,195,239,240]. Creating artificial nucleases (TAL effector nucleases or TALENs) by fusing the TAL repeats to the cleavage domain of the *FokI* nuclease has met with promising success[173,191,192,241-244]. A recent large-scale study demonstrated activity for 84 of 96 TALENs constructed to target 96 endogenous genes in the human cell line[244]. The ease of assembly of TALENs, their high success rate and their ability to target almost any DNA sequence give them an edge over the ZFNs for gene targeting. However, the detailed DNA binding specificities and thermodynamics of

DNA binding for TALENs have not been characterized. Further, except for it being less toxic than ZFNs, there is no information available about its off-target activity. Moreover unlike ZFNs, TALENs are much larger in size and contain highly repetitive modules that may pose a concern of viral packaging and delivery for gene therapy based applications.

Creating disease models in zebrafish

Zebrafish has served as an excellent model system for studying development. In the past 15 years, zebrafish has also been utilized for modeling human diseases mainly through reverse genetics, transgenesis, chemical treatments, and nutritional control[245]. The emergence of ZFNs has opened possibilities for creating disease models through targeted gene disruption in zebrafish. Using our archive of 2F-modules described in Chapter IV in combination with the published single finger modules[53] we intend to create zebrafish models of type 2 diabetes and obesity and atherosclerosis.

Metabolism is one of the most fundamental systems for organism survival, dysfunction of which can result in an array of diseases collectively termed as metabolic disorders. In 2008 more than 30% of the population in US suffered from atleast one of the metabolic disorders and the subsequent disease complications, which is a significant increase from 19% in 1997 (The Social Report, 2010). The ever-increasing rate of metabolic disorders warrants a deeper understanding into their molecular mechanisms and development of novel therapeutics for combating these disorders. Although mouse models have greatly contributed to our understanding of metabolic disorders, performing large-scale genetic

and chemical screens in this model is both challenging and expensive. In contrast, zebrafish are small sized vertebrates that closely resemble humans in development and metabolic processes. Owing to their small size, zebrafish can be stored at higher densities than mice and therefore are more economical to raise in large numbers. Further, over their lifetime, zebrafish can produce thousands of offspring that are optically clear during early developmental stages allowing high-resolution visualization of biological processes. These features make zebrafish a valuable system for creating metabolic disease models. In this direction, we created ZFNs and TALENs to mutate several genes in zebrafish that when knocked-out in mice induce obesity, type 2 diabetes and atherosclerosis.

We demonstrated activity of ZFNs for targeting *mc3r* (*melanocortin receptor 3*) and *mc4r* (*melanocortin receptor 4*) genes in zebrafish in Chapter IV and identified founders that harbor mutations in these genes. In mice, *mc4r* deletion leads to increased food uptake resulting in increased adiposity and linear growth[246,247]. Although *mc3r* has been shown to play a role in energy homeostasis, its exact function is still unknown[247]. In zebrafish, overexpression of *agrp* (*agouti-related peptide*), an antagonist of melanocortin receptors, results in increased body weight[248]. Also, the *mc4r* mutant fish obtained through TILLING show increased linear growth as compared to the wild type demonstrating that the melanocortin system is conserved in the zebrafish and when blocked can disturb the energy homeostasis[247-249]. Moreover, polymorphisms and copy number variations in the *mc4r* gene have been linked to body size and onset of maturity in the fish from the *Xiphophorus* genus[250]. Therefore, we expect that the *mc4r* and *mc3r*

mutants created using ZFNs would induce obesity in zebrafish upon feeding a high-fat diet or *ad libitum* **(Table 5-1)**.

To create other models of obesity we employed TALENs to disrupt the *lepa* and *lepb* genes that presumably code for leptin. Leptin, a hormone secreted predominantly by the adiopocytes, is important for the regulation of food intake and energy homeostasis in mice and humans[251,252]. *leptin* knockout in mice results in excessive food uptake, leading to obesity[251]. Humans deficient in *leptin* show early onset of obesity, high fat mass, and impaired satiety with marked hyperphagia[253]. Unlike humans and mice, zebrafish have two copies of *leptin* (*lepa* and *lepb*) that are only 18% similar to the human *leptin*[254]. Since the functions of the two leptin genes have not been characterized in fish, we mutated both of them using TALENs. The double mutants of leptin genes are expected to be hyperphagic and thus obese.

We mutated another component of the leptin signaling, the *leptin receptor* (*lepr*) using ZFNs which is expressed mainly in the hypothalamus and is required for leptin signalling[255]. *lepr* mutants in mice are severely diabetic and also develop obesity[255].

In our attempts to create diabetes models and simultaneously disrupt micro-RNA function in zebrafish, we constructed ZFNs that could mutate *miR-375-1* and *miR-375-2*. In mice, *miR-375* knockout results in hyperglycemia, increased pancreatic α-cell mass and decreased β-cell mass[256]. In zebrafish, somewhat consistent with the mouse knockout results, inhibition of *miR-375-1* and *miR-375-2* by morpholinos affects

Table 5-1

Gene	Disorder	Status
mc4r	Obesity	Heterozygous animals available
mc3r	Obesity	Heterozygous animals available
lepa / lepb	Obesity	Active TALENs available
lepr	Type 2 Diabetes	Active ZFNs available
miR-375-1 / miR-375-2	Type 2 Diabetes	Active ZFNs available
apoeb	Atherosclerosis	Founders Identified

Table 5-1: Summary of genes successfully targeted using ZFNs or TALENs to create zebrafish metabolic disease models.

pancreatic islet development. Therefore, the ZFN-induced mutants of *miR-375-1* and *miR-375-2* would be used to study diabetes and pancreatic development.

Finally, to create atherosclerosis models, we intend to mutate the *apolipoproteinE* genes (*apoEa and apoEb*). Mice and humans harbor a single copy of the *apoE* gene deletion of which results in hypercholesterolemia and aortic atherosclerotic plaque formation. Zebrafish model of atherosclerosis would provide advantage over mouse models since aortic plaques in fish can be visualized in live animals as compared to post-mortem in mice[257]. Although we could create active ZFNs for targeting *apoEb* gene, *apoEa* gene may also have to be disrupted to model atherosclerosis.

Since nutrition can have a dramatic influence on the progression and control of each of these diseases, we will test different feeding regimens ranging in fat and protein contents and the type of fats on these mutants. Moreover, for the atherosclerosis model, we will feed high cholesterol diet supplemented with fluorophore-labeled cholesterol that would help visualize lipid deposition in zebrafish vasculature[257]. This 'nutrition screening' would also allow us to study the effect of diet on early adipogenesis[258258]. Finally, these mutants can be utilized to perform both forward-genetic screens and small-chemical screens to identify new drug targets and therapies for control of metabolic disorders.

Summary

Zinc finger nucleases have shown a tremendous potential for gene targeting in a variety of cell lines and model organisms. Consequently, their use in gene therapy based applications is currently being evaluated in clinical and preclinical trials. Our efforts to optimize the inter-finger interactions between zinc fingers improved both the success rate and the targeting range of ZFNs resulting in ~80% success rate and more than 5-fold targeting density than the previously published zinc finger archives. However, there is both room and need for the improvement in the zinc finger design to enhance the activity and precision of ZFNs. Recently, TALENs have emerged as another tool for targeted gene editing that appear to be free of the limitations of context dependent interactions. Owing to their high success rate and virtually unlimited targeting range, TALENs may outperform ZFNs for gene targeting in model organisms but ZFNs will continue to be employed in gene therapy based applications.

APPENDIX

Table A-1:

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
Asn	TAATTA	gaAAcg	5	10	1	RSDTLEA// QRGNLTR
Asn	TAATTA	gaAAcg	5	10	2	RSDTLMV// QRGNLTR
Asn	TAATTA	gaAAcg	5	10	3	RSDTLVA// QRGNLTR
Asn	TAATTA	gaAAcg	5	10	4	RSDTLNA// QRGNLTR
Asn	TAATTA	gaAAcg	5	10	5	RSDTLAA// QRGNLTR
Asn	TAATTA	gaAAcg	5	10	6	RSDTLRV// QRGNLTR
Asn	TAATTA	gaAAcg	5	10	7	RSDTLLA// QRGNLTR
Asn	TAATTA	gaAAcg	5	10	8	RSDTLKA// QTGNLTR
Asn	TAATTA	gaAAcg	5	10	9	RSDTLKA// QRGNLTR
Asn	TAATTA	gaAAcg	5	10	10	RSDTLKQ// QVGNLTR
Asn	TAATTA	gaAAcg	10	10	1	RSDTLQA// QRGNLTR
Asn	TAATTA	gaAAcg	10	10	2	RSDTLQA// QRGNLTR
Asn	TAATTA	gaAAcg	10	10	3	RSDTLQQ// QRGNLTR
Asn	TAATTA	gaAAcg	10	10	4	RSDTLMA// QRGNLTR
Asn	TAATTA	gaAAcg	10	10	5	RSDTLVQ// QRGNLTR
Asn	TAATTA	gaAAcg	10	10	6	RSDTLMA// QRGNLTR
Asn	TAATTA	gaAAcg	10	10	7	RSDTLAQ// QRGNLTR
Asn	TAATTA	gaAAcg	10	10	8	RSDTLAA// QRGNLTR
Asn	TAATTA	gaAAcg	10	10	9	RSDTLAQ// QSGNLTR
Asn	TAATTA	gaAAcg	10	10	10	RSDTLTQ// QRGNLTR
His	TAATTA	gaAAcg	2	0	1	RSDTLLV// VLQHLTR
His	TAATTA	gaAAcg	2	0	2	RSDTLDA// GLEHLTR
His	TAATTA	gaAAcg	2	0	3	RSDTLVQ// QRGHLTR
His	TAATTA	gaAAcg	2	0	4	RSDTLDQ// QRIHLTR
His	TAATTA	gaAAcg	2	0	5	
His	TAATTA	gaAAcg	2	0	6	RSDTLRV// QVGHLTR
His	TAATTA	gaAAcg	2	0	7	RSDTLAG// ESSHLTR
His	TAATTA	gaAAcg	2	0	8	RSDTLRT// QRVHLTR
His	TAATTA	gaAAcg	2	0	9	RSDTLKQ// QRVHLTR
His	TAATTA	gaAAcg	2	0	10	RSDTLRV// QSGHLTR
His	TAATTA	gaAAcg	5	10	1	RSDTLKQ// QRIHLTR
His	TAATTA	gaAAcg	5	10	2	
His	TAATTA	gaAAcg	5	10	3	RSDTLTQ// QRGHLTR
His	TAATTA	gaAAcg	5	10	4	RSDTLQQ// QRVHLTR
His	TAATTA	gaAAcg	5	10	5	RSDTLRQ// QRVHLTR
His	TAATTA	gaAAcg	5	10	6	
His	TAATTA	gaAAcg	5	10	7	RSDTLTQ// QRGHLTR
His	TAATTA	gaAAcg	5	10	8	
Asn	TAATTA	gaACcg	5		1	RSDTLDA// QSSNLTR
Asn	TAATTA	gaACcg	5		2	RSDTLIT// QGGNLTR
Asn	TAATTA	gaACcg	5		3	RSDTLAD// QAGNLTR
Asn	TAATTA	gaACcg	5		4	RSDTLKA// VRGNLTR
Asn	TAATTA	gaACcg	5		5	RSDTLKE// QRANLTR
Asn	TAATTA	gaACcg	5		6	RSDTLAD// QHGNLTR
Asn	TAATTA	gaACcg	5		7	RSDTLDA// QAGNLTR
Asn	TAATTA	gaACcg	5		8	RSDTLVA// QRGNLTR
Asn	TAATTA	gaACcg	10	10	1	RSDTLMA// QSGNLTR

Table A-1 contd.

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -11234**56**//-112**3**456
Asn	TAATTA	gaACcg	10	10	2	RSDTLTA// QKCNLTR
Asn	TAATTA	gaACcg	10	10	3	RSDTLKQ// QCGNLTR
Asn	TAATTA	gaACcg	10	10	4	RSDTLKQ// QCGNLTR
Asn	TAATTA	gaACcg	10	10	5	RSDTLKE// QHSNLTR
Asn	TAATTA	gaACcg	10	10	6	RSDTLVE// QRGNLTR
Asn	TAATTA	gaACcg	10	10	7	RSDTLLQ// QRSNLTR
Asn	TAATTA	gaACcg	10	10	8	RSDTLIE// QRGNLTR
HIS	TAATTA	gaACcg	2	10	1	RSDTLAT// QRGHLTR
HIS	TAATTA	gaACcg	2	10	2	RSDTLTA// QGGHLTR
HIS	TAATTA	gaACcg	2	10	3	RSDTLRE// QRGHLTR
HIS	TAATTA	gaACcg	2	10	4	RSDTLRA// QGGHLTR
HIS	TAATTA	gaACcg	2	10	5	RSDTLKA// QGGHLTR
HIS	TAATTA	gaACcg	2	10	6	RSDTLQA// QQGHLTR
HIS	TAATTA	gaACcg	2	10	7	RSDTLTQ// QSGHLTR
HIS	TAATTA	gaACcg	2	10	8	RSDTLLD// QRGHLTR
His	TAATTA	gaACcg	5	10	1	RSDTLKE// QRIHLTR
His	TAATTA	gaACcg	5	10	2	RSDTLKA// QGGHLTR
His	TAATTA	gaACcg	5	10	3	RSDTLAA// QSGHLTR
His	TAATTA	gaACcg	5	10	4	RSDTLKA// QGGHLTR
His	TAATTA	gaACcg	5	10	5	RSDTLRA// QRGHLTR
His	TAATTA	gaACcg	5	10	6	RSDTLAE// QRGHLTR
His	TAATTA	gaACcg	5	10	7	RSDTLKI// QRGHLTR
His	TAATTA	gaACcg	5	10	8	RSDTLRA// QGGHLTR
Asn	TAATTA	gaAGcg	10- URA	0	1	RSDTLVR// QLSNLTR
Asn	TAATTA	gaAGcg	10- URA	0	2	RSDTLAR// QGCNLTR
Asn	TAATTA	gaAGcg	10- URA	0	3	RSDTLLR// QKCNLTR
Asn	TAATTA	gaAGcg	10- URA	0	4	RSDTLVR// QEGNLTR
Asn	TAATTA	gaAGcg	10- URA	0	5	RSDTLKK// QTCNLTR
Asn	TAATTA	gaAGcg	10- URA	0	6	RSDTLDR// QGGNLTR
Asn	TAATTA	gaAGcg	25-URA	0	1	RSDTLQR// QKSNLTR
Asn	TAATTA	gaAGcg	25-URA	0	2	RSDTLLR// QNSNLTR
Asn	TAATTA	gaAGcg	25-URA	0	3	RSDTLRR// QGANLTR
Asn	TAATTA	gaAGcg	25-URA	0	4	RSDTLQR// QGANLTR
Asn	TAATTA	gaAGcg	25-URA	0	5	RSDTLAR// QCSNLTR
Asn	TAATTA	gaAGcg	25-URA	0	6	RSDTLLR// QGANLTR
Asn	TAATTA	gaAGcg	25-URA	0	7	RSDTLLK// QSCNLTR
Asn	TAATTA	gaAGcg	25-URA	0	8	RSDTLKR// XXCNLTR
His	TAATTA	gaAGcg	5	10	1	RSDTLKR// QVAHLTR
His	TAATTA	gaAGcg	5	10	2	RSDTLVR// QSSHLTR
His	TAATTA	gaAGcg	5	10	3	RSDTLVR// QNGHLTR
His	TAATTA	gaAGcg	5	10	4	RSDTLLR// QSVHLTR
His	TAATTA	gaAGcg	5	10	5	RSDTLRR// QCYHLTR
His	TAATTA	gaAGcg	5	10	6	RSDTLRR// QRGHLTR
His	TAATTA	gaAGcg	5	10	7	RSDTLER// QRGHLTR
His	TAATTA	gaAGcg	5	10	8	RSDTLIR// QAGHLTR
His	TAATTA	gaAGcg	10	10	1	RSDTLVR// QSGHLTR
His	TAATTA	gaAGcg	10	10	2	RSDTLRR// QAYHLTR

Table A-1

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
His	TAATTA	gaAGcg	10	10	3	
His	TAATTA	gaAGcg	10	10	4	RSDTLKK// QSGHLTR
His	TAATTA	gaAGcg	10	10	5	RSDTLAR// QGGHLTR
His	TAATTA	gaAGcg	10	10	6	RSDTLRR// QSYHLTR
His	TAATTA	gaAGcg	10	10	7	RSDTLAR// QSHHLTR
His	TAATTA	gaAGcg	10	10	8	RSDTLAR// QQYHLTR
Asn	TAATTA	gaATcg	10	10	1	RSDTLRL// QRGNLTR
Asn	TAATTA	gaATcg	10	10	2	RSDTLKV// QYGNLTR
Asn	TAATTA	gaATcg	10	10	3	RSDTLTV// QRSNLTR
Asn	TAATTA	gaATcg	10	10	4	RSDTLEA// QRSNLTR
Asn	TAATTA	gaATcg	10	10	5	RSDTLAA// QGGNLTR
Asn	TAATTA	gaATcg	10	10	6	RSDTLAA// QSGNLTR
Asn	TAATTA	gaATcg	10	10	7	RSDTLKI// VRSNLTR
Asn	TAATTA	gaATcg	10	10	8	RSDTLTA// QKGNLTR
Asn	TAATTA	gaATcg	25	10	1	RSDTLRT// QRSNLTR
Asn	TAATTA	gaATcg	25	10	2	RSDTLRA// QRGNLTR
Asn	TAATTA	gaATcg	25	10	3	RSDTLKA// QRSNLTR
Asn	TAATTA	gaATcg	25	10	4	RSDTLKQ// QSSNLTR
Asn	TAATTA	gaATcg	25	10	5	
Asn	TAATTA	gaATcg	25	10	6	RSDTLQ(A/G)// QXSNLTR
His	TAATTA	gaATcg	5	10	1	RSDTLVI// QRIHLTR
His	TAATTA	gaATcg	5	10	2	RSDTLKV// QRCHLTR
His	TAATTA	gaATcg	5	10	3	RSDTLKA// QRIHLTR
His	TAATTA	gaATcg	5	10	4	RSDTLKT// QRIHLTR
His	TAATTA	gaATcg	5	10	5	
His	TAATTA	gaATcg	5	10	6	
His	TAATTA	gaATcg	5	10	7	
His	TAATTA	gaATcg	5	10	8	
His	TAATTA	gaATcg	10	10	1	RSDTLKI// QKVHLTR
His	TAATTA	gaATcg	10	10	2	RSDTLKI// QRVHLTR
His	TAATTA	gaATcg	10	10	3	RSDTLKV// QRIHLTR
His	TAATTA	gaATcg	10	10	4	RSDTLKV// QRIHLTR
His	TAATTA	gaATcg	10	10	5	RSDTLKI// QGIHLTR
His	TAATTA	gaATcg	10	10	6	RSDTLKV// QRIHLTR
His	TAATTA	gaATcg	10	10	7	RSDTLKV// QRIHLTR
His	TAATTA	gaATcg	10	10	8	RSDTLKV// QRIHLTR
Asn	TAATTA	gaCAcg	5	10	1	RSDTLAE// CARNLTR
Asn	TAATTA	gaCAcg	5	10	2	RSDTLAE// SRRNLTR
Asn	TAATTA	gaCAcg	5	10	3	RSDTLAE// VARNLTR
Asn	TAATTA	gaCAcg	5	10	4	RSDTLML// CRSNLTR
Asn	TAATTA	gaCAcg	5	10	5	RSDTLSE// AARNLTR
Asn	TAATTA	gaCAcg	5	10	6	RSDTLME// CRSNLTR
Asn	TAATTA	gaCAcg	5	10	7	RSDTLAE// TTRNLTR
Asn	TAATTA	gaCAcg	5	10	8	RSDTLRE// VSRNLTR
Asn	TAATTA	gaCAcg	10	10	1	RSDTLKE// VGRNLTR
Asn	TAATTA	gaCAcg	10	10	2	RSDTLKE// VLRNLTR
Asn	TAATTA	gaCAcg	10	10	3	RSDTLGR// ARRNLTR

Table A-1

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
Asn	TAATTA	gaCAcg	10	10	4	RSDTLAE// VRRNLTR
Asn	TAATTA	gaCAcg	10	10	5	RSDTLKE// VSRNLTR
Asn	TAATTA	gaCAcg	10	10	6	RSDTLQE// TARNLTR
Asn	TAATTA	gaCAcg	10	10	7	RSDTLAE// ARRNLTR
Asn	TAATTA	gaCAcg	10	10	8	RSDTLKQ// CKPNLTR
Asn	TAATTA	gaCAcg	10	10	9	RSDTLVE// CKPNLTR
Asn	TAATTA	gaCAcg	25	10	1	RSDTLKE// GRSNLTR
Asn	TAATTA	gaCAcg	25	10	2	RSDTLKD// IRRNLTR
Asn	TAATTA	gaCAcg	25	10	3	RSDTLKE// RRSNLTR
Asn	TAATTA	gaCAcg	25	10	4	RSDTLKQ// DRRNLTR
Asn	TAATTA	gaCAcg	25	10	5	RSDTLKE// SKSNLTR
Asn	TAATTA	gaCAcg	25	10	6	RSDTLKE// VRRNLTR
Asn	TAATTA	gaCAcg	25	10	7	RSDTLKE// CRRNLTR
Asn	TAATTA	gaCAcg	25	10	8	RSDTLKQ// DKRNLTR
HIS	TAATTA	gaCAcg	5	10	1	RSDTLND// TRRHLTR
HIS	TAATTA	gaCAcg	5	10	2	RSDTLKE// TRRHLTR
HIS	TAATTA	gaCAcg	5	10	3	RSDTLAE// TRRHLTR
HIS	TAATTA	gaCAcg	5	10	4	RSDTLKR// ERGHLTR
HIS	TAATTA	gaCAcg	5	10	5	
HIS	TAATTA	gaCAcg	5	10	6	RSDTLKE// ARRHLTR
HIS	TAATTA	gaCAcg	5	10	7	
HIS	TAATTA	gaCAcg	5	10	8	RSDTLKE// ARRHLTR
Asn	TAATTA	gaCCcg	10	10	1	RSDTLKE// YRSNLTR
Asn	TAATTA	gaCCcg	10	10	2	
Asn	TAATTA	gaCCcg	10	10	3	RSDTLRE// CRSNLTR
Asn	TAATTA	gaCCcg	10	10	4	
Asn	TAATTA	gaCCcg	10	10	5	RSDTLKD// IRSNLTR
Asn	TAATTA	gaCCcg	10	10	6	RSDTLKD// CHRNLTR
Asn	TAATTA	gaCCcg	10	10	7	RSDTLAE// GRSNLTR
Asn	TAATTA	gaCCcg	10	10	8	RSDTLLE// CRSNLTR
His	TAATTA	gaCCcg	10	10	1	
His	TAATTA	gaCCcg	10	10	2	RSDTLAT// DRSHLTR
His	TAATTA	gaCCcg	10	10	3	RSDTLDP// LYEHLTR
His	TAATTA	gaCCcg	10	10	4	RSDTLKD// TRKHLTR
His	TAATTA	gaCCcg	10	10	5	RSDTLKE// LRRHLTR
His	TAATTA	gaCCcg	10	10	6	RSDTLKA// ERGHLTR
His	TAATTA	gaCCcg	10	10	7	RSDTLKE// LRRHLTR
His	TAATTA	gaCCcg	10	10	8	
Asn	TAATTA	gaCGcg	10	10	1	RSDTLKQ// CASNLTR
Asn	TAATTA	gaCGcg	10	10	2	RSDTLKR// CRGNLTR
Asn	TAATTA	gaCGcg	10	10	3	RSDTLKR// EASNLTR
Asn	TAATTA	gaCGcg	10	10	4	RSDTLKR// DRRNLTR
Asn	TAATTA	gaCGcg	10	10	5	RSDTLKL// GRSNLTR
Asn	TAATTA	gaCGcg	10	10	6	RSDTLAR// EGGNLTR
Asn	TAATTA	gaCGcg	10	10	7	RSDTLKR// DRGNLTR
Asn	TAATTA	gaCGcg	10	10	8	RSDTLVR// ERGNLTR
HIS	TAATTA	gaCGcg	5	10	1	RSDTLRR// ESGHLTR
HIS	TAATTA	gaCGcg	5	10	2	RSDTLKR// EGGHLTR
HIS	TAATTA	gaCGcg	5	10	3	RSDTLRR// ERGHLTR

Table A-1

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
HIS	TAATTA	gaCGcg	5	10	4	RSDTLLR// ERGHLTR
HIS	TAATTA	gaCGcg	5	10	5	RSDTLKR// EGGHLTR
HIS	TAATTA	gaCGcg	5	10	6	RSDTLLR// ESGHLTR
HIS	TAATTA	gaCGcg	5	10	7	
HIS	TAATTA	gaCGcg	5	10	8	RSDTLLR// ERGHLTR
His	TAATTA	gaCGcg	10	10	1	RSDTLKR// ERGHLTR
His	TAATTA	gaCGcg	10	10	2	RSDTLRR// ESGHLTR
His	TAATTA	gaCGcg	10	10	3	RDSTLKR// EQGHTLR
His	TAATTA	gaCGcg	10	10	4	RSDTLKR// ERGHLTR
His	TAATTA	gaCGcg	10	10	5	RSDTLKR// EGGHLTR
His	TAATTA	gaCGcg	10	10	6	RSDTLRR// ERGHLTR
His	TAATTA	gaCGcg	10	10	7	RSDTLRK// EKFHLTR
His	TAATTA	gaCGcg	10	10	8	RSDTLRR// EKGHLTR
Asn	TAATTA	gaCTcg	5	10	1	RSDTLSL// CRANLTR
Asn	TAATTA	gaCTcg	5	10	2	RSDTLKG// DRSNLTR
Asn	TAATTA	gaCTcg	5	10	3	RSDTLKL// GGSNLTR
Asn	TAATTA	gaCTcg	5	10	4	RSDTLAV// DRSNLTR
Asn	TAATTA	gaCTcg	5	10	5	RSDTLKM// NASNLTR
Asn	TAATTA	gaCTcg	5	10	6	RSDTLVR// DPCNLTR
Asn	TAATTA	gaCTcg	5	10	7	RSDTLKE// GRSNLTR
Asn	TAATTA	gaCTcg	5	10	8	RSDTLRD// CRSNLTR
Asn	TAATTA	gaCTcg	5	10	9	RSDTLKE// SKSNLTR
Asn	TAATTA	gaCTcg	10	10	1	RSDTLKL// CGSNLTR
Asn	TAATTA	gaCTcg	10	10	2	RSDTLRL// CSSNLTR
Asn	TAATTA	gaCTcg	10	10	3	RSDTLVL// CKSNLTR
Asn	TAATTA	gaCTcg	10	10	4	RSDTLAG// DRSNLTR
Asn	TAATTA	gaCTcg	10	10	5	RSDTLKL// CASNLTR
Asn	TAATTA	gaCTcg	10	10	6	RSDTLKG// DRCNLTR
Asn	TAATTA	gaCTcg	10	10	7	RSDTLQL// CRSNLTR
Asn	TAATTA	gaCTcg	10	10	8	RSDTLAL// CRCNLTR
His	TAATTA	gaCTcg	10	10	1	RSDTLQT// GDLHQTR
His	TAATTA	gaCTcg	10	10	2	RSDTLLR// HYAHLTR
His	TAATTA	gaCTcg	10	10	3	RSDTLTR// SPCHLTR
His	TAATTA	gaCTcg	10	10	4	RSDTLKG// GLLHLTR
His	TAATTA	gaCTcg	10	10	5	RSDTLRL// CYSNLTR
His	TAATTA	gaCTcg	10	10	6	RSDTLLA// RPVHLTR
His	TAATTA	gaCTcg	10	10	7	bad read
His	TAATTA	gaCTcg	10	10	8	RSDTLRL// CYSNLTR
His	TAATTA	gaCTcg	10	10	9	RSDTLPE// SGDHLTR
His	TAATTA	gaCTcg	10	10	10	bad read
His	TAATTA	gaCTcg	10	10	11	RSDTLGR// VESHLTR
His	TAATTA	gaCTcg	10	10	12	RSDTLGG// GDHHLTR
His	TAATTA	gaCTcg	5	10	1	RSDTLGR// G*RHLTR
His	TAATTA	gaCTcg	5	10	2	RSDTLEV// VSPHLTR
His	TAATTA	gaCTcg	5	10	3	bad read
Asn	TAATTA	gaGAcg	10	10	1	RSDTLRE// TRRNLTR
Asn	TAATTA	gaGAcg	10	10	2	RSDTLAE// AKRNLTR
Asn	TAATTA	gaGAcg	10	10	3	RSDTLKE// CSRNLTR
Asn	TAATTA	gaGAcg	10	10	4	RSDTLAN// RKGNLTR

Table A-1

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
Asn	TAATTA	gaGAcg	10	10	5	RSDTLVE// VHRNLTR
Asn	TAATTA	gaGAcg	10	10	6	RSDTLKE// VRRNLTR
Asn	TAATTA	gaGAcg	10	10	7	RSDTLRE// AARNLTR
Asn	TAATTA	gaGAcg	10	10	8	
HIS	TAATTA	gaGAcg	5	10	1	RSDTLKE// TTRHLTR
HIS	TAATTA	gaGAcg	5	10	2	RSDTLAD// VRRHLTR
HIS	TAATTA	gaGAcg	5	10	3	RSDTLKE// VSRHLTR
HIS	TAATTA	gaGAcg	5	10	4	RSDTLVE// RKRHLTR
HIS	TAATTA	gaGAcg	5	10	5	RSDTLRE// VRRHLTR
HIS	TAATTA	gaGAcg	5	10	6	RSDTLKE// VGRHLTR
HIS	TAATTA	gaGAcg	5	10	7	RSDTLKE// RRRHLTR
HIS	TAATTA	gaGAcg	5	10	8	RSDTLRD// VRRHLTR
His	TAATTA	gaGAcg	10	10	1	RSDTLKE// VRRHLTR
His	TAATTA	gaGAcg	10	10	2	RSDTLKE// TKRHLTR
His	TAATTA	gaGAcg	10	10	3	RSDTLKE// VRRHLTR
His	TAATTA	gaGAcg	10	10	4	RSDTLKE// VARHLTR
His	TAATTA	gaGAcg	10	10	5	RSDTLKE// VRRHLTR
His	TAATTA	gaGAcg	10	10	6	
His	TAATTA	gaGAcg	10	10	7	RDTLKE// VRRHLTR
His	TAATTA	gaGAcg	10	10	8	
Asn	TAATTA	gaGCcg	25	10	1	RSDTLKA// KRSNLTR
Asn	TAATTA	gaGCcg	25	10	2	RSDTLRS// RRFNLTR
Asn	TAATTA	gaGCcg	25	10	3	RSDTLKA// KRYNLTR
Asn	TAATTA	gaGCcg	25	10	4	
Asn	TAATTA	gaGCcg	25	10	5	RSDTLRS// RAYNLTR
Asn	TAATTA	gaGCcg	25	10	6	RSDTLAA// RNSNLTR
Asn	TAATTA	gaGCcg	25	10	7	RSDTLRS// KKYNLTR
Asn	TAATTA	gaGCcg	25	10	8	RSDTLRA// RNFNLTR
Asn	TAATTA	gaGCcg	25-Ura	0	1	RSDTLKE// KGFNLTR
Asn	TAATTA	gaGCcg	25-Ura	0	2	RSDTLKA// ARYNLTR
Asn	TAATTA	gaGCcg	25-Ura	0	3	RSDTLKE// RRSNLTR
Asn	TAATTA	gaGCcg	25-Ura	0	4	RSDTLKE// KRYNLTR
Asn	TAATTA	gaGCcg	25-Ura	0	5	RSDTLRD// KRFNLTR
Asn	TAATTA	gaGCcg	25-Ura	0	6	RSDTLRE// KSGNLTR
Asn	TAATTA	gaGCcg	25-Ura	0	7	RSDTLKE// KACNLTR
Asn	TAATTA	gaGCcg	25-Ura	0	8	RSDTLKA// QRFNLTR
Asn	TAATTA	gaGCcg	25-Ura	0	9	RSDTLKE// RSSNLTR
His	TAATTA	gaGCcg	5	10	1	RSDTLKE// RKGHLTR
His	TAATTA	gaGCcg	5	10	2	RSDTLKE// RRYHLTR
His	TAATTA	gaGCcg	5	10	3	
His	TAATTA	gaGCcg	5	10	4	RSDTLRD// RRGHLTR
His	TAATTA	gaGCcg	5	10	5	RSDTLAD// RRSHLTR
His	TAATTA	gaGCcg	5	10	6	RSDTLAD// RSSHLTR
His	TAATTA	gaGCcg	5	10	7	
His	TAATTA	gaGCcg	5	10	8	RSDTLRD// RRQHLTR
His	TAATTA	gaGCcg	10	10	1	
His	TAATTA	gaGCcg	10	10	2	RSDTLKE// RRSHLTR
His	TAATTA	gaGCcg	10	10	3	RSDTLKE// RSSHLTR
His	TAATTA	gaGCcg	10	10	4	

Table A-1

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
His	TAATTA	gaGCcg	10	10	5	RSDTLKE// RSSHLTR
His	TAATTA	gaGCcg	10	10	6	RSDTLKE// RRTHLTR
His	TAATTA	gaGCcg	10	10	7	RSDTLRD// RRQHLTR
His	TAATTA	gaGCcg	10	10	8	RSDTLKE// RRSHLTR
Asn	TAATTA	gaGGcg	25-Ura	0	1	RSDTLRR// VQYNLTR
Asn	TAATTA	gaGGcg	25-Ura	0	2	RSDTLIR// RAENLTR
Asn	TAATTA	gaGGcg	25-Ura	0	3	RSDTLKR// CRFNLTR
Asn	TAATTA	gaGGcg	25-Ura	0	4	RSDTLKR// AQGNLTR
Asn	TAATTA	gaGGcg	25-Ura	0	5	RSDTLKR// TTGNLTR
Asn	TAATTA	gaGGcg	25-Ura	0	6	RSDTLAR// GPQNLTR
Asn	TAATTA	gaGGcg	25-Ura	0	7	RSDTLVR// TRFNLTR
Asn	TAATTA	gaGGcg	25-Ura	0	8	RSDTLAR// AAYNLTR
Asn	TAAAGG	gaGGcg	10	10	1	RSDTLER// RTDNLTR
Asn	TAAAGG	gaGGcg	10	10	2	RSDTLER// RCDNLTR
Asn	TAAAGG	gaGGcg	10	10	3	RSDTLKR// RIDNLTR
Asn	TAAAGG	gaGGcg	10	10	4	RSDTLLR// RQDNLTR
Asn	TAAAGG	gaGGcg	10	10	5	RSDTLAA// FRRNLTR
Asn	TAAAGG	gaGGcg	10	10	6	RSDTLVR// RQDHLTR
Asn	TAAAGG	gaGGcg	10	10	7	RSDTLLR// RTDNLTR
Asn	TAAAGG	gaGGcg	10	10	8	RSDTLER// RHDNLTR
Asn	TAATTA	gaGGcg	50	10	1	RSDTLER// RRSNLTR
Asn	TAATTA	gaGGcg	50	10	2	RSDTLGH// *ECNLTR
Asn	TAATTA	gaGGcg	50	10	3	RSDTLAR// RSCNLTR
Asn	TAATTA	gaGGcg	50	10	4	RSDTLKL// RRYNLTR
Asn	TAATTA	gaGGcg	50	10	5	RSDTLKR// RGSNLTR
Asn	TAATTA	gaGGcg	50	10	6	RSDTLKR// RNYNLTR
Asn	TAATTA	gaGGcg	50	10	7	RSDTLVR// RRYNLR
Asn	TAATTA	gaGGcg	50	10	8	RSDTLRR// RRYNLTR
Asn	TAATTA	gaGGcg	50	10	9	RSDTLVE// KKYNLTR
Asn	TAATTA	gaGGcg	50	10	10	RSDTLAR// SRFNLTR
Asn	TAATTA	gaGGcg	50	10	11	RSDTLAR// RRYNLTR
Asn	TAATTA	gaGGcg	50	10	12	RSDTLAR// RRFNLTR
His	TAATTA	gaGGcg	10	10	1	RSDTLRR// RSCHLTR
His	TAATTA	gaGGcg	10	10	2	RSDTLAR// RFDHLTR
His	TAATTA	gaGGcg	10	10	3	RSDTLER// RQCHLTR
His	TAATTA	gaGGcg	10	10	4	RSDTLMR// RFDHLTR
His	TAATTA	gaGGcg	10	10	5	
His	TAATTA	gaGGcg	10	10	6	
His	TAATTA	gaGGcg	10	10	7	
His	TAATTA	gaGGcg	10	10	8	RSDTLQR// RGCHLTR
His	TAATTA	gaGGcg	25	10	1	
His	TAATTA	gaGGcg	25	10	2	RSDTLVR// RAEHLTR
His	TAATTA	gaGGcg	25	10	3	RSDTLSR// RREHLTR
His	TAATTA	gaGGcg	25	10	4	RSDTLAR// RAEHLTR
His	TAATTA	gaGGcg	25	10	5	RSDTLVR// RLDHLTR
His	TAATTA	gaGGcg	25	10	6	RSDTLIR// RYDHLTR
His	TAATTA	gaGGcg	25	10	7	
His	TAATTA	gaGGcg	25	10	8	RSDTLKR// RSCHLTR
Asn	TAATTA	gaGTcg	5	10	1	RSDTLKE// RSCNLTR

Table A-1

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
Asn	TAATTA	gaGTcg	5	10	2	RSDTLRE// KACNLTR
Asn	TAATTA	gaGTcg	5	10	3	RSDTLKV// TSSNLTR
Asn	TAATTA	gaGTcg	5	10	4	RSDTLLA// KGCNLTR
Asn	TAATTA	gaGTcg	5	10	5	RSDTLRG// KSCNLTR
Asn	TAATTA	gaGTcg	5	10	6	RSDTLKA// RADNLTR
Asn	TAATTA	gaGTcg	5	10	7	RSDTLKE// VRRNLTR
Asn	TAATTA	gaGTcg	5	10	8	RSDTLKL// SGSNLTR
Asn	TAATTA	gaGTcg	5	10	9	
Asn	TAATTA	gaGTcg	5	10	10	RSDTLVE// VRRNLTR
Asn	TAATTA	gaGTcg	10	10	1	RSDTLME// KSCNLTR
Asn	TAATTA	gaGTcg	10	10	2	RSDTLME// KSCNLTR
Asn	TAATTA	gaGTcg	10	10	3	RSDTLLE// IKRNLTR
Asn	TAATTA	gaGTcg	10	10	4	RSDTLIE// IKRNLTR
Asn	TAATTA	gaGTcg	10	10	5	
Asn	TAATTA	gaGTcg	10	10	6	RSDTLVE// KSCNLTR
Asn	TAATTA	gaGTcg	10	10	7	RSDTLRE// KQCNLTR
Asn	TAATTA	gaGTcg	10	10	8	RSDTLKE// KGCNLTR
Asn	TAATTA	gaGTcg	10	10	9	RSDTLIR// ASSNLTR
Asn	TAATTA	gaGTcg	10	10	10	RSDTLVE// KGCNLTR
Asn	TAATTA	gaGTcg	10	0	1	RSDTLKE// KRCNLTR
Asn	TAATTA	gaGTcg	10	0	2	RSDTLRV// RSGNLTR
Asn	TAATTA	gaGTcg	10	0	3	RSDTLRE// KSCNLTR
Asn	TAATTA	gaGTcg	10	0	4	RSDTLME// KRCNLTR
Asn	TAATTA	gaGTcg	10	0	5	RSDTLKE// KSCNLTR
Asn	TAATTA	gaGTcg	10	0	6	RSDTLME// KRCNLTR
Asn	TAATTA	gaGTcg	10	0	7	RSDTLKE// KGCNLTR
Asn	TAATTA	gaGTcg	10	0	8	RSDTLVE// KSCNLTR
Asn	TAATTA	gaGTcg	10	0	9	RSDTLVE// KRCNLTR
Asn	TAATTA	gaGTcg	10	0	10	RSDTLRE// KRCNLTR
His	TAATTA	gaGTcg	5	10	1	RSDTLQA// AHAHLTR
His	TAATTA	gaGTcg	5	10	2	RSDTLVG// VRQHLTR
His	TAATTA	gaGTcg	5	10	3	RSDTLPM// RSRHLTR
His	TAATTA	gaGTcg	5	10	4	RSDTLLV// GAVHLTR
His	TAATTA	gaGTcg	5	10	5	RSDTLVK// RSDHLTR
His	TAATTA	gaGTcg	5	10	6	
His	TAATTA	gaGTcg	5	10	7	RSDTLLK// RGDHLTR
His	TAATTA	gaGTcg	5	10	8	
His	TAATTA	gaGTcg	5	10	9	RSDTLQK// RSDHLTR
His	TAATTA	gaGTcg	2	0	1	RSDTLLL// RSDHLTR
His	TAATTA	gaGTcg	2	0	2	RSDTLLL// RRDHLTR
His	TAATTA	gaGTcg	2	0	3	RSDTLLK// GPGHLTR
His	TAATTA	gaGTcg	2	0	4	RSDTLVR// GIQHLTR
His	TAATTA	gaGTcg	2	0	5	RSDTLKE// HLHHLTR
Asn	TAATTA	gaTAcg	10-Ura	0	1	RSDTLKV// VRGNLTR
Asn	TAATTA	gaTAcg	10-Ura	0	2	RSDTLKQ// AAGNLTR
Asn	TAATTA	gaTAcg	10-Ura	0	3	RSDTLLT// VRGNLTR
Asn	TAATTA	gaTAcg	10-Ura	0	4	RSDTLLV// TRGNLTR
Asn	TAATTA	gaTAcg	10-Ura	0	5	RSDTLAV// VRGNLTR
Asn	TAATTA	gaTAcg	10-Ura	0	6	RSDTLAA// IRGNLTR

Table A-1

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
Asn	TAATTA	gaTAcg	25-URA	0	1	RSDTLKA// VAGNLTR
Asn	TAATTA	gaTAcg	25-URA	0	2	RSDTLKA// VRGNLTR
Asn	TAATTA	gaTAcg	25-URA	0	3	RSDTLKD// VRANLTR
Asn	TAATTA	gaTAcg	25-URA	0	4	RSDTLKV// TVGNLTR
Asn	TAATTA	gaTAcg	25-URA	0	5	RSDTLKV// APGNLTR
Asn	TAATTA	gaTAcg	25-URA	0	6	RSDTLKI// AKGNLTR
Asn	TAATTA	gaTAcg	25-URA	0	7	RSDTLKA// IRANLTR
Asn	TAATTA	gaTAcg	25-URA	0	8	RSDTLKA// VAGNLTR
His	TAATTA	gaTAcg	5	10	1	
His	TAATTA	gaTAcg	5	10	2	RSDTLLD// LRRHLTR
His	TAATTA	gaTAcg	5	10	3	
His	TAATTA	gaTAcg	5	10	4	
His	TAATTA	gaTAcg	5	10	5	
His	TAATTA	gaTAcg	5	10	6	RSDTLRE// LRRHLTR
His	TAATTA	gaTAcg	5	10	7	
His	TAATTA	gaTAcg	5	10	8	RSDTLKD// LRRHLTR
His	TAATTA	gaTAcg	10	10	1	RSDTLKD// LKRHLTR
His	TAATTA	gaTAcg	10	10	2	RSDTLKE// LRRHLTR
His	TAATTA	gaTAcg	10	10	3	RSDTLKE// LKRHLTR
His	TAATTA	gaTAcg	10	10	4	RSDTLKE// LKRHLTR
His	TAATTA	gaTAcg	10	10	5	RSDTLKE// LRRHLTR
His	TAATTA	gaTAcg	10	10	6	RSDTLKE// LKRHLTR
His	TAATTA	gaTAcg	10	10	7	RSDTLMV// ARCNLTR
His	TAATTA	gaTAcg	10	10	8	RSDTLVR// LRCYLTR
His	TAATTA	gaTAcg	10	10	9	RSDTLRE// LKRHLTR
Asn	TAATTA	gaTCcg	10	10	1	RSDTLKE// LKRNLTR
Asn	TAATTA	gaTCcg	10	10	2	RSDTLKD// RRSNLTR
Asn	TAATTA	gaTCcg	10	10	3	RSDTLKD// RKTNLTR
Asn	TAATTA	gaTCcg	10	10	4	
Asn	TAATTA	gaTCcg	10	10	5	RSDTLKD// RSSNLTR
Asn	TAATTA	gaTCcg	10	10	6	RSDTLLM// GGSNLTR
Asn	TAATTA	gaTCcg	10	10	7	RSDTLKE// LNRNLTR
Asn	TAATTA	gaTCcg	10	10	8	RSDTLRD// RSTNLTR
His	TAATTA	gaTCcg	10	10	1	RSDTLKR// ERSHLTR
His	TAATTA	gaTCcg	10	10	2	
His	TAATTA	gaTCcg	10	10	3	RSDTLKD// LKRHLTR
His	TAATTA	gaTCcg	10	10	4	RSDTLKE// IKRHLTR
His	TAATTA	gaTCcg	10	10	5	
His	TAATTA	gaTCcg	10	10	6	RSDTLKV// VKAHLTR
His	TAATTA	gaTCcg	10	10	7	RSDTLKD// TRGHLTR
His	TAATTA	gaTCcg	10	10	8	RSDTLKV// ARGHLTR
His	TAATTA	gaTCcg	5 -Ura	0	1	RSDTLKA// VRNHLTR
His	TAATTA	gaTCcg	5 -Ura	0	2	RSDTLRE// LKRHLTR
His	TAATTA	gaTCcg	5 -Ura	0	3	RSDTLKD// IQRHLTR
His	TAATTA	gaTCcg	5 -Ura	0	4	RDSTLKV// ARGHLTR
His	TAATTA	gaTCcg	5 -Ura	0	5	RSDTLKE// CRRHLTR
His	TAATTA	gaTCcg	5 -Ura	0	6	RDSTLKV// VRAHLTR
His	TAATTA	gaTCcg	5 -Ura	0	7	RSDTLKV// ASGHLTR
His	TAATTA	gaTCcg	5 -Ura	0	8	RSDTLQD// NKRHLTR

Table A-1

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
Asn	TAATTA	gaTGcg	25-Ura	0	1	RSDTLQR// VGSNLTR
Asn	TAATTA	gaTGcg	25-Ura	0	2	
Asn	TAATTA	gaTGcg	25-Ura	0	3	RSDTLVR// HAFNLTR
Asn	TAATTA	gaTGcg	25-Ura	0	4	RSDTLAR// CRGNLTR
Asn	TAATTA	gaTGcg	25-Ura	0	5	RSDTLGR// VRCNLTR
Asn	TAATTA	gaTGcg	25-Ura	0	6	RSDTLAR// VRANLTR
Asn	TAATTA	gaTGcg	25-Ura	0	7	RSDTLVR// CRHNLTR
Asn	TAATTA	gaTGcg	25-Ura	0	8	RSDTLQR// VGGNLTR
Asn	TAATTA	gaTGcg	10	10	1	RSDTLMR// IRSNLTR
Asn	TAATTA	gaTGcg	10	10	2	RSDTLRR// CRFNLTR
Asn	TAATTA	gaTGcg	10	10	3	RSDTLAT// VRGNLTR
Asn	TAATTA	gaTGcg	10	10	4	RSDTLER// VKSNLTR
Asn	TAATTA	gaTGcg	10	10	5	RSDTLGR// VRANLTR
Asn	TAATTA	gaTGcg	10	10	6	
Asn	TAATTA	gaTGcg	10	10	7	
Asn	TAATTA	gaTGcg	10	10	8	RSDTLRA// TGGNLTR
HIS	TAATTA	gaTGcg	5	10	1	
HIS	TAATTA	gaTGcg	5	10	2	RSDTLVR// LRFHLTR
HIS	TAATTA	gaTGcg	5	10	3	RSDTLER// LSFHLTR
HIS	TAATTA	gaTGcg	5	10	4	RSDTLRR// ISFHLTR
HIS	TAATTA	gaTGcg	5	10	5	RSDTLAR// FRFHLTR
HIS	TAATTA	gaTGcg	5	10	6	RSDTLKR// LPFHLTR
HIS	TAATTA	gaTGcg	5	10	7	RSDTLTR// LRYHLTR
HIS	TAATTA	gaTGcg	5	10	8	RSDTLKR// LTFHLTR
His	TAATTA	gaTGcg	10	10	1	RSDTLRR// IRFHLTR
His	TAATTA	gaTGcg	10	10	2	RSDTLKR// CGFHLTR
His	TAATTA	gaTGcg	10	10	3	RSDTLRR// LPFHLTR
His	TAATTA	gaTGcg	10	10	4	RSDTLAR// FRFHLTR
His	TAATTA	gaTGcg	10	10	5	RSDTLGR// LRFHLTR
His	TAATTA	gaTGcg	10	10	6	RSDTLVR// LRYHLTR
His	TAATTA	gaTGcg	10	10	7	RSDTLRR// LSFHLTR
His	TAATTA	gaTGcg	10	10	8	RSDTLQR// VQFHLTR
Asn	TAATTA	gaTTcg	10	10	1	RSDTLRV// VRSNLTR
Asn	TAATTA	gaTTcg	10	10	2	RSDTLKQ// ARSNLTR
Asn	TAATTA	gaTTcg	10	10	3	RSDTLAI// VSSNLTR
Asn	TAATTA	gaTTcg	10	10	4	RSDTLKV// VRGNLTR
Asn	TAATTA	gaTTcg	10	10	5	RSDTLLV// TSSNLTR
Asn	TAATTA	gaTTcg	10	10	6	RSDTLAV// VRSNLTR
Asn	TAATTA	gaTTcg	10	10	7	RSDTLRT// TKSNLTR
Asn	TAATTA	gaTTcg	10	10	8	RSDTLAT// VSSNLTR
His	TAATTA	gaTTcg	10	10	1	bad read
His	TAATTA	gaTTcg	10	10	2	RSDTLLE// SGGHLTR
His	TAATTA	gaTTcg	10	10	3	RSDTLGT// RVLHLTR
His	TAATTA	gaTTcg	10	10	4	RSDTLAE// SAGHLTR
His	TAATTA	gaTTcg	10	10	5	RSDTLRR// AAAHLTR
His	TAATTA	gaTTcg	10	10	6	RSDTLLE// CARHLTR
His	TAATTA	gaTTcg	10	10	7	RSDTLRT// Mutations
His	TAATTA	gaTTcg	10	10	8	RSDTLRR// GGGHLTR
His	TAATTA	gaTTcg	10	10	9	RSDTLKS// VGSNLTR

Table A-1

1352 lib Asn+3F2/ His+3F2	Homeodomain Binding site	ZFP Binding site	3-AT conc (mM)	IPTG (uM)	Clone #	Recognition helices F1(VNS)// F2(NNW) -1123456//-1123456
His	TAATTA	gaTTcg	10	10	10	RSDTLLS// FSGHLTR
His	TAATTA	gaTTcg	10	10	11	RSDTLER// STCHLTR
His	TAATTA	gaTTcg	10	10	12	mutations
His	TAATTA	gaTTcg	5	10	1	RSDTLLR// TLSHLTR
His	TAATTA	gaTTcg	5	10	2	RSDTLLE// GATHLTR
His	TAATTA	gaTTcg	5	10	3	RSDTLPV// KCGHLTR
His	TAATTA	gaTTcg	5	10	4	RSDTLGP// HYEHLTR

!"#$%&'():*+,-&./&"$$&01(2.34$%,&.#-"+5%3&"/-%6&7)8(,%$%9-+.5,&/6.2&-:%&',5;<10($+#6"6=&"53&-:%&8+,;<10($+#6"6=. Recognition helix sequences (-1, 1, 2, 3, 4, 5, 6) for F1 and F2 for selected 2F-modules are shown for 2F-modules selected for zinc finger binding sites (gaNNcg where NN represents the 2bp-interface). The amino acid at position 3 of F2 can either be Asn or His depending on the zinc finger library used. The selection conditions (3-AT, IPTG and Uracil concentrations) are given. For selection of 2F-modules that bind the G-G interface with the Asn+3F2-library library, a mutant homeodomain binding site (TAAAGG) was used.

Table A-2

List of all 2F-modules in the archive. Recognition helix sequences for F1 and F2 of 2F-modules characterized in this study are shown.

Name	F1: -1123456	F2: -1123456	Note	Target site
2FM-1	RSDTLTQ	QRGNLTR	Selected	GAAACG
2FM-1	RSDTLTQ	QRGNLTR	Selected	GAAATG
2FM-1-QRG	QRGTLTQ	QRGNLTR	Rationally Designed	GAAACA
2FM-1-QRG	QRGTLTQ	QRGNLTR	Rationally Designed	GAAATA
2FM-2	RSDTLVE	QRGNLTR	Selected	GAACCG
2FM-2	RSDTLVE	QRGNLTR	Selected	GAACTG
2FM-2-QRG	QRGTLVE	QRGNLTR	Rationally Designed	GAACCA
2FM-2-QRG	QRGTLVE	QRGNLTR	Rationally Designed	GAACTA
2FM-3	RSDTLQR	QKSNLTR	Selected	GAAGCG
2FM-3	RSDTLQR	QKSNLTR	Selected	GAAGTG
2FM-3-QRG	QRGTLQR	QKSNLTR	Selected	GAAGCA
2FM-3-QRG	QRGTLQR	QKSNLTR	Selected	GAAGTA
2FM-4	RSDTLKG	QRCNLTR	Rationally Designed	GAATCG
2FM-4	RSDTLKG	QRCNLTR	Rationally Designed	GAATTG
2FM-4	RSDTLKG	QRCNLTR	Rationally Designed	GAATAG
2FM-4-QRG	QRGTLKG	QRCNLTR	Rationally Designed	GAATCA
2FM-4-QRG	QRGTLKG	QRCNLTR	Rationally Designed	GAATTA
2FM-4-QRG	QRGTLKG	QRCNLTR	Rationally Designed	GAATAA
2FM-5	RSDTLKQ	DKGNLTR	Rationally Designed	GACACG
2FM-5	RSDTLKQ	DKGNLTR	Rationally Designed	GACATG
2FM-5-QRG	QRGTLKQ	DKGNLTR	Rationally Designed	GACACA
2FM-5-QRG	QRGTLKQ	DKGNLTR	Rationally Designed	GACATA
2FM-6	RSDTLMV	DRSNLTR	Rationally Designed	GACCCG
2FM-6	RSDTLMV	DRSNLTR	Rationally Designed	GACCTG
2FM-6-QRG	QRGTLMV	DRSNLTR	Rationally Designed	GACCCA
2FM-6-QRG	QRGTLMV	DRSNLTR	Rationally Designed	GACCTA
2FM-7	RSDTLER	DRGNLTR	Selected	GACGCG
2FM-7	RSDTLER	DRGNLTR	Selected	GACGTG
2FM-7-QRG	QRGTLER	DRGNLTR	Selected	GACGCA
2FM-7-QRG	QRGTLER	DRGNLTR	Selected	GACGTA
2FM-8	RSDTLKG	DRCNLTR	Selected	GACTTG
2FM-8	RSDTLKG	DRCNLTR	Selected	GACTCG
2FM-8-QRG	QRGTLKG	DRCNLTR	Selected	GACTTA
2FM-8-QRG	QRGTLKG	DRCNLTR	Selected	GACTCA
2FM-9	RSDTLVE	RKRNLTR	Rationally Designed	GAGACG
2FM-9	RSDTLVE	RKRNLTR	Rationally Designed	GAGATG
2FM-9-QRG	QRGTLVE	RKRNLTR	Rationally Designed	GAGACA

2FM-9-QRG	QRGTLVE	RKRNLTR	Rationally Designed	GAGATA
2FM-10	RSDTLKE	RSSNLTR	Selected	GAGCCG
2FM-10	RSDTLKE	RSSNLTR	Selected	GAGCTG
2FM-10-QRG	QRGTLKE	RSSNLTR	Selected	GAGCCA
2FM-10-QRG	QRGTLKE	RSSNLTR	Selected	GAGCTA
2FM-11	RSDTLIR	RAENLTR	Selected	GAGGCG
2FM-11	RSDTLIR	RAENLTR	Selected	GAGGTG
2FM-11-QRG	QRGTLIR	RAENLTR	Selected	GAGGCA
2FM-11-QRG	QRGTLIR	RAENLTR	Selected	GAGGTA
2FM-12	RSDTLKE	KGCNLTR	Selected	GAGTCG
2FM-12	RSDTLKE	KGCNLTR	Selected	GAGTTG
2FM-12-QRG	QRGTLKE	KGCNLTR	Selected	GAGTCA
2FM-12-QRG	QRGTLKE	KGCNLTR	Selected	GAGTTA
2FM-13	RSDTLKQ	AAGNLTR	Selected	GATACG
2FM-13	RSDTLKQ	AAGNLTR	Selected	GATATG
2FM-13-QRG	QRGTLKQ	AAGNLTR	Selected	GATACA
2FM-13-QRG	QRGTLKQ	AAGNLTR	Selected	GATATA
2FM-14	RSDTLLE	LKGHLTR	Rationally Designed	GATCCG
2FM-14	RSDTLLE	LKGHLTR	Rationally Designed	GATCTG
2FM-14-QRG	QRGTLLE	LKGHLTR	Rationally Designed	GATCCA
2FM-14-QRG	QRGTLLE	LKGHLTR	Rationally Designed	GATCTA
2FM-15	RSDTLMR	IRSNLTR	Selected	GATGCG
2FM-15	RSDTLMR	IRSNLTR	Selected	GATGTG
2FM-15-QRG	QRGTLMR	IRSNLTR	Selected	GATGCA
2FM-15-QRG	QRGTLMR	IRSNLTR	Selected	GATGTA
2FM-16	RSDTLRT	TKSNLTR	Selected	GATTCG
2FM-16	RSDTLRT	TKSNLTR	Selected	GATTTG
2FM-16-QRG	QRGTLRT	TKSNLTR	Selected	GATTCA
2FM-17	RSDTLTQ	QRGHLTR	Selected	GGAACG
2FM-17	RSDTLTQ	QRGHLTR	Selected	GGAATG
2FM-17-QRG	QRGTLTQ	QRGHLTR	Selected	GGAACA
2FM-17-QRG	QRGTLTQ	QRGHLTR	Selected	GGAATA
2FM-18	RSDTLRE	QRGHLTR	Selected	GGACCG
2FM-18	RSDTLRE	QRGHLTR	Selected	GGACTG
2FM-18-QRG	QRGTLRE	QRGHLTR	Selected	GGACCA
2FM-18-QRG	QRGTLRE	QRGHLTR	Selected	GGACTA
2FM-19	RSDTLVR	QSGHLTR	Selected	GGAGCG
2FM-19	RSDTLVR	QSGHLTR	Selected	GGAGTG
2FM-19-QRG	QRGTLVR	QSGHLTR	Selected	GGAGCA
2FM-19-QRG	QRGTLVR	QSGHLTR	Selected	GGAGTA
2FM-20	RSDTLKG	QRCHLTR	Rationally Designed	GGATCG
2FM-20	RSDTLKG	QRCHLTR	Rationally Designed	GGATTG
2FM-20	RSDTLKG	QRCHLTR	Rationally Designed	GGATAG

2FM-20-QRG	QRGTLKG	QRCHLTR	Rationally Designed	GGATCA
2FM-20-QRG	QRGTLKG	QRCHLTR	Rationally Designed	GGATTA
2FM-20-QRG	QRGTLKG	QRCHLTR	Rationally Designed	GGATAA
2FM-1014	RSDTLKE	ARRNLTR	Selected	GGCATG
2FM-1014	RSDTLKE	ARRNLTR	Selected	GGCACG
2FM-1014-QRG	QRGTLKE	ARRNLTR	Selected	GGCATA
2FM-1014-QRG	QRGTLKE	ARRNLTR	Selected	GGCACA
2FM-21	RSDTLMV	DRSHLTR	Rationally Designed	GGCCCG
2FM-21	RSDTLMV	DRSHLTR	Rationally Designed	GGCCTG
2FM-21-QRG	QRGTLMV	DRSHLTR	Rationally Designed	GGCCCA
2FM-21-QRG	QRGTLMV	DRSHLTR	Rationally Designed	GGCCTA
2FM-23	RSDTLLR	ESGHLTR	Selected	GGCGCG
2FM-23	RSDTLLR	ESGHLTR	Selected	GGCGTG
2FM-23-QRG	QRGTLLR	ESGHLTR	Selected	GGCGCA
2FM-23-QRG	QRGTLLR	ESGHLTR	Selected	GGCGTA
2FM-24	RSDTLKG	DRCHLTR	Rationally Designed	GGCTCG
2FM-24	RSDTLKG	DRCHLTR	Rationally Designed	GGCTTG
2FM-24-QRG	QRGTLKG	DRCHLTR	Rationally Designed	GGCTCA
2FM-24-QRG	QRGTLKG	DRCHLTR	Rationally Designed	GGCTTA
2FM-25	RSDTLVE	RKRHLTR	Selected	GGGACG
2FM-25	RSDTLVE	RKRHLTR	Selected	GGGATG
2FM-25-QRG	QRGTLVE	RKRHLTR	Selected	GGGACA
2FM-25-QRG	QRGTLVE	RKRHLTR	Selected	GGGATA
2FM-26	RSDTLKE	RSSHLTR	Selected	GGGCCG
2FM-26	RSDTLKE	RSSHLTR	Selected	GGGCTG
2FM-26-QRG	QRGTLKE	RSSHLTR	Selected	GGGCCA
2FM-26-QRG	QRGTLKE	RSSHLTR	Selected	GGGCTA
2FM-27	RSDTLAR	RAEHLTR	Selected	GGGGCG
2FM-27	RSDTLAR	RAEHLTR	Selected	GGGGTG
2FM-27-QRG	QRGTLAR	RAEHLTR	Selected	GGGGCA
2FM-27-QRG	QRGTLAR	RAEHLTR	Selected	GGGGTA
2FM-28	RSDTLLL	RSDHLTR	Selected	GGGTTG
2FM-28-QRG	QRGTLLL	RSDHLTR	Selected	GGGTTA
2FM-29	RSDTLKQ	AAGHLTR	Rationally Designed	GGTACG
2FM-29	RSDTLKQ	AAGHLTR	Rationally Designed	GGTATG
2FM-29-QRG	QRGTLKQ	AAGHLTR	Rationally Designed	GGTACA
2FM-22	RSDTLLE	SGGHLTR	Selected	GGTCCG
2FM-22	RSDTLLE	SGGHLTR	Selected	GGTCTG
2FM-22-QRG	QRGTLLE	SGGHLTR	Selected	GGTCCA
2FM-22-QRG	QRGTLLE	SGGHLTR	Selected	GGTCTA
2FM-31	RSDTLRR	IRFHLTR	Selected	GGTGTG
2FM-31	RSDTLRR	IRFHLTR	Selected	GGTGCG
2FM-31-QRG	QRGTLRR	IRFHLTR	Selected	GGTGTA

2FM-31-QRG	QRGTLRR	IRFHLTR	Selected	GGTGCA
2FM-33	RSDTLTQ	QRGDLTR	Site directed mutagenesis	GCAACG
2FM-33-QRG	QRGTLTQ	QRGDLTR	Site directed mutagenesis	GCAACA
2FM-33	RSDTLTQ	QRGDLTR	Site directed mutagenesis	GCAATG
2FM-33-QRG	QRGTLTQ	QRGDLTR	Site directed mutagenesis	GCAATA
2FM-34	RSDTLTQ	QRGELTR	Site directed mutagenesis	GTAACG
2FM-34-QRG	QRGTLTQ	QRGELTR	Site directed mutagenesis	GTAACA
2FM-34	RSDTLTQ	QRGELTR	Site directed mutagenesis	GTAATG
2FM-34-QRG	QRGTLTQ	QRGELTR	Site directed mutagenesis	GTAATA
2FM-35	RSDTLAE	CARDLTR	Site directed mutagenesis	GCCACG
2FM-35-QRG	QRGTLAE	CARDLTR	Site directed mutagenesis	GCCACA
2FM-36	RSDTLKQ	AAGDLTR	Site directed mutagenesis	GCTACG
2FM-36-QRG	QRGTLKQ	AAGDLTR	Site directed mutagenesis	GCTACA
2FM-36	RSDTLKQ	AAGDLTR	Site directed mutagenesis	GCTATG
2FM-36-QRG	QRGTLKQ	AAGDLTR	Site directed mutagenesis	GCTATA
2FM-37	RSDHLTQ	QRGNLTR	Site directed mutagenesis	GAAAGG
2FM-37-QRG	QRGHLTQ	QRGNLTR	Site directed mutagenesis	GAAAGA
2FM-37	RSDHLTQ	QRGNLTR	Site directed mutagenesis	GAATGG
2FM-37-QRG	QRGHLTQ	QRGNLTR	Site directed mutagenesis	GAATGA
2FM-38	RSDHLKQ	AAGNLTR	Site directed mutagenesis	GATAGG
2FM-38-QRG	QRGHLKQ	AAGNLTR	Site directed mutagenesis	GATAGA
2FM-39	RSDNLKQ	AAGNLTR	Site directed mutagenesis	GATAAG
2FM-39-QRG	QRGNLKQ	AAGNLTR	Site directed mutagenesis	GATAAA
2FM-40	RSDTLVE	RKRSLTR	Site directed mutagenesis	GTGACG
2FM-40-QRG	QRGTLVE	RKRSLTR	Site directed mutagenesis	GTGACA
2FM-40	RSDTLVE	RKRSLTR	Site directed mutagenesis	GTGATG
2FM-40-QRG	QRGTLVE	RKRSLTR	Site directed mutagenesis	GTGATA
2FM-41	RSDNLTQ	QRGNLTR	Site directed	GAAAAG

			mutagenesis	
2FM-41-QRG	QRGNLTQ	QRGNLTR	Site directed mutagenesis	GAAAAA
2FM-42	RSDTLKQ	DKRSLTR	Site directed mutagenesis	GTCACG
2FM-42-QRG	QRGTLKQ	DKRSLTR	Site directed mutagenesis	GTCACA
2FM-48	DKGTLTQ	QRGNLTR	N-terminal Cap	GAAACT
2FM-48	DKGTLTQ	QRGNLTR	N-terminal Cap	GAAATT
2FM-49	DKRTLTQ	QRGNLTR	N-terminal Cap	GAAACC
2FM-49	DKRTLTQ	QRGNLTR	N-terminal Cap	GAAATC
2FM-51	DKGTLVE	RKRHLTR	N-terminal Cap	GGGACT
2FM-52	DKGTLKE	RSSNLTR	N-terminal Cap	GAGCCT
2FM-52	DKGTLKE	RSSNLTR	N-terminal Cap	GAGCTT
2FM-1131	DKGTLVE	QRGNLTR	N-terminal Cap	GAACCT

Table A-3: Primer sequences for ZFN assembly.

Primer Name	Sequence (5' to 3')
F0Fn	CCCAGTCACGACGTTGTAAAACGGTACCAAGCCCTATAAATGTCCTGAATG
F0Rn	ACACGCGTATGGCTTCTCACCGGTGTGCGTA
F1Fn	TGAGAAGCCATACGCGTGTCCTGTCGAGTCCTGT
F1Rn	GCATTGAAACGGTTTTTGCCCTGTGTGAATC
F2Fn	GCAAAAACCGTTTCAATGCCGCATCTGCATG
F2Rn	ACAGGCGAAGGGCTTTTCTCCTGTGTGGGTG
F3Fn	AGAAAAGCCCTTCGCCTGTGACATCTGCGG
F3RnLRGS	AGCGGATAACAATTTCACACAGGATCCACGGAGGTGGATCTTGGTGTG
F3RnTGPGAAGS	AGCGGATAACAATTTCACACAGGATCCTGCAGCACCAGGGCCAGTGTGGATCTTGGTGTG
F1(noF0)Fn	CCCAGTCACGACGTTGTAAAACGGTACCCGCCCATATGCTTGCCC
2FM-F0Fn	CCCAGTCACGACGTTGTAAAACGGTACCAAACCGTATGCTTGCCCTGTC
2FM-F1Rn	GCATTGAAACGGTTTTTGCCCTGTGTGGGTCCTGATGTG
2FM-F1Fn	TGAGAAGCCATACGCGTGTCCTGTCGAGTCCTGTGAC
2FM-F2Rn	ACAGGCGAAGGGCTTTTCTCCTGTGTGGGTCCTGATGTG
2FM-F2Fn	GCAAAAACCGTTTCAATGCCCTGTCGAGTCCTGCGAC
2FM-F3RnLRGS	AGCGGATAACAATTTCACACAGGATCCACGGAGGTGGGTCCTGATGTG
2FM-F3RnTGPGAAGS	AGCGGATAACAATTTCACACAGGATCCTGCAGCACCAGGGCCAGTGTGGGTCCTGATGTG
2FM-F1(noF0)Fn	CCCAGTCACGACGTTGTAAAACGGTACCAAACCGTATGCTTGCCCTG
2FM-F0-QRG(**X**)Fn	CCGTATGCTTGCCCTGTCGAGTCCTGCGACCGCCGCTTCTCCcagcgcggc**NNN**CT
2FM-F1-QRG(**X**)Fn	TGAGAAGCCATACGCGTGTCCTGTCGAGTCCTGTGACCGCCGCTTCTCCcagcgcggc**NNN**CT
2FM-F2-QRG(**X**)Fn	GCAAAAACCGTTTCAATGCCCTGTCGAGTCCTGCGACCGCCGCTTCTCCcagcgcggc**NNN**CT
2FM-F1(noF0)-QRG(**X**)Fn	TTGTAAAACGGTACCAAACCGTATGCTTGCCCTGTCGAGTCCTGCGACCGCCGCTTCTCCcagcgcggc**NNN**CT
2FM-NT-in-Fn	CGTTGTAAAACGGTACCAAACCTTATGCTTGCCCTGTC
2FM-NT-out-Fn	ACGTTGTAAAACGGTACCAAACCT
2FM-CT-out-Rn	AACAATTTCACACAGGATCCACG

NOTE: For QRG(X) primers in place of NNN use ACN if X (F1 position 3) is Thr, use AAY if X (F1 position 3) is Asn, use CAC if X (F1 position 3) is His.

Table A-4:

Gene	Genotyping assay used	Genotyping Forward primer for RFLP analysis or CelI assay (5' to 3')	Genotyping Reverse primer for RFLP analysis or CelI assay (5' to 3')	Forward Primer for Illumina Sequencing (5' to 3')	Reverse Primer for Illumina Sequencing (5' to 3')	Restriction Enzyme site used for Illumina sample preparation	5' Tag for Counting InDels	3' Tag for Counting InDels
dab2ip	RFLP - *SfcI*	CAGGGTACCAC TTCTCCAC	CAGCCTATATGCC CGCAC	CGGCATACGAGCTCTTCC GATCTCCACTTCTCCACC AGCTGC	GCGGTCCAGAGCGGTACCGTCC	*Hpy188I*	GTACCGTCCAT	TCGGAC
hey2	RFLP - *XcmI*	CAGCCCCAGCG TTACAGC	CTGCTGACCGAA GCAGGC	CAAGCAGAAGACGGCAT ACGAGCTCTTCCGATCTC TGCTGACCGAAGCAGGC	CTGCTGACCGAAGCAGGC	*Hpy166II*	AACCAT	GAACTG
rock1	RFLP - *Hpy188I*	GAGATGGTGGA GTCTTTCTC	GTATTGTCTGCAG GGAGTCTC	CAAGCAGAAGACGGCATAC ACGAGCTCTTCCGATCTG AGATGGTGGAGTCTTTCT C	GTATTGTCTGCAGGGAGTCTC	*StyI* HF	CAAGGCCGA	GGCACG
zgc77041	Cel1	GGAGCAAATGT AAGGCAAACC	ATTGTTACATTTT CAAAGATGCTG	GGAGCAAATGTAAGGCA AACC	CAAGCAGAAGACGGCATACGAG CTCTTCCGATCTATTGTTACATT TTCAAAGATGCTG	*SnaBI*	GTACCCAT	CTGGAG
dclk2	RFLP - *AvaI*	GACACGGCGTA CACAAGCC	GAACCAGCGCTA TCACTTAAG	CAAGCAGAAGACGGCAT ACGAGCTCTTCCGATCTG ACACGGCGTACACAAGC C	GGCAGCGGCCGGCTCCC	*NaeI*	GGCTCCCATTCCGG	ACGGGA
mc4r	Cel1	CAGCCTCCTGGA GAACATCC	TCACGGTTGGTCA GGTTGC	CAAGAACCTACATTCCCC TATGAACTTCTTC	CGGCATACGAGCTCTTCCGATCT CATAGAGTCAAACACGTTGTC	*XmnI*	TCTTCATCTGC	GCAGAC
lrp8	RFLP - *MwoI*	GAGGCTGTGAG TATCTGTGC	GAAAGTGTGCAG TATGAGTAAAC	CACTCACCCAAATACACC GGTACCTGCC	CGGCATACGAGCTCTTCCGATCT CCAAATTTTACTCACAACAATG	*Acc65I*	GTACCTGCCCC	CTGGGC
mc3r	Cel1	TTCTTCTCGCCA GACTTCAC	CACCAGTAGAAT GAGGTGGAG	CCCGGCGGCTCCTGGTGC TGGGTACCCAGCTC	CGGCATACGAGCTCTTCCGATCT GCAGAGGCAGAGCGGATG	*Acc65I*	GTACCCAGCTCCAC	GCATGA
apoeb	RFLP - *Hpy188III*	CCACCCAGAAA CTGGGCGC	GGTAAGTGTGGA GCTCTTAAGC	GAAGCTGGAGGAGACAG CCGGGTACCTAC	CGGCATACGAGCTCTTCCGATCT GGTAAGTGTGGAGCTCTTAAGC	*Acc65I*	GTACCTACGC	GGGCCG
lepr	Cel1	AGGTGGACCGG CACACAAC	CACAATTCTTACA AACATCAC	CGGCATACGAGCTCTTCC GATCTGGCGCACCTGTCA ATCTGC	CATTACACCAACAAAAGAGACC AGGTACCTTCC	*Acc65I*	GTACCTTCCAC	GAATTG
irs2	RFLP - *Hpy188III*	GTTCACACTCTT CTAAACTGTG	CCTTTTGAAACCC CCTGGTTG	GTTTTCTCAACGAACAGA GAAAGGTACCATG	CGGCATACGAGCTCTTCCGATCT CCTTTTGAAACCCCCTGGTTG	*Acc65I*	GAATGTACCATGCC	GAAAAG

Table A-4: Sequences of the genotyping primers used for lesions detection in zebrafish embryos. For the analysis by Illumina sequencing the restriction enzymes used for truncating the PCR product near the ZFN site for adaptor ligation are indicated. The unique 5' and 3' tags employed for distinguishing and counting sequences containing InDels for each target site are listed.

Table A-5: Sequences for barcoded adapters

Barcode	Strand 1 Sequence (no phosphorylation)	Strand 2 Sequence (no phosphorylation)
TT	aaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTttT
TG	caAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtgT
TC	gaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtcT
TA	taAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtaT
GT	acAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgtT
GG	ccAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTggT
GC	gcAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgcT
GA	tcAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgaT
CT	agAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTctT
CG	cgAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcgT
CC	ggAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTccT
CA	tgAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcaT
AT	atAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTatT
AG	ctAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTagT
AC	gtAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTacT
AA	ttAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTaaT

Table A-6: List of 2F-modules selected using the B2H system.

Clone ID	9bp Binding Site	F1	F2	F3
R10-3	GAAGAGGCGG	RSDTLAR	RRENLIR	LSSNLTR
R10-5	GAAGAGGCGG	RSDTLAR	RNENLLR	QGPNLSR
R12-1	GAAGGGGCGG	RSDTLAR	RAEHLTN	QHPNLTR
R12-2	GAAGGGGCGG	RSDTLAR	RAEHLTN	QAPNLGR
R13-1	GAAGTGGCGG	RSDTLAR	RRNILQN	LSSNLTR
R15-1	GACGCGGCGG	RSDTLAR	RTDDLKR	DPSNLRR
R15-4	GACGCGGCGG	RSDTLAR	RRDDLTR	EGGNLMR
R16-2	GACGGGGCGG	RSDTLAR	RVDHLHR	GGDNLVR
R17-1	GACGTGGCGG	RSDTLAR	RRQILRN	DPSNLRR
R17-4	GACGTGGCGG	RSDTLAR	RPQILIN	DPSNLRR
R18-4	GAGGAGGCGG	RSDTLAR	RPDNLGR	RHDQLTR
R18-5	GAGGAGGCGG	RSDTLAR	RPDNLGR	RVDNLPR
R19-1	GAGGCGGCGG	RSDTLAR	RRESLVR	RDDHLGR
R19-2	GAGGCGGCGG	RSDTLAR	REDTLTR	RHDQLTR
R20-1	GAGGGGGCGG	RSDTLAR	RKAHLKN	RRDNLLR
R20-4	GAGGGGGCGG	RSDTLAR	RRAHLGN	RQDNLQR
R21-1	GAGGTGGCGG	RSDTLAR	RRQILRN	RRDNLLR
R21-2	GAGGTGGCGG	RSDTLAR	RRSILAN	RGDNLAR
R22-1	GATGAGGCGG	RSDTLAR	RPDNLGR	VVNNLAR
R22-4	GATGAGGCGG	RSDTLAR	RVDNLGR	ISHNLAR
R23-4	GATGCGGCGG	RSDTLAR	RQDDLTR	LSQNLGR

R24-3	GATGGGGCGG	RSDTLAR	RAAHLDN	VTNNLKR
R24-4	GATGGGGCGG	RSDTLAR	RNTHLDN	VTNNLKR
R25-3	GATGTGGCGG	RSDTLAR	RRSILAN	VVSNLRR
R27-2	GCAGCGGCGG	RSDTLAR	RRDDLLR	QGGTLRR
R27-5	GCAGCGGCGG	RSDTLAR	RADSLPR	QGGTLRR
R28-1	GCAGGGGCGG	RSDTLAR	RQEHLVR	QGGTLRR
R28-5	GCAGGGGCGG	RSDTLAR	RREHLAR	QGGTLRR
R29-2	GCAGTGGCGG	RSDTLAR	RREVLMN	QGGTLRR
R29-5	GCAGTGGCGG	RSDTLAR	RSEVLAN	QGGTLRR
R30-1	GCCGAGGCGG	RSDTLAR	RADNLAR	EHRGLKR
R30-2	GCCGAGGCGG	RSDTLAR	RGDNLVR	GRSDLTR
R30-3	GCCGAGGCGG	RSDTLAR	RPDNLGR	DHSNLSR
R34-2	GCGGAGGCGG	RSDTLAR	RRENLKR	RTDSLPR
R34-3	GCGGAGGCGG	RSDTLAR	RQDNLGR	RHQGLHH
R34-4	GCGGAGGCGG	RSDTLAR	RQDNLGR	RREGLGR
R44-3	GCGGCGGCGG	RSDTLAR	RADSLPR	RTDSLPR
R44-4	GCGGCGGCGG	RSDTLAR	RSDDLRR	RTDSLPR
R46-2	GCGGTGGCGG	RSDTLAR	RRQILLN	RPDGLAR
R46-3	GCGGTGGCGG	RSDTLAR	RRNILQN	RLDMLAR
R47-3	GCTGAGGCGG	RSDTLAR	RQDNLGR	VSNTLTR
R47-4	GCTGAGGCGG	RSDTLAR	RQDNLGR	LGHTLNR
R48-4	GCTGCGGCGG	RSDTLAR	RADGLTR	LKHDLGR
R48-5	GCTGCGGCGG	RSDTLAR	RRDDLTR	LGHTLNR
R49-2	GCTGGGGCGG	RSDTLAR	RNDHLTN	VTNSLTR

R49-3	GCTGGGGCGG	RSDTLAR	RSAHLQN	VKNTLTR
R50-1	GCTGTGGCGG	RSDTLAR	RVEVLTN	VRNTLTR
R50-5	GCTGTGGCGG	RSDTLAR	RTEVLAN	VGASLKR
R51-1	GGAGAGGCGG	RSDTLAR	RSDNLGK	QTTHLSR
R51-2	GGAGAGGCGG	RSDTLAR	RPDNLVR	QGGHLAR
R51-3	GGAGAGGCGG	RSDTLAR	RPDNLGR	KKDTLGN
R52-2	GGAGCGGCGG	RSDTLAR	RTDMLAR	QGGHLKR
R52-5	GGAGCGGCGG	RSDTLAR	RRDILLR	QGGHLKR
R54-3	GGAGTGGCGG	RSDTLAR	RREVLMN	QTTHLSR
R54-5	GGAGTGGCGG	RSDTLAR	RREVLVN	QSQHLVR
R55-1	GGCGAGGCGG	RSDTLAR	RQDNLGR	KRVSLNL
R55-5	GGCGAGGCGG	RSDTLAR	RADNLGR	DPSHLPR
R56-1	GGCGCGGCGG	RSDTLAR	RRDDLQR	ETGHLKR
R57-1	GGCGGGGCGG	RSDTLAR	RGEHLTR	ESGHLKR
R58-1	GGCGTGGCGG	RSDTLAR	RADSLPR	ERRGLHR
R58-5	GGCGTGGCGG	RSDTLAR	RRDLLHN	KNISLNH
R59-1	GGGGAGGCGG	RSDTLAR	RTDNLDR	RIDKLGG
R59-3	GGGGAGGCGG	RSDTLAR	RPDNLGR	RVSHLQR
R60-5	GGGGCGGCGG	RSDTLAR	RQDDLTR	RRXGLGR
R61-3	GGGGGGGCGG	RSDTLAR	RREHLTR	RNDKLVP
R62-1	GGGGTGGCGG	RSDTLAR	RREVLMN	RNHGLVR
R62-2	GGGGTGGCGG	RSDTLAR	RREVLEN	RNHGLVR
R63-1	GGTGAGGCGG	RSDTLAR	RLDNLDR	HTHRLVS
R63-5	GGTGAGGCGG	RSDTLAR	RRENLKR	IRHHLKR

R64-4	GGTGCGGCGG	RSDTLAR	RPDDLRR	AGGGLAR
R64-5	GGTGCGGCGG	RSDTLAR	REDGLHR	HTHRLVS
R65-2	GGTGGGGCGG	RSDTLAR	RQEHLVR	HTHRLVS
R66-3	GGTGTGGCGG	RSDTLAR	RVEVLTN	IKHHLGR
R66-4	GGTGTGGCGG	RSDTLAR	RRSILAN	IRHHLKR
R69-3	GTAGGGGCGG	RSDTLAR	RQEHLVR	QHSSLSR
R70-1	GTAGTGGCGG	RSDTLAR	RKQILNN	QGGALQR
R70-4	GTAGTGGCGG	RXDXLAR	RAGILTN	QRGSLGR
R71-1	GTCGAGGCGG	RSDTLAR	RGDNLGR	DLSSLPR
R71-3	GTCGAGGCGG	RSDTLAR	RRENLKR	DQTVLRR
R72-1	GTCGCGGCGG	RSDTLAR	RSDDLRR	ESGALRR
R73-1	GTCGGGGCGG	RSDTLAR	RQEHLVR	EGGALKR
R73-2	GTCGGGGCGG	RSDTLAR	RQEHLVR	DRTPLNR
R74-2	GTCGTGGCGG	RSDTLAR	RTDGLVR	ERRSLGR
R74-4	GTCGTGGCGG	RSDTLAR	RPDNLGR	DRTPLQR
R75-4	GTGGAGGCGG	RSDTLAR	RDDNLQR	RPDALPR
R75-5	GTGGAGGCGG	RSDTLAR	RQDNLGR	RDANLAT
R76-2	GTGGCGGCGG	RSDTLAR	RPDDLRR	RPDALPR
R76-3	GTGGCGGCGG	RSDTLAR	REDTLTR	RGANLNL
R77-1	GTGGGGGCGG	RSDTLAR	RVEHLNN	RMDALMR
R77-3	GTGGGGGCGG	RSDTLAR	RVDHLHR	RGDPLHR
R78-3	GTGGTGGCGG	RSDTLAR	RTEILRN	RHTSLTR
R78-4	GTGGTGGCGG	RSDTLAR	RRDTLRR	RRTILVN
R79-1	**GTTGAGGCGG**	**RSDTLAR**	**RQDNLGR**	**ARHRLIP**

R79-2	GTTGAGGCGG	RSDTLAR	RRENLIR	IRTSLKR
R79-3	GTTGAGGCGG	RSDTLAR	RADNLGR	ARHNLVP
R80-2	GTTGCGGCGG	RSDTLAR	RADSLPR	IRTSLKR
R80-3	GTTGCGGCGG	RSDTLAR	RADTLRK	HHNSLTR
R81-3	GTTGGGGCGG	RSDTLAR	RAEHLTN	INHSLRR
R81-4	GTTGGGGCGG	RSDTLAR	RAAHLDN	VNSSLGR
R82-1	GTTGTGGCGG	RSDTLAR	RRQILSN	HHNSLTR
R82-2	GTTGTGGCGG	RSDTLAR	RRNILQN	HHNSLTR

Figure A-1

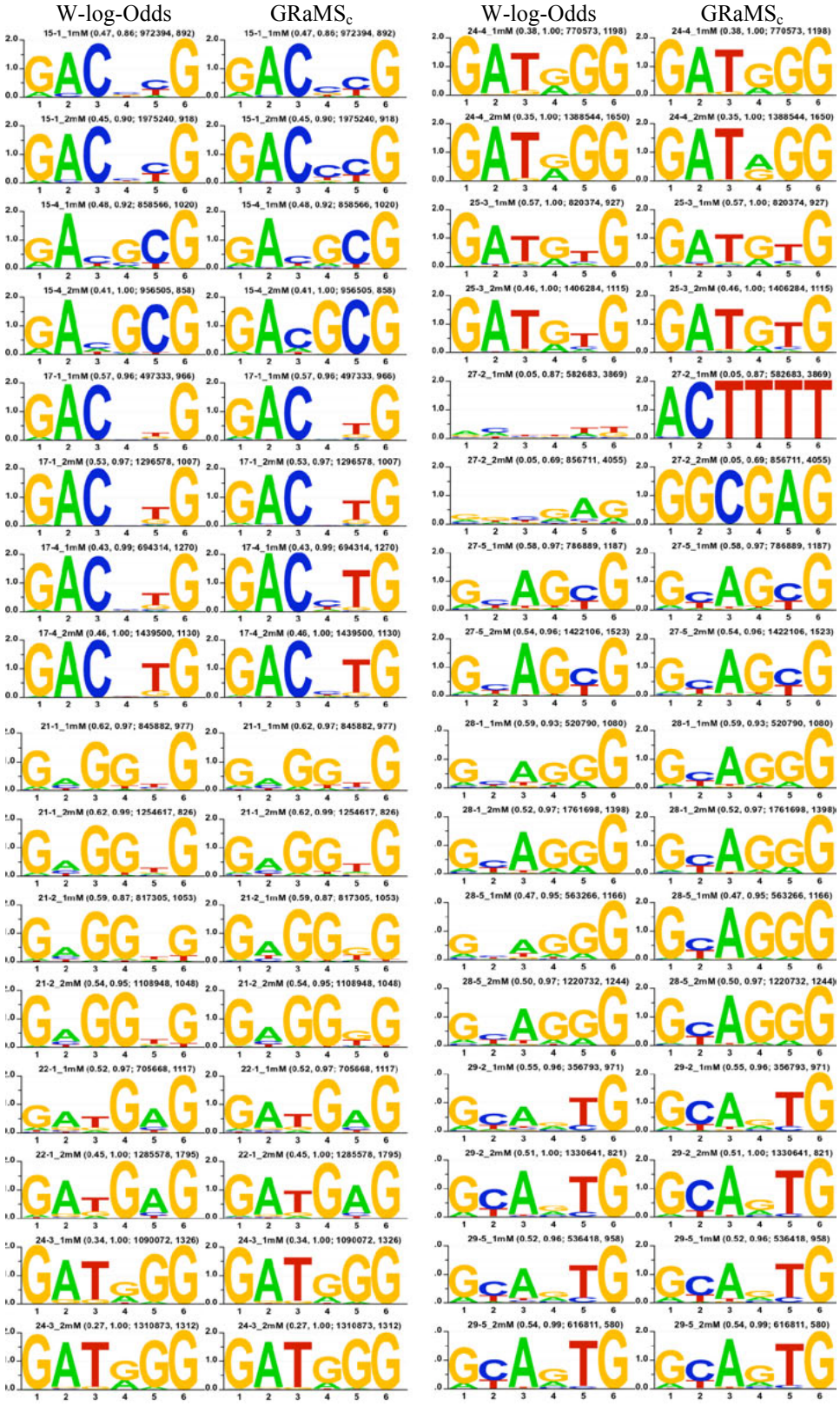


Figure A-1 contd.

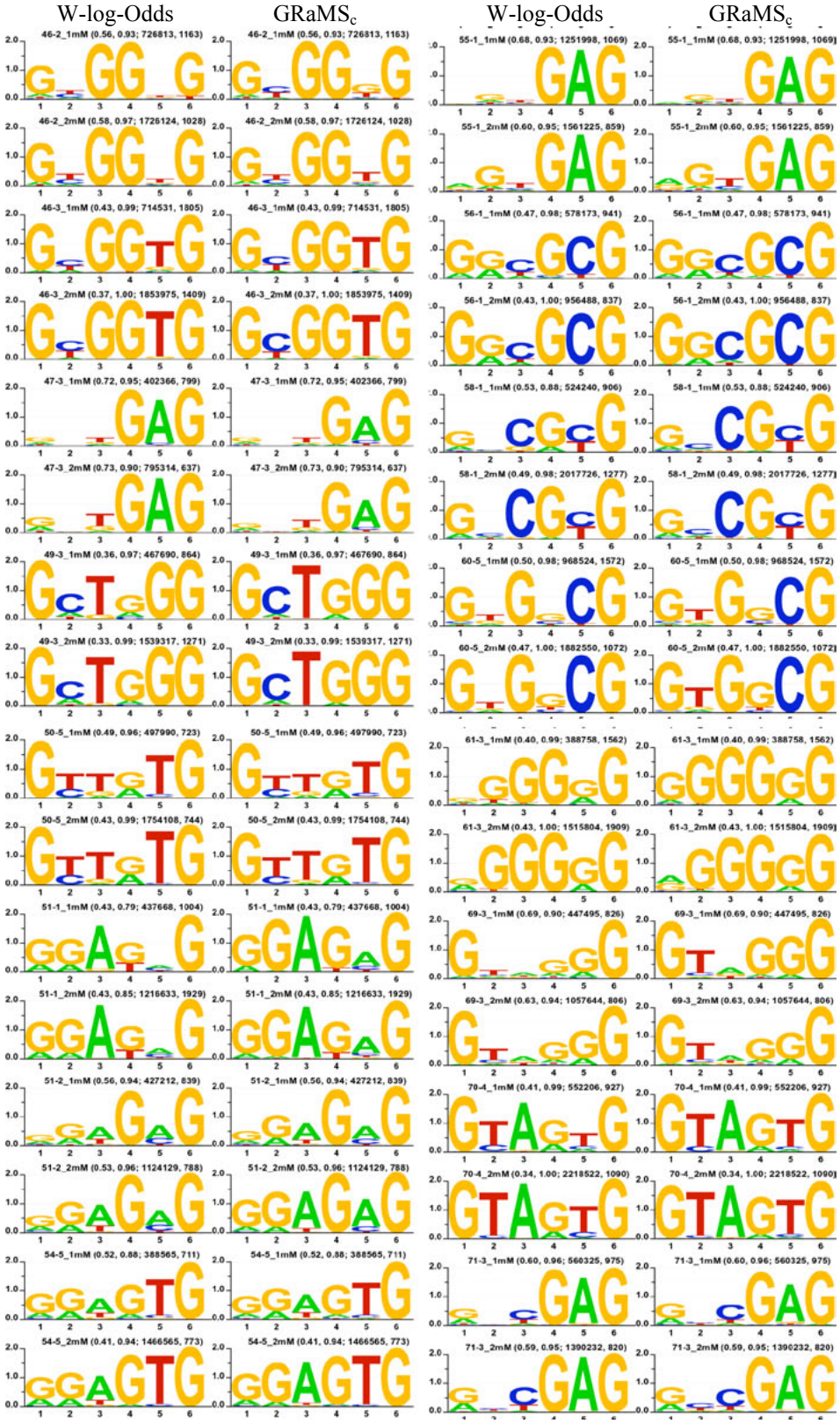


Figure A-1 contd.

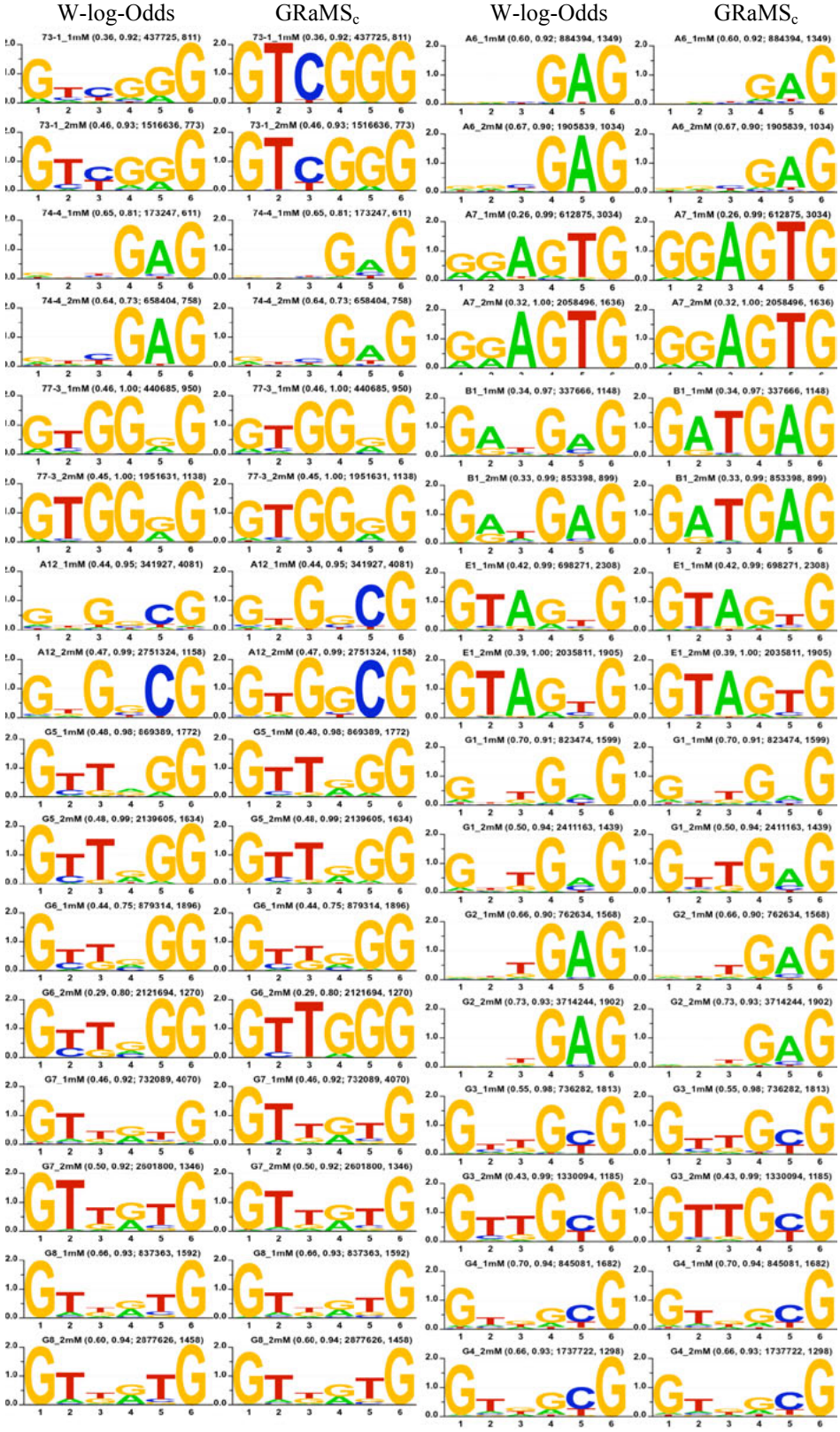


Figure A-1 contd.

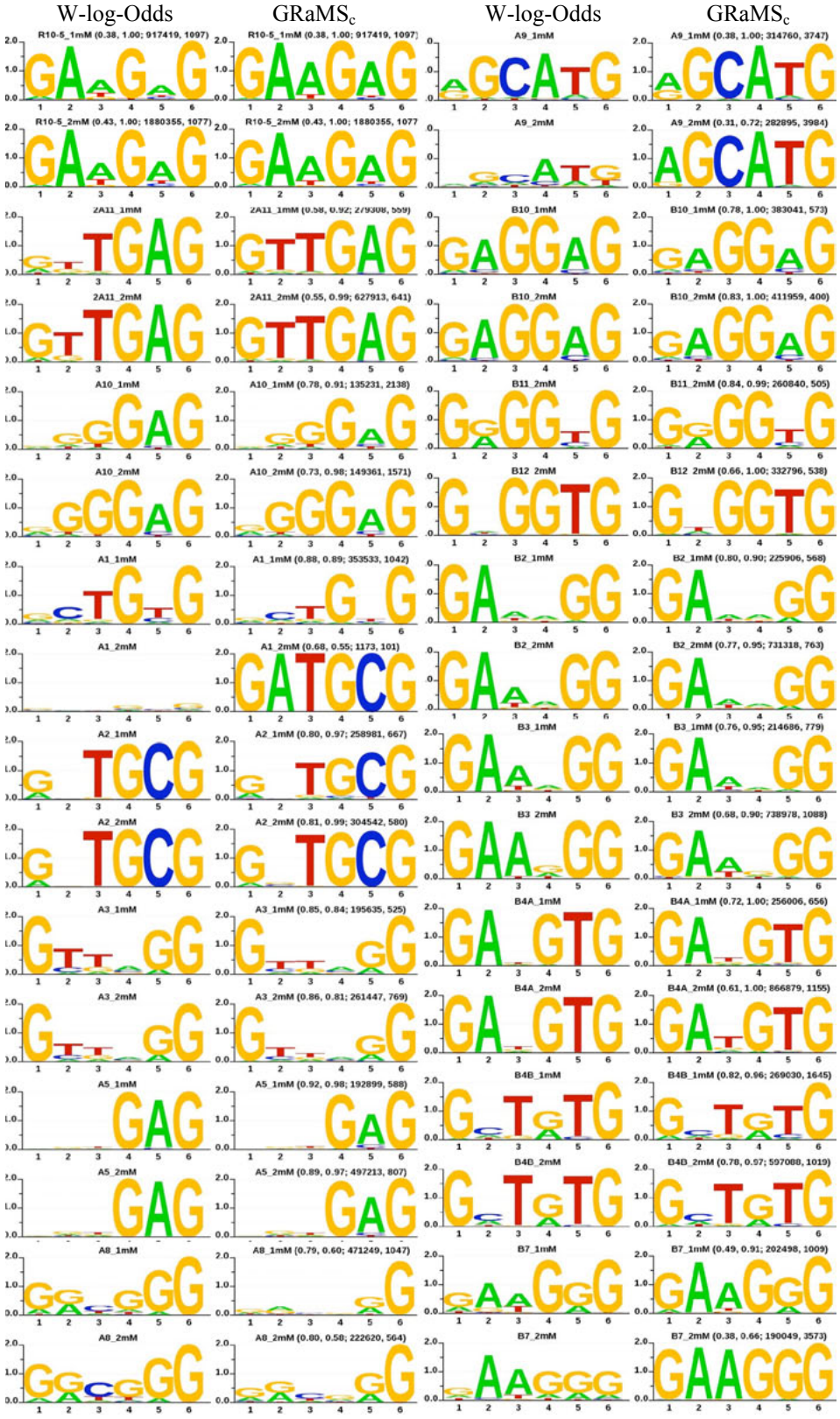


Figure A-1 contd.

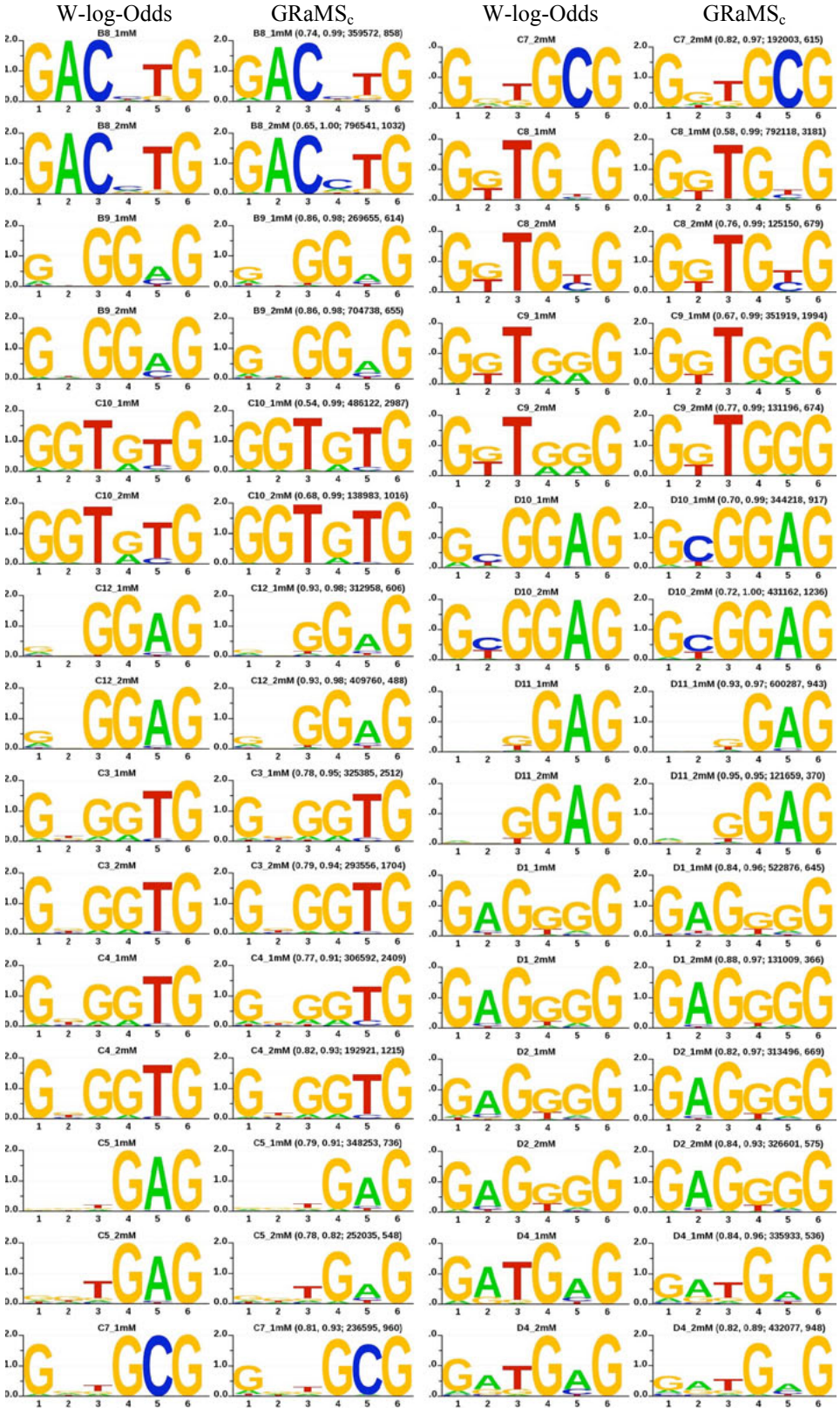


Figure A-1 contd.

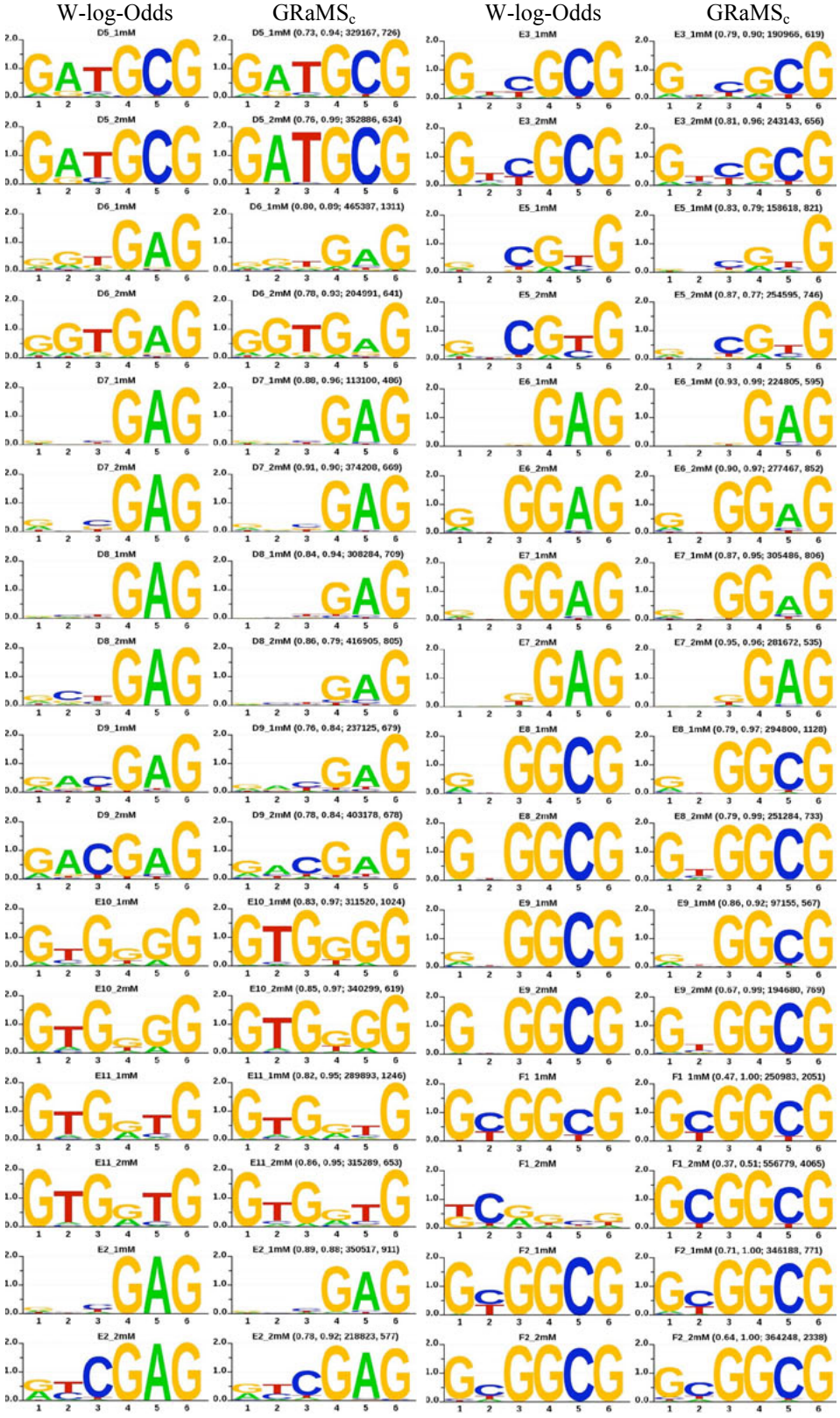


Figure A-1 contd.

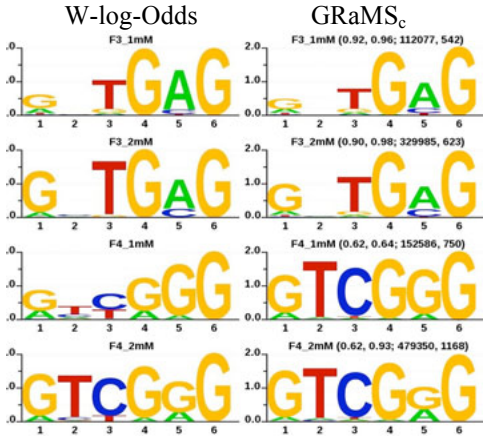


Figure A-1: Binding site specificities of B2H-selected 2F-modules. The binding site specificities of B2H-selected 2F-modules were determined using the CV-B1H system.

REFERENCES

1 Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717-728 (1998).

2 Thomas, M. C. & Chiang, C. M. The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology* **41**, 105-178, doi:10.1080/10409230600648736 (2006).

3 Lopez-Bigas, N., De, S. & Teichmann, S. A. Functional protein divergence in the evolution of Homo sapiens. *Genome biology* **9**, R33, doi:10.1186/gb-2008-9-2-r33 (2008).

4 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).

5 Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

6 Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 19428-19433, doi:10.1073/pnas.0709013104 (2007).

7 Hahn, M. W., Han, M. V. & Han, S. G. Gene family evolution across 12 Drosophila genomes. *PLoS genetics* **3**, e197, doi:10.1371/journal.pgen.0030197 (2007).

8 Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195 (2000).

9 Hillier, L. W. *et al.* Genomics in C. elegans: so many genes, such a little worm. *Genome research* **15**, 1651-1660, doi:10.1101/gr.3729105 (2005).

10 Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. & Teichmann, S. A. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* **14**, 283-291, doi:10.1016/j.sbi.2004.05.004 (2004).

11 Reece-Hoyes, J. S. *et al.* A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome biology* **6**, R110, doi:10.1186/gb-2005-6-13-r110 (2005).

12 Nowick, K. & Stubbs, L. Lineage-specific transcription factors and the evolution of gene regulatory networks. *Brief Funct Genomics* **9**, 65-78, doi:10.1093/bfgp/elp056 (2010).

13 van Nimwegen, E. Scaling laws in the functional content of genomes. *Trends in genetics : TIG* **19**, 479-484 (2003).

14 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10**, 252-263, doi:10.1038/nrg2538 (2009).

15 Ellenberger, T., Fass, D., Arnaud, M. & Harrison, S. C. Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. *Genes Dev* **8**, 970-980 (1994).

16 Ma, P. C., Rould, M. A., Weintraub, H. & Pabo, C. O. Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell* **77**, 451-459 (1994).
17 Ledent, V. & Vervoort, M. The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis. *Genome research* **11**, 754-770, doi:10.1101/gr.177001 (2001).
18 Lewis, E. B. A gene complex controlling segmentation in Drosophila. *Nature* **276**, 565-570 (1978).
19 Noyes, M. B. *et al.* Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**, 1277-1289, doi:10.1016/j.cell.2008.05.023 (2008).
20 Wilson, D. S. & Desplan, C. Structural basis of Hox specificity. *Nat Struct Biol* **6**, 297-300, doi:10.1038/7524 (1999).
21 Nowick, K., Hamilton, A. T., Zhang, H. & Stubbs, L. Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. *Mol Biol Evol* **27**, 2606-2617, doi:10.1093/molbev/msq157 (2010).
22 Emerson, R. O. & Thomas, J. H. Adaptive evolution in zinc finger transcription factors. *PLoS genetics* **5**, e1000325, doi:10.1371/journal.pgen.1000325 (2009).
23 Nolte, R. T., Conlin, R. M., Harrison, S. C. & Brown, R. S. Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 2938-2943 (1998).
24 Renda, M. *et al.* Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *The Journal of biological chemistry* **282**, 33336-33345, doi:10.1074/jbc.M706213200 (2007).
25 Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in genetics : TIG* **17**, 520-527 (2001).
26 Iuchi, S. Three classes of C2H2 zinc finger proteins. *Cellular and molecular life sciences : CMLS* **58**, 625-635 (2001).
27 Iuchi, S. K., N (Eds.). *Zinc Finger Proteins: From Atomic Contact to Cellular Function*. (Landes Biosciences, 2005).
28 Brayer, K. J. & Segal, D. J. Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem Biophys* **50**, 111-131, doi:10.1007/s12013-008-9008-5 (2008).
29 Miller, J., McLachlan, A. D. & Klug, A. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J* **4**, 1609-1614 (1985).

30 Collins, T., Stone, J. R. & Williams, A. J. All in the family: the BTB/POZ, KRAB, and SCAN domains. *Molecular and cellular biology* **21**, 3609-3615, doi:10.1128/MCB.21.11.3609-3615.2001 (2001).
31 Margolin, J. F. *et al.* Kruppel-associated boxes are potent transcriptional repression domains. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 4509-4513 (1994).
32 Cook, T., Gebelein, B. & Urrutia, R. Sp1 and its likes: biochemical and functional predictions for a growing family of zinc finger transcription factors. *Annals of the New York Academy of Sciences* **880**, 94-102 (1999).
33 Liang, H. L. *et al.* The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature* **456**, 400-403, doi:10.1038/nature07388 (2008).
34 Dynan, W. S. & Tjian, R. Isolation of transcription factors that discriminate between different promoters recognized by RNA polymerase II. *Cell* **32**, 669-680 (1983).
35 Dynan, W. S. & Tjian, R. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35**, 79-87 (1983).
36 Briggs, M. R., Kadonaga, J. T., Bell, S. P. & Tjian, R. Purification and biochemical characterization of the promoter-specific transcription factor, Sp1. *Science* **234**, 47-52 (1986).
37 Tan, N. Y. & Khachigian, L. M. Sp1 phosphorylation and its regulation of gene transcription. *Molecular and cellular biology* **29**, 2483-2488, doi:10.1128/MCB.01828-08 (2009).
38 Adamson, E. D. & Mercola, D. Egr1 transcription factor: multiple roles in prostate tumor cell growth and survival. *Tumour Biol* **23**, 93-102 (2002).
39 Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211, doi:10.1016/j.cell.2009.06.001 (2009).
40 Martin, D. *et al.* Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nature structural & molecular biology* **18**, 708-714, doi:10.1038/nsmb.2059 (2011).
41 Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**, 836-840, doi:10.1126/science.1183439 (2010).
42 Myers, S. *et al.* Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**, 876-879, doi:10.1126/science.1182363 (2010).
43 Hanas, J. S., Hazuda, D. J., Bogenhagen, D. F., Wu, F. Y. & Wu, C. W. Xenopus transcription factor A requires zinc for binding to the 5 S RNA gene. *The Journal of biological chemistry* **258**, 14120-14125 (1983).
44 Ginsberg, A. M., King, B. O. & Roeder, R. G. Xenopus 5S gene transcription factor, TFIIIA: characterization of a cDNA clone and measurement of RNA levels throughout development. *Cell* **39**, 479-489 (1984).

45 Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**, 809-817 (1991).
46 Vincent, A., Colot, H. V. & Rosbash, M. Sequence and structure of the serendipity locus of Drosophila melanogaster. A densely transcribed region including a blastoderm-specific gene. *Journal of molecular biology* **186**, 149-166 (1985).
47 Rosenberg, A. S., Mizuochi, T. & Singer, A. Analysis of T-cell subsets in rejection of Kb mutant skin allografts differing at class I MHC. *Nature* **322**, 829-831, doi:10.1038/322829a0 (1986).
48 Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* **29**, 183-212, doi:10.1146/annurev.biophys.29.1.183 (2000).
49 Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A. & Wright, P. E. Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science* **245**, 635-637 (1989).
50 Elrod-Erickson, M., Benson, T. E. & Pabo, C. O. High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure* **6**, 451-464 (1998).
51 Christy, B. & Nathans, D. DNA binding site of the growth factor-inducible protein Zif268. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 8737-8741 (1989).
52 Wolfe, S. A., Grant, R. A., Elrod-Erickson, M. & Pabo, C. O. Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. *Structure* **9**, 717-723 (2001).
53 Zhu, C. *et al.* Evaluation and application of modularly assembled zinc-finger nucleases in zebrafish. *Development* **138**, 4555-4564, doi:10.1242/dev.066779 (2011).
54 Cook, W. J. *et al.* Mutations in the zinc-finger region of the yeast regulatory protein ADR1 affect both DNA binding and transcriptional activation. *The Journal of biological chemistry* **269**, 9374-9379 (1994).
55 Choo, Y. & Klug, A. A role in DNA binding for the linker sequences of the first three zinc fingers of TFIIIA. *Nucleic acids research* **21**, 3341-3346 (1993).
56 Jantz, D. & Berg, J. M. Reduction in DNA-binding affinity of Cys2His2 zinc finger proteins by linker phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 7589-7593, doi:10.1073/pnas.0402191101 (2004).
57 Rizkallah, R., Alexander, K. E. & Hurt, M. M. Global mitotic phosphorylation of C2H2 zinc finger protein linker peptides. *Cell Cycle* **10**, 3327-3336, doi:10.4161/cc.10.19.17619 (2011).
58 Klug, A. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu Rev Biochem* **79**, 213-231, doi:10.1146/annurev-biochem-010909-095056 (2010).

59 Moore, M., Klug, A. & Choo, Y. Improved DNA binding specificity from polyzinc finger peptides by using strings of two-finger units. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 1437-1441, doi:10.1073/pnas.98.4.1437 (2001).

60 Kim, J. S. & Pabo, C. O. Getting a handhold on DNA: design of poly-zinc finger proteins with femtomolar dissociation constants. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 2812-2817 (1998).

61 Moore, M., Choo, Y. & Klug, A. Design of polyzinc finger peptides with structured linkers. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 1432-1436, doi:10.1073/pnas.98.4.1432 (2001).

62 Hockemeyer, D. *et al.* Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature biotechnology* **27**, 851-857, doi:10.1038/nbt.1562 (2009).

63 Li, H. *et al.* In vivo genome editing restores haemostasis in a mouse model of haemophilia. *Nature* **475**, 217-221, doi:10.1038/nature10177 (2011).

64 Handel, E. M., Alwin, S. & Cathomen, T. Expanding or restricting the target site repertoire of zinc-finger nucleases: the inter-domain linker as a major determinant of target site selectivity. *Mol Ther* **17**, 104-111, doi:10.1038/mt.2008.233 (2009).

65 Cui, X. *et al.* Targeted integration in rat and mouse embryos with zinc-finger nucleases. *Nature biotechnology* **29**, 64-67, doi:10.1038/nbt.1731 (2011).

66 Fairall, L., Schwabe, J. W., Chapman, L., Finch, J. T. & Rhodes, D. The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* **366**, 483-487, doi:10.1038/366483a0 (1993).

67 Houbaviy, H. B., Usheva, A., Shenk, T. & Burley, S. K. Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 13577-13582 (1996).

68 Stoll, R. *et al.* Structure of the Wilms tumor suppressor protein zinc finger domain bound to DNA. *Journal of molecular biology* **372**, 1227-1245, doi:10.1016/j.jmb.2007.07.017 (2007).

69 Pavletich, N. P. & Pabo, C. O. Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science* **261**, 1701-1707 (1993).

70 Desjarlais, J. R. & Berg, J. M. Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 7345-7349 (1992).

71 Desjarlais, J. R. & Berg, J. M. Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 2256-2260 (1993).

72 Thukral, S. K., Morrison, M. L. & Young, E. T. Mutations in the zinc fingers of ADR1 that change the specificity of DNA binding and transactivation. *Molecular and cellular biology* **12**, 2784-2792 (1992).
73 Choo, Y., Sanchez-Garcia, I. & Klug, A. In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature* **372**, 642-645, doi:10.1038/372642a0 (1994).
74 Liu, Q., Segal, D. J., Ghiara, J. B. & Barbas, C. F., 3rd. Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 5525-5530 (1997).
75 Segal, D. J., Dreier, B., Beerli, R. R. & Barbas, C. F., 3rd. Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2758-2763 (1999).
76 Jamieson, A. C., Kim, S. H. & Wells, J. A. In vitro selection of zinc fingers with altered DNA-binding specificity. *Biochemistry* **33**, 5689-5695 (1994).
77 Jamieson, A. C., Wang, H. & Kim, S. H. A zinc finger directory for high-affinity DNA recognition. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 12834-12839 (1996).
78 Choo, Y. & Klug, A. Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 11163-11167 (1994).
79 Rebar, E. J. & Pabo, C. O. Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **263**, 671-673 (1994).
80 Wu, H., Yang, W. P. & Barbas, C. F., 3rd. Building zinc fingers by selection: toward a therapeutic application. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 344-348 (1995).
81 Greisman, H. A. & Pabo, C. O. A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* **275**, 657-661 (1997).
82 Dreier, B., Beerli, R. R., Segal, D. J., Flippin, J. D. & Barbas, C. F., 3rd. 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. *The Journal of biological chemistry* **276**, 29466-29478, doi:10.1074/jbc.M102604200 (2001).
83 Dreier, B. *et al.* 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. *The Journal of biological chemistry* **280**, 35588-35597, doi:10.1074/jbc.M506654200 (2005).
84 Isalan, M., Klug, A. & Choo, Y. Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry* **37**, 12026-12033, doi:10.1021/bi981358z (1998).

85 Isalan, M., Klug, A. & Choo, Y. A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nature biotechnology* **19**, 656-660, doi:10.1038/90264 (2001).
86 Wolfe, S. A., Greisman, H. A., Ramm, E. I. & Pabo, C. O. Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *Journal of molecular biology* **285**, 1917-1934, doi:10.1006/jmbi.1998.2421 (1999).
87 Cheng, X., Boyer, J. L. & Juliano, R. L. Selection of peptides that functionally replace a zinc finger in the Sp1 transcription factor by using a yeast combinatorial library. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 14120-14125 (1997).
88 Bae, K. H. *et al.* Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nature biotechnology* **21**, 275-280, doi:10.1038/nbt796 (2003).
89 Meng, X., Brodsky, M. H. & Wolfe, S. A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature biotechnology* **23**, 988-994, doi:10.1038/nbt1120 (2005).
90 Meng, X. & Wolfe, S. A. Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nature protocols* **1**, 30-45, doi:10.1038/nprot.2006.6 (2006).
91 Meng, X., Smith, R. M., Giesecke, A. V., Joung, J. K. & Wolfe, S. A. Counter-selectable marker for bacterial-based interaction trap systems. *Biotechniques* **40**, 179-184 (2006).
92 Noyes, M. B. *et al.* A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic acids research* **36**, 2547-2560, doi:10.1093/nar/gkn048 (2008).
93 Meng, X., Thibodeau-Beganny, S., Jiang, T., Joung, J. K. & Wolfe, S. A. Profiling the DNA-binding specificities of engineered Cys2His2 zinc finger domains using a rapid cell-based method. *Nucleic acids research* **35**, e81, doi:10.1093/nar/gkm385 (2007).
94 Meng, X., Noyes, M. B., Zhu, L. J., Lawson, N. D. & Wolfe, S. A. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nature biotechnology* **26**, 695-701, doi:10.1038/nbt1398 (2008).
95 Joung, J. K., Ramm, E. I. & Pabo, C. O. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 7382-7387, doi:10.1073/pnas.110149297 (2000).
96 Foley, J. E. *et al.* Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN). *PloS one* **4**, e4348, doi:10.1371/journal.pone.0004348 (2009).
97 Hurt, J. A., Thibodeau, S. A., Hirsh, A. S., Pabo, C. O. & Joung, J. K. Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based

selection. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12271-12276, doi:10.1073/pnas.2135381100 (2003).
98 Maeder, M. L. *et al.* Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Molecular cell* **31**, 294-301, doi:10.1016/j.molcel.2008.06.016 (2008).
99 Maeder, M. L., Thibodeau-Beganny, S., Sander, J. D., Voytas, D. F. & Joung, J. K. Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nature protocols* **4**, 1471-1501, doi:10.1038/nprot.2009.98 (2009).
100 Sander, J. D. *et al.* Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nature methods* **8**, 67-69, doi:10.1038/nmeth.1542 (2011).
101 Beerli, R. R., Segal, D. J., Dreier, B. & Barbas, C. F., 3rd. 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. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14628-14633 (1998).
102 Segal, D. J. *et al.* Evaluation of a modular strategy for the construction of novel polydactyl zinc finger DNA-binding proteins. *Biochemistry* **42**, 2137-2148, doi:10.1021/bi026806o (2003).
103 Ramirez, C. L. *et al.* Unexpected failure rates for modular assembly of engineered zinc fingers. *Nature methods* **5**, 374-375, doi:10.1038/nmeth0508-374 (2008).
104 Sander, J. D., Zaback, P., Joung, J. K., Voytas, D. F. & Dobbs, D. An affinity-based scoring scheme for predicting DNA-binding activities of modularly assembled zinc-finger proteins. *Nucleic acids research* **37**, 506-515, doi:10.1093/nar/gkn962 (2009).
105 Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure* **4**, 1171-1180 (1996).
106 Sander, J. D. *et al.* Predicting success of oligomerized pool engineering (OPEN) for zinc finger target site sequences. *BMC Bioinformatics* **11**, 543, doi:10.1186/1471-2105-11-543 (2010).
107 Stormo, G. D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nature reviews. Genetics* **11**, 751-760, doi:10.1038/nrg2845 (2010).
108 Choo, Y. & Klug, A. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 11168-11172 (1994).
109 Oliphant, A. R., Brandl, C. J. & Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence

oligonucleotides: analysis of yeast GCN4 protein. *Molecular and cellular biology* **9**, 2944-2949 (1989).
110 Blackwell, T. K. & Weintraub, H. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science* **250**, 1104-1110 (1990).
111 Fields, D. S., He, Y., Al-Uzri, A. Y. & Stormo, G. D. Quantitative specificity of the Mnt repressor. *Journal of molecular biology* **271**, 178-194 (1997).
112 Roulet, E. *et al.* High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nature biotechnology* **20**, 831-835, doi:10.1038/nbt718 (2002).
113 Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology* **26**, 808-816, doi:10.1038/nbt1410 (2008).
114 Doyon, Y. *et al.* Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nature biotechnology* **26**, 702-708, doi:10.1038/nbt1409 (2008).
115 Soldner, F. *et al.* Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell* **146**, 318-331, doi:10.1016/j.cell.2011.06.019 (2011).
116 Hauschild, J. *et al.* Efficient generation of a biallelic knockout in pigs using zinc-finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 12013-12017, doi:10.1073/pnas.1106422108 (2011).
117 Zhao, Y., Granas, D. & Stormo, G. D. Inferring binding energies from selected binding sites. *PLoS computational biology* **5**, e1000590, doi:10.1371/journal.pcbi.1000590 (2009).
118 Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**, 1429-1435, doi:10.1038/nbt1246 (2006).
119 Bulyk, M. L., Huang, X., Choo, Y. & Church, G. M. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 7158-7163, doi:10.1073/pnas.111163698 (2001).
120 Cullum, R., Alder, O. & Hoodless, P. A. The next generation: using new sequencing technologies to analyse gene regulation. *Respirology* **16**, 210-222, doi:10.1111/j.1440-1843.2010.01899.x (2011).
121 MacQuarrie, K. L., Fong, A. P., Morse, R. H. & Tapscott, S. J. Genome-wide transcription factor binding: beyond direct target regulation. *Trends in genetics : TIG* **27**, 141-148, doi:10.1016/j.tig.2011.01.001 (2011).
122 Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408-1419, doi:10.1016/j.cell.2011.11.013 (2011).

123 Christensen, R. G. *et al.* A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. *Nucleic acids research* **39**, e83, doi:10.1093/nar/gkr239 (2011).
124 Matthews, B. W. Protein-DNA interaction. No code for recognition. *Nature* **335**, 294-295, doi:10.1038/335294a0 (1988).
125 Choo, Y. & Klug, A. Physical basis of a protein-DNA recognition code. *Curr Opin Struct Biol* **7**, 117-125 (1997).
126 Benos, P. V., Lapedes, A. S. & Stormo, G. D. Probabilistic code for DNA recognition by proteins of the EGR family. *Journal of molecular biology* **323**, 701-727 (2002).
127 Jamieson, A. C., Miller, J. C. & Pabo, C. O. Drug discovery with engineered zinc-finger proteins. *Nat Rev Drug Discov* **2**, 361-368, doi:10.1038/nrd1087 (2003).
128 Sadowski, I., Ma, J., Triezenberg, S. & Ptashne, M. GAL4-VP16 is an unusually potent transcriptional activator. *Nature* **335**, 563-564, doi:10.1038/335563a0 (1988).
129 Ruben, S. M. *et al.* Isolation of a rel-related human cDNA that potentially encodes the 65-kD subunit of NF-kappa B. *Science* **254**, 11 (1991).
130 Thiesen, H. J., Bellefroid, E., Revelant, O. & Martial, J. A. Conserved KRAB protein domain identified upstream from the zinc finger region of Kox 8. *Nucleic acids research* **19**, 3996 (1991).
131 Reynolds, L. *et al.* Repression of the HIV-1 5' LTR promoter and inhibition of HIV-1 replication by using engineered zinc-finger transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 1615-1620, doi:10.1073/pnas.252770699 (2003).
132 Papworth, M. *et al.* Inhibition of herpes simplex virus 1 gene expression by designer zinc-finger transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 1621-1626, doi:10.1073/pnas.252773399 (2003).
133 Ren, D., Collingwood, T. N., Rebar, E. J., Wolffe, A. P. & Camp, H. S. PPARgamma knockdown by engineered transcription factors: exogenous PPARgamma2 but not PPARgamma1 reactivates adipogenesis. *Genes Dev* **16**, 27-32, doi:10.1101/gad.953802 (2002).
134 Zhang, L. *et al.* Synthetic zinc finger transcription factor action at an endogenous chromosomal site. Activation of the human erythropoietin gene. *The Journal of biological chemistry* **275**, 33850-33860, doi:10.1074/jbc.M005341200 (2000).
135 Liu, P. Q. *et al.* Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. Activation of vascular endothelial growth factor A. *The Journal of biological chemistry* **276**, 11323-11334, doi:10.1074/jbc.M011172200 (2001).
136 Laganiere, J. *et al.* An engineered zinc finger protein activator of the endogenous glial cell line-derived neurotrophic factor gene provides

functional neuroprotection in a rat model of Parkinson's disease. *J Neurosci* **30**, 16469-16474, doi:10.1523/JNEUROSCI.2440-10.2010 (2010).

137 Beerli, R. R., Dreier, B. & Barbas, C. F., 3rd. Positive and negative regulation of endogenous genes by designed transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 1495-1500, doi:10.1073/pnas.040552697 (2000).

138 Urnov, F. D., Rebar, E. J., Reik, A. & Pandolfi, P. P. Designed transcription factors as structural, functional and therapeutic probes of chromatin in vivo. Fourth in review series on chromatin dynamics. *EMBO reports* **3**, 610-615, doi:10.1093/embo-reports/kvf140 (2002).

139 Snowden, A. W., Gregory, P. D., Case, C. C. & Pabo, C. O. Gene-specific targeting of H3K9 methylation is sufficient for initiating repression in vivo. *Current biology : CB* **12**, 2159-2166 (2002).

140 Snowden, A. W. *et al.* Repression of vascular endothelial growth factor A in glioblastoma cells using engineered zinc finger transcription factors. *Cancer Res* **63**, 8968-8976 (2003).

141 Rebar, E. J. *et al.* Induction of angiogenesis in a mouse model using engineered transcription factors. *Nat Med* **8**, 1427-1432, doi:10.1038/nm795 (2002).

142 Price, S. A. *et al.* Gene transfer of an engineered transcription factor promoting expression of VEGF-A protects against experimental diabetic neuropathy. *Diabetes* **55**, 1847-1854, doi:10.2337/db05-1060 (2006).

143 Eisenstein, M. Sangamo's lead zinc-finger therapy flops in diabetic neuropathy. *Nature biotechnology* **30**, 121-123, doi:10.1038/nbt0212-121a (2012).

144 Sugisaki, H. & Kanazawa, S. New restriction endonucleases from Flavobacterium okeanokoites (FokI) and Micrococcus luteus (MluI). *Gene* **16**, 73-78 (1981).

145 Li, L., Wu, L. P. & Chandrasegaran, S. Functional domains in Fok I restriction endonuclease. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 4275-4279 (1992).

146 Wah, D. A., Hirsch, J. A., Dorner, L. F., Schildkraut, I. & Aggarwal, A. K. Structure of the multimodular endonuclease FokI bound to DNA. *Nature* **388**, 97-100, doi:10.1038/40446 (1997).

147 Kim, Y. G. & Chandrasegaran, S. Chimeric restriction endonuclease. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 883-887 (1994).

148 Kim, Y. G., Smith, J., Durgesha, M. & Chandrasegaran, S. Chimeric restriction enzyme: Gal4 fusion to FokI cleavage domain. *Biological chemistry* **379**, 489-495 (1998).

149 Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 1156-1160 (1996).

150 Huang, B., Schaeffer, C. J., Li, Q. & Tsai, M. D. Splase: a new class IIS zinc-finger restriction endonuclease with specificity for Sp1 binding sites. *Journal of protein chemistry* **15**, 481-489 (1996).
151 Bitinaite, J., Wah, D. A., Aggarwal, A. K. & Schildkraut, I. FokI dimerization is required for DNA cleavage. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 10570-10575 (1998).
152 Wah, D. A., Bitinaite, J., Schildkraut, I. & Aggarwal, A. K. Structure of FokI has implications for DNA cleavage. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 10564-10569 (1998).
153 Szczepek, M. *et al.* Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nature biotechnology* **25**, 786-793, doi:10.1038/nbt1317 (2007).
154 Miller, J. C. *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* **25**, 778-785, doi:10.1038/nbt1319 (2007).
155 Doyon, Y. *et al.* Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nature methods* **8**, 74-79, doi:10.1038/nmeth.1539 (2011).
156 Guo, J., Gaj, T. & Barbas, C. F., 3rd. Directed evolution of an enhanced and highly efficient FokI cleavage domain for zinc finger nucleases. *Journal of molecular biology* **400**, 96-107, doi:10.1016/j.jmb.2010.04.060 (2010).
157 Ramirez, C. L. *et al.* Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects. *Nucleic acids research*, doi:10.1093/nar/gks179 (2012).
158 Wang, J. *et al.* Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome research*, doi:10.1101/gr.122879.111 (2012).
159 Cornu, T. I. *et al.* DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases. *Mol Ther* **16**, 352-358, doi:10.1038/sj.mt.6300357 (2008).
160 Gupta, A., Meng, X., Zhu, L. J., Lawson, N. D. & Wolfe, S. A. Zinc finger protein-dependent and -independent contributions to the in vivo off-target activity of zinc finger nucleases. *Nucleic acids research* **39**, 381-392, doi:10.1093/nar/gkq787 (2011).
161 Kim, H. J., Lee, H. J., Kim, H., Cho, S. W. & Kim, J. S. Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. *Genome research* **19**, 1279-1288, doi:10.1101/gr.089417.108 (2009).
162 Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nature reviews. Genetics* **11**, 636-646, doi:10.1038/nrg2842 (2010).
163 Smith, J. *et al.* Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. *Nucleic acids research* **28**, 3361-3369 (2000).

164 Orlando, S. J. *et al.* Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. *Nucleic acids research* **38**, e152, doi:10.1093/nar/gkq512 (2010).
165 Chen, F. *et al.* High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nature methods* **8**, 753-755, doi:10.1038/nmeth.1653 (2011).
166 Jasin, M. Genetic manipulation of genomes with rare-cutting endonucleases. *Trends in genetics : TIG* **12**, 224-228 (1996).
167 Bibikova, M., Golic, M., Golic, K. G. & Carroll, D. Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* **161**, 1169-1175 (2002).
168 Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763, doi:10.1126/science.1078395

300/5620/763 [pii] (2003).
169 Zou, J. *et al.* Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell Stem Cell* **5**, 97-110, doi:10.1016/j.stem.2009.05.023 (2009).
170 Lombardo, A. *et al.* Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nature biotechnology* **25**, 1298-1306, doi:10.1038/nbt1353 (2007).
171 Urnov, F. D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**, 646-651, doi:10.1038/nature03556 (2005).
172 Lawson, N. D. & Wolfe, S. A. Forward and reverse genetic approaches for the analysis of vertebrate development in the zebrafish. *Dev Cell* **21**, 48-64, doi:10.1016/j.devcel.2011.06.007 (2011).
173 Wood, A. J. *et al.* Targeted genome editing across species using ZFNs and TALENs. *Science* **333**, 307, doi:10.1126/science.1207773 (2011).
174 Geurts, A. M. *et al.* Knockout rats via embryo microinjection of zinc-finger nucleases. *Science* **325**, 433, doi:10.1126/science.1172447 (2009).
175 Carbery, I. D. *et al.* Targeted genome modification in mice using zinc-finger nucleases. *Genetics* **186**, 451-459, doi:10.1534/genetics.110.117002 (2010).
176 Yang, D. *et al.* Generation of PPARgamma mono-allelic knockout pigs via zinc-finger nucleases and nuclear transfer cloning. *Cell research* **21**, 979-982, doi:10.1038/cr.2011.70 (2011).
177 Zhang, F. *et al.* High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12028-12033, doi:10.1073/pnas.0914991107 (2010).
178 Shukla, V. K. *et al.* Precise genome modification in the crop species Zea mays using zinc-finger nucleases. *Nature* **459**, 437-441, doi:10.1038/nature07992 (2009).

179 Wright, D. A. *et al.* High-frequency homologous recombination in plants mediated by zinc-finger nucleases. *Plant J* **44**, 693-705, doi:10.1111/j.1365-313X.2005.02551.x (2005).
180 Townsend, J. A. *et al.* High-frequency modification of plant genes using engineered zinc-finger nucleases. *Nature* **459**, 442-445, doi:10.1038/nature07845 (2009).
181 Beumer, K. J. *et al.* Efficient gene targeting in Drosophila by direct embryo injection with zinc-finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 19821-19826, doi:10.1073/pnas.0810475105 (2008).
182 Lee, H. J., Kim, E. & Kim, J. S. Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome research* **20**, 81-89, doi:10.1101/gr.099747.109 (2010).
183 Holt, N. *et al.* Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nature biotechnology* **28**, 839-847, doi:10.1038/nbt.1663 (2010).
184 Pattanayak, V., Ramirez, C. L., Joung, J. K. & Liu, D. R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods* **8**, 765-770, doi:10.1038/nmeth.1670 (2011).
185 Gabriel, R. *et al.* An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nature biotechnology* **29**, 816-823, doi:10.1038/nbt.1948 (2011).
186 van Rensburg, R. *et al.* Chromatin structure of two genomic sites for targeted transgene integration in induced pluripotent stem cells and hematopoietic stem cells. *Gene therapy*, doi:10.1038/gt.2012.25 (2012).
187 Siekmann, A. F., Standley, C., Fogarty, K. E., Wolfe, S. A. & Lawson, N. D. Chemokine signaling guides regional patterning of the first embryonic artery. *Genes Dev* **23**, 2272-2277, doi:10.1101/gad.1813509 (2009).
188 Cifuentes, D. *et al.* A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science* **328**, 1694-1698, doi:10.1126/science.1190809 (2010).
189 Bussmann, J., Wolfe, S. A. & Siekmann, A. F. Arterial-venous network formation during brain vascularization involves hemodynamic regulation of chemokine signaling. *Development* **138**, 1717-1726, doi:10.1242/dev.059881 (2011).
190 Silva, G. *et al.* Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. *Current gene therapy* **11**, 11-27 (2011).
191 Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29**, 143-148, doi:10.1038/nbt.1755 (2011).
192 Sander, J. D. *et al.* Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature biotechnology* **29**, 697-698, doi:10.1038/nbt.1934 (2011).

193 Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research* **39**, e82, doi:10.1093/nar/gkr218 (2011).
194 Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509-1512, doi:10.1126/science.1178811 (2009).
195 Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**, 1501, doi:10.1126/science.1178817 (2009).
196 Cathomen, T. & Joung, J. K. Zinc-finger nucleases: the next generation emerges. *Mol Ther* **16**, 1200-1207, doi:10.1038/mt.2008.114 (2008).
197 Lloyd, A., Plaisier, C. L., Carroll, D. & Drews, G. N. Targeted mutagenesis using zinc-finger nucleases in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 2232-2237, doi:10.1073/pnas.0409339102 (2005).
198 Zhang, F. *et al.* High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America*, doi:0914991107 [pii]

10.1073/pnas.0914991107 (2010).
199 Osakabe, K., Osakabe, Y. & Toki, S. Site-directed mutagenesis in Arabidopsis using custom-designed zinc finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America*, doi:1000234107 [pii]

10.1073/pnas.1000234107 (2010).
200 Mashimo, T. *et al.* Generation of knockout rats with X-linked severe combined immunodeficiency (X-SCID) using zinc-finger nucleases. *PloS one* **5**, e8870, doi:10.1371/journal.pone.0008870 (2010).
201 Alper, J. One-off therapy for HIV. *Nature biotechnology* **27**, 300, doi:10.1038/nbt0409-300 (2009).
202 Carroll, D., Morton, J. J., Beumer, K. J. & Segal, D. J. Design, construction and in vitro testing of zinc finger nucleases. *Nature protocols* **1**, 1329-1341, doi:10.1038/nprot.2006.231 (2006).
203 Bibikova, M. *et al.* Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Molecular and cellular biology* **21**, 289-297, doi:10.1128/MCB.21.1.289-297.2001 (2001).
204 Liu, P. Q. *et al.* Generation of a triple-gene knockout mammalian cell line using engineered zinc-finger nucleases. *Biotechnology and bioengineering*, doi:10.1002/bit.22654 (2009).
205 Alwin, S. *et al.* Custom zinc-finger nucleases for use in human cells. *Mol Ther* **12**, 610-617, doi:10.1016/j.ymthe.2005.06.094 (2005).
206 Pruett-Miller, S. M., Reading, D. W., Porter, S. N. & Porteus, M. H. Attenuation of zinc finger nuclease toxicity by small-molecule regulation of protein levels. *PLoS genetics* **5**, e1000376, doi:10.1371/journal.pgen.1000376 (2009).
207 Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18**, 6097-6100 (1990).

208 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).
209 Benjamini, Y. H., Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B*, 289-300 (1995).
210 Le Provost, F. *et al.* Zinc finger nuclease technology heralds a new era in mammalian transgenesis. *Trends Biotechnol* **28**, 134-141, doi:10.1016/j.tibtech.2009.11.007 (2010).
211 Yan, Z., Sun, X. & Engelhardt, J. F. Progress and prospects: techniques for site-directed mutagenesis in animal models. *Gene therapy* **16**, 581-588, doi:10.1038/gt.2009.16 (2009).
212 Carroll, D. Progress and prospects: zinc-finger nucleases as gene therapy agents. *Gene therapy* **15**, 1463-1468, doi:10.1038/gt.2008.145 (2008).
213 Elrod-Erickson, M. & Pabo, C. O. Binding studies with mutants of Zif268. Contribution of individual side chains to binding affinity and specificity in the Zif268 zinc finger-DNA complex. *The Journal of biological chemistry* **274**, 19281-19285 (1999).
214 Shimizu, Y. *et al.* Adding fingers to an engineered zinc finger nuclease can reduce activity. *Biochemistry* **50**, 5033-5041, doi:10.1021/bi200393g (2011).
215 Westerfield. *The Zebrafish Book*. (University of Oregon Press, 1993).
216 Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **2**, 28-36 (1994).
217 Ihaka, R. a. G., R. R: A Language for Data Analysis and Graphics 1996. *Journal of Computational and Graphical Statistics* **5**, 299-314 (1996).
218 Liu, X., Brutlag, D. L. & Liu, J. S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 127-138 (2001).
219 Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).
220 Moré, J. *The Levenberg-Marquardt algorithm: Implementation and theory.*, (Springer, 1978).
221 Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics*, 431-441 (1963).
222 Levenberg, K. A method for the solution of certain problems in least squares. *Quart. Applied Math.*, 164-168 (1944).
223 Berg, O. G. & von Hippel, P. H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology* **193**, 723-750 (1987).

224 Benos, P. V., Bulyk, M. L. & Stormo, G. D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic acids research* **30**, 4442-4451 (2002).
225 Miller, J. C. & Pabo, C. O. Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *Journal of molecular biology* **313**, 309-315, doi:10.1006/jmbi.2001.4975 (2001).
226 Dreier, B., Segal, D. J. & Barbas, C. F., 3rd. 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, doi:10.1006/jmbi.2000.4133 (2000).
227 Liu, Q., Xia, Z., Zhong, X. & Case, C. C. Validated zinc finger protein designs for all 16 GNN DNA triplet targets. *The Journal of biological chemistry* **277**, 3850-3856, doi:10.1074/jbc.M110669200 (2002).
228 Mandell, J. G. & Barbas, C. F., 3rd. Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. *Nucleic acids research* **34**, W516-523, doi:10.1093/nar/gkl209 (2006).
229 Wright, D. A. *et al.* Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly. *Nature protocols* **1**, 1637-1652, doi:10.1038/nprot.2006.259 (2006).
230 Kim, S., Lee, M. J., Kim, H., Kang, M. & Kim, J. S. Preassembled zinc-finger arrays for rapid construction of ZFNs. *Nature methods* **8**, 7, doi:10.1038/nmeth0111-7a (2011).
231 Ryan, M. P., Jones, R. & Morse, R. H. SWI-SNF complex participation in transcriptional activation at a step subsequent to activator binding. *Molecular and cellular biology* **18**, 1774-1782 (1998).
232 Rebar, E. J., Greisman, H. A. & Pabo, C. O. Phage display methods for selecting zinc finger proteins with novel DNA-binding specificities. *Methods Enzymol* **267**, 129-149 (1996).
233 Liu, J. & Stormo, G. D. Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics* **24**, 1850-1857, doi:10.1093/bioinformatics/btn331 (2008).
234 Persikov, A. V., Osada, R. & Singh, M. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics* **25**, 22-29, doi:10.1093/bioinformatics/btn580 (2009).
235 Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).
236 Pomerantz, J. L., Wolfe, S. A. & Pabo, C. O. Structure-based design of a dimeric zinc finger protein. *Biochemistry* **37**, 965-970, doi:10.1021/bi972464o (1998).
237 Wolfe, S. A., Grant, R. A. & Pabo, C. O. Structure of a designed dimeric zinc finger protein bound to DNA. *Biochemistry* **42**, 13401-13409, doi:10.1021/bi034830b (2003).

238 Romer, P. *et al.* Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. *Science* **318**, 645-648, doi:10.1126/science.1144958 (2007).
239 Mak, A. N., Bradley, P., Cernadas, R. A., Bogdanove, A. J. & Stoddard, B. L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335**, 716-719, doi:10.1126/science.1216211 (2012).
240 Deng, D. *et al.* Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335**, 720-723, doi:10.1126/science.1215670 (2012).
241 Hockemeyer, D. *et al.* Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**, 731-734, doi:10.1038/nbt.1927 (2011).
242 Mahfouz, M. M. *et al.* De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 2623-2628, doi:10.1073/pnas.1019533108 (2011).
243 Fonfara, I., Curth, U., Pingoud, A. & Wende, W. Creating highly specific nucleases by fusion of active restriction endonucleases and catalytically inactive homing endonucleases. *Nucleic acids research* **40**, 847-860, doi:10.1093/nar/gkr788 (2012).
244 Reyon, D. *et al.* FLASH assembly of TALENs for high-throughput genome editing. *Nature biotechnology*, doi:10.1038/nbt.2170 (2012).
245 Lieschke, G. J. & Currie, P. D. Animal models of human disease: zebrafish swim into view. *Nature reviews. Genetics* **8**, 353-367, doi:10.1038/nrg2091 (2007).
246 Huszar, D. *et al.* Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* **88**, 131-141 (1997).
247 Renquist, B. J., Lippert, R. N., Sebag, J. A., Ellacott, K. L. & Cone, R. D. Physiological roles of the melanocortin MC receptor. *European journal of pharmacology* **660**, 13-20, doi:10.1016/j.ejphar.2010.12.025 (2011).
248 Song, Y. & Cone, R. D. Creation of a genetic model of obesity in a teleost. *FASEB J* **21**, 2042-2049, doi:10.1096/fj.06-7503com (2007).
249 Zhang, C., Forlano, P. M. & Cone, R. D. AgRP and POMC neurons are hypophysiotropic and coordinately regulate multiple endocrine axes in a larval teleost. *Cell metabolism* **15**, 256-264, doi:10.1016/j.cmet.2011.12.014 (2012).
250 Lampert, K. P. *et al.* Determination of onset of sexual maturation and mating behavior by melanocortin receptor 4 polymorphisms. *Current biology : CB* **20**, 1729-1734, doi:10.1016/j.cub.2010.08.029 (2010).
251 Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425-432, doi:10.1038/372425a0 (1994).
252 Friedman, J. M. & Halaas, J. L. Leptin and the regulation of body weight in mammals. *Nature* **395**, 763-770, doi:10.1038/27376 (1998).

253 Montague, C. T. *et al.* Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387**, 903-908, doi:10.1038/43185 (1997).
254 Gorissen, M., Bernier, N. J., Nabuurs, S. B., Flik, G. & Huising, M. O. Two divergent leptin paralogues in zebrafish (Danio rerio) that originate early in teleostean evolution. *The Journal of endocrinology* **201**, 329-339, doi:10.1677/JOE-09-0034 (2009).
255 Tartaglia, L. A. *et al.* Identification and expression cloning of a leptin receptor, OB-R. *Cell* **83**, 1263-1271 (1995).
256 Poy, M. N. *et al.* miR-375 maintains normal pancreatic alpha- and beta-cell mass. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 5813-5818, doi:10.1073/pnas.0810550106 (2009).
257 Stoletov, K. *et al.* Vascular lipid accumulation, lipoprotein oxidation, and macrophage lipid uptake in hypercholesterolemic zebrafish. *Circulation research* **104**, 952-960, doi:10.1161/CIRCRESAHA.108.189803 (2009).
258 Flynn, E. J., 3rd, Trent, C. M. & Rawls, J. F. Ontogeny and nutritional control of adipogenesis in zebrafish (Danio rerio). *Journal of lipid research* **50**, 1641-1652, doi:10.1194/jlr.M800590-JLR200 (2009).