

University of Massachusetts Medical School

**eScholarship@UMMS**

---

GSBS Dissertations and Theses

Graduate School of Biomedical Sciences

---

2013-05-24

## Expanding the Known DNA-binding Specificity of Homeodomains for Utility in Customizable Sequence-specific Nucleases: A Dissertation

Stephanie W. Chu

*University of Massachusetts Medical School*

### Let us know how access to this document benefits you.

Follow this and additional works at: [https://escholarship.umassmed.edu/gsbs\\_diss](https://escholarship.umassmed.edu/gsbs_diss)



Part of the [Biochemistry Commons](#), and the [Molecular Genetics Commons](#)

---

#### Repository Citation

Chu SW. (2013). Expanding the Known DNA-binding Specificity of Homeodomains for Utility in Customizable Sequence-specific Nucleases: A Dissertation. GSBS Dissertations and Theses. <https://doi.org/10.13028/M2DW3Z>. Retrieved from [https://escholarship.umassmed.edu/gsbs\\_diss/684](https://escholarship.umassmed.edu/gsbs_diss/684)

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in GSBS Dissertations and Theses by an authorized administrator of eScholarship@UMMS. For more information, please contact [Lisa.Palmer@umassmed.edu](mailto:Lisa.Palmer@umassmed.edu).

*Graduate School of Biomedical Sciences*

*GSBS Dissertations*

*University of Massachusetts Medical School Year 2013*

EXPANDING THE KNOWN DNA-BINDING SPECIFICITY OF  
HOMEODOMAINS FOR UTILITY IN CUSTOMIZABLE  
SEQUENCE-SPECIFIC NUCLEASES

Stephanie Chu

University of Massachusetts Medical School

EXPANDING THE KNOWN DNA-BINDING SPECIFICITY OF  
HOMEODOMAINS FOR UTILITY IN CUSTOMIZABLE  
SEQUENCE-SPECIFIC NUCLEASES

A Dissertation Presented

by

Stephanie Chu

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

May 24<sup>th</sup>, 2013

Department of Biochemistry and Molecular Pharmacology

Program in Gene Function and Expression

# EXPANDING THE KNOWN DNA-BINDING SPECIFICITY OF HOMEODOMAINS FOR UTILITY IN CUSTOMIZABLE SEQUENCE-SPECIFIC NUCLEASES

A Dissertation Presented By  
Stephanie Chu

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

Ernest Fraenkel, Ph.D., Member of Committee

Dan Bolon, Ph.D., Member of Committee

Sean Ryder, Ph.D., Member of Committee

Marian Walhout, 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

Charles Sagerstrom, 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  
May 24<sup>th</sup>, 2013

## **Acknowledgement**

My advisor, Scot Wolfe, for challenging me and believing in me to be the scientist I am today.

My thesis committee members, Charles Sagerstrom, Dan Bolon, Sean Ryder, and Marian Walhout for attending my committee meetings with helpful advice and comments.

Ernest Fraenkel for coming to the defense.

Members of the Wolfe lab, past and present.

Members of the Ryder lab, Lawson lab, and Brodsky lab.

Friends that I have met while at UMass: Ankit, Sarah, Victoria, and Dave.

Good friends that have watch each other grow and support each other over the past 8 years: Jadyn, Jen, and Leah.

Keri for still knowing me the best, and who never ceases to amaze me.

Aaron W. Reinke, whom I could not done this without.

## **Abstract**

Homeodomains (HDs) are a large family of DNA-binding domains contained in transcription factors that are most notable for regulating body development and patterning in metazoans. HDs consist of three alpha helices preceded by an N-terminal arm, where the third helix (the recognition helix) and the N-terminal arm are responsible for defining DNA-binding specificity. Here we attempted to engineer the HDs by fully randomizing positions in the recognition helix to specify each of the 64 possible 3' triplet sites (i.e. TAANNN). We recovered HD variants that preferentially recognize or are compatible with 44 of the possible sites, a dramatic increase from the previously observed range of specificities. Many of these HD variants contain combinations of novel specificity determinants that are uncommon or absent in extant HDs, where these determinants can be grafted into alternate HD backbones with an accompanying alteration in their specificity. The identified determinates expand our understanding of HD recognition, allowing for the creation of more explicit recognition models for this family. Additionally, we demonstrate that HDs can recognize a broader range of DNA sequences than anticipated, thus raising questions about the fitness barrier that restricts the evolution HD-DNA recognition in nature. Finally, these new HD variants have utility as DNA-binding domains to direct targeting of customizable sequence-specific nuclease as demonstrated by site-specific lesions created in zebrafish. Thus HDs can guide sequence-specific enzymatic function precisely and predictably within a complex genome when used in engineered artificial enzymes.

# Table of Contents

ABSTRACT	v
TABLE OF CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF ABBREVIATIONS	xii
PREFACE	xiv
<b>CHAPTER I</b>	<b>INTRODUCTION TO HOMEODOMAINS</b>
	1
Molecular Biology Perpetuates Life	2
DNA-binding Domains	3
History and Biology of Homeodomains	4
General Homeodomains Characteristic	6
Molecular Interaction Between the Homeodomain and DNA	7
Functional Residues for DNA-binding Specificity	11
Exploring the Recognition Potential of Homedomains	12
Summary	17
<b>CHAPTER II</b>	<b>18</b>
<b>EXPLORING THE DNA-RECOGNITION POTENTIAL OF HOMEODOMAINS</b>	
Chapter II has been publish previously as:	
Stephanie W. Chu, Marcus B. Noyes, Ryan G. Christensen, Brian G. Pierce, Lihua J. Zhu, Zhiping Weng, Gary D. Stormo, and Scot A. Wolfe (2012). Exploring the DNA-recognition potential of homeodomains. Genome Research 22, 1889-1898	
Introduction	19

Results	22
Discussion	52
Material and Methods	56
<b>CHAPTER III            INTRODUCTION TO GENOMIC TARGETING</b>	<b>64</b>
Advancing biology, biotechnology, and medicine through targeted genome editing and targeted gene regulation	65
Tools to target specific genomic sites	68
Targeted gene regulation by artificial transcription factors	72
Genome editing by customizable sequence-directed endonucleases	73
Previous gene targeting utilizing homeodomains	79
Summary	79
<b>CHAPTER IV</b>	<b>81</b>
<b>UTILIZING ENGINEERED HOMEODOMAINS IN CUSTOMIZABLE SEQUENCE- SPECIFIC NUCLEASES</b>	
Introduction	82
Results	83
Discussion	99
Material and Methods	102
<b>CHAPTER V            DESCRIPTION OF METHODS USED</b>	<b>105</b>
Bacterial-One Hybrid System For Selections	106
Identifying Target Sites or DBDs From the B1H System	107
Mutual Information Analysis of Amino Acid-Base Interactions	109
Electrophoretic Mobility Shift Assays & Competition Binding Assays to	



Determine Equilibrium Dissociation Constants	110
Superimposition to Estimate Distance Spanning Two DBDs	111
Bacterial-One Hybrid Activity Assay	112
Yeast-Based Nuclease Assay	112
Nuclease Treatment of Zebrafish	113
LacZalpha Blue-White Assay for Lesion Identification	115
<b>CHAPTER VI            GENERAL DISCUSSION</b>	116
Expanding Homeodomain Sequence Specificity	117
Limitations of Homeodomain Recognition Potential	119
Evolutionary Implications of Expanding Homeodomain Specificity	120
Future Directions for Broadening Homeodomain Specificity	122
Homeodomains in Artificial Nucleases	123
Limitations of Homeodomains in Artificial Nuclease	124
Future Direction of Homeodomains in Artificial Nuclease	126
Overall Utility of Homeodomains as DNA-binding Domains	127
<b>APPENDIX</b>	
<b>REFERENCES</b>	

## **List of Tables**

Table 2-1	Mutual Information analysis of the selected homeodomain-binding site combinations
Table 2-2	Equilibrium dissociation constants of homeodomain variants
Table 3-1	The advantages and disadvantage of the different types of customizable sequence-directed endonucleases.
Table 4-1	Linkers identified between the ZF and HD from B1H selections
Table 4-2	Sequences tested in the yeast reporter assay
Table 4-3	nZFHD constructs used for zebrafish.

## List of Figures

- Figure 1-1 Cartoon representation of the En HD-DNA complex
- Figure 1-2 Cartoons of the multiple En variant-DNA structures show different possible interactions between the HD and DNA
- Figure 1-3 Previously published chart that catalog HD specificity determinants
- Figure 1-4 Previously published clustering of sequence specificity groups based on fly HD sequence specificity appears limited
- Figure 2-1 Structure of the *engrailed* HD and distribution of HD recognition residues
- Figure 2-2 The B1H selection system
- Figure 2-3 Stringency used to select HDs for different target sites.
- Figure 2-4 Logos representing the sequences of the recovered HDs from each target site selection
- Figure 2-5 DNA-binding specificity of selected HD variants.
- Figure 2-6 Selected HDs with favorable recognition preferences for each target site
- Figure 2-7 Diversity in the specificity of extant HDs.
- Figure 2-8 Robust specificity determinants observed in the selected HDs.
- Figure 2-9 Determination of the dissociation constant for each HD variant.
- Figure 2-10 Determination of the dissociation constant for different binding sites through cold competition.
- Figure 2-11 Robust function of the new specificity determinants.

- Figure 2-12 Robust function of these New specificity determinants.
- Figure 2-13 New specificity determinants function with 5' specificity alterations.
- Figure 2-14 Modeling of HD variants.
- Figure 2-15 Additional modeling of HD variants.
- Figure 2-16 Limited diversity at the key recognition positions is observed in extant HDs
- Figure 2-17 MSE contribution per position in refined RF recognition models
- Figure 3-1 Possible modes of repair after a DSB is created in the genome by customizable sequence-directed endonucleases
- Figure 3-2 Cas9/CRISPR utilizes the sgRNA to direct site-specific DNA cleavage
- Figure 3-3 Structure of ZFP and TALE interacting with DNA and cartoon illustrating their mode of specificity (Gersbach)
- Figure 3-4 Cartoon representation of the general architecture of ZFNs and TALENs
- Figure 4-1 Schematic of nZFHD, nZF, and ZFN.
- Figure 4-2 Optimization of the linker between the ZF and the HD
- Figure 4-3 Models with spacings between the ZF and HD
- Figure 4-4 Stringency of selected linkers between the ZF and HD
- Figure 4-5 Models estimating the distance between fusing the nuclease to the N-terminus of the ZF
- Figure 4-6 Yeast activity assay showing relative activity for nZFHDs
- Figure 4-7 Utilizing nZFHD in zebrafish

## **List of Abbreviations**

DNA	Deoxyribonucleic acid
DBD	DNA-binding domain
HD	Homeodomain
ZF	Zinc finger
ZFP	Zinc-finger protein
bHLH	Basic helix-loop-helix
En	Engrailed
B1H	Bacterial-one hybrid
ZFN	Zinc-finger nuclease
ZFHD	Zinc finger homeodomain
nZFHD	Nuclease zinc finger homeodomain
TF	Transcription factor
DSB	Double-stranded Break
NHEJ	Non-homologous end joining
HDR	Homology-directed repair
TALE	Transcription activator-like effector
TALEN	Transcription activator-like effector nucleases

## **List of Third Party Copyrighted Material**

Chapter 3, Figure 2 – modified from: Jinek, M., East, A., Cheng, A., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. *eLife*, 2:e00471. Creative Commons Attribution License

Chapter 3, Figure 3 – modified from Perez-Pinera, P., Ousterout, D.G., and Gersbach, C.A. (2012). Advances in targeted genome editing. (2012). *Curr Opin Chem Biol*, 16(3-4):268-77. License number: 3182140503449

Chapter 3, Figure 4 – modified from: Joung, J.K., and Sander, J.D. (2013). TALENs: a widely applicable technology for targeted genome editing. (2013). *Nat Rev Mol Cell Biol*, 14(1):49-55. License number: 3182210707047

## **PREFACE**

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

Chapter II has been publish previously as:

Stephanie W. Chu, Marcus B. Noyes, Ryan G. Christensen, Brian G. Pierce, Lihua J. Zhu, Zhiping Weng, Gary D. Stormo, and Scot A. Wolfe (2012). Exploring the DNA-recognition potential of homeodomains. *Genome Research* 22, 1889-1898

**CHAPTER I**  
**INTRODUCTION**



## **Molecular Biology Perpetuates Life**

Life, along with its beauty (a subjective value of living as perceived by the individual), is the motivation that perpetuates our existence (the objective state of living as perceived by a society). It is this positive feedback cycle that permits us to study life, to further molecular biology, thereby creating a greater understanding of life's beauty. Nonetheless, from a biologically scientific standpoint, life is merely an object that contains in itself an innate self-sustained system; although its exact definition is debatable (Koshland 2002). From a purely reductionist interpretation, where the cell is regarded as the basic unit of life, life is the result of genetic material and how that genetic material is regulated. For cells to exist as live entities, genomes must be regulated for biological processes to occur, which permits for the necessary dynamics in molecular biology that allow for our lives to thrive.

To manipulate individual cells, and thus whole organisms, to study living entities, they must be fully understood through the molecular parts that drive it. Doing so allows for the further understanding of how life functions, enabling diseases and illnesses to be cured and even prevented, thus allowing for the perpetuation of life. Life requires that genomes be regulated, where *trans*-acting factors act on *cis*-regulatory elements. These two distinct *cis* and *trans* components are intertwined as dynamic constituents that permit the processes of cell growth, cell division, cell differentiation, and even cell death.

The *trans*-acting factors encompass general transcription factors and gene-specific transcription factors that regulate gene expression by interacting on *cis*-regulatory elements, which permits cells to function and survive. General

transcription factors are those factors that regulate a basal level of transcription to almost all genes (Thomas and Chiang 2006). Gene-specific transcription factors act on specific genes to regulate a particular biological process (Brivanlou and Darnell 2002). As the complexity of an organism increases, the number of transcription factors increase, which further increases the gene regulatory network complexity (van Nimwegen 2003). The evolution of transcription factors (and *cis*-regulatory elements (Schmidt et al. 2010), and the complexity that is created by this expansion enables biological diversity to occur. Consequently, the molecular progression of transcription-factor evolution can give rise to the intense yet subtle beauty of organismal evolution (Babu et al. 2004).

### **DNA-binding Domains**

Sequence-specific transcription factors contain at least one DNA-binding domain (DBD) to facilitate target recognition within the genome. DBDs discern different DNA sequences through reversible intermolecular protein-DNA interactions read from the sequence, shape, and inherent complexities contained in double-stranded DNA (Rohs et al. 2010). DBDs are grouped into families of related structures that are utilized for recognition, where DBDs within a given family recognize DNA with similar mechanisms (Ades and Sauer 1995). The three most common DBD families constitute 80 percent of human transcription factors. Starting with the most common, they are: C2H2 zinc fingers domains (ZFs), homeodomains (HDs), and basic-helix-loop-helix domains (bHLHs) (Vaquerizas et al. 2009). ZFs, each of 30 amino acids, fold into a beta beta alpha motif around a zinc

ion to recognize a three to four base pair sequence, where ZFs can be fused in an average of 8.5 tandem arrays as a zinc finger protein (ZFP) to recognize longer sequences (Enuameh et al. 2013b). The HD, of 60 amino acids each, consists of a bundle of three alpha helices preceded by an N-terminal arm to recognize a core six base-pair site, however, they typically function with other cofactors/HDs to recognize their targets (Gehring et al. 1994b). bHLHs, each consisting of 60 amino acids, fold into two helices joined by a variable loop, where it homo- or heterodimerizes into a four helix bundle with another bHLHs to recognize a six base-pair sequence with its basic helix region (Grove et al. 2009). This sample of structure and function in protein-DNA recognition observed in the top three DBD families is only cross section of a greater diversity that is observed in the remaining families.

## **History and Biology of Homeodomains**

HDs were initially discovered in *Drosophila* as a homeobox contained in a homeotic gene, which is a gene involved in programming specific cell lineage that ultimately give rise to body parts (McGinnis et al. 1984). Aberrant function of certain homeotic genes in flies results in segment transformation during development including the incorrect development of legs instead of antennae and the development of first legs into third legs (Harellrigg and Kaufman 1983). It was subsequently shown that it was the homeobox (which encodes the HD) within the homeotic gene that is responsible for developmental regulation through its DNA-binding properties (Kuziora and McGinnis 1989; Mann and Hogness 1990).

Moreover, changes in the HD sequence can affect its DNA-binding properties, which can lead to differential gene regulation (Otting et al. 1990). HDs have since been implicated in a broad spectrum of biological processes and found to be broadly represented across eukaryotes.

The HD is best known to be encoded by genes of the HOX clusters where the genes within the cluster regulate anterior and posterior body development.

*Drosophila* have one HOX cluster while vertebrates have four. The four arose from the duplication of a single cluster, and the paralogous HD sequences within these clusters are highly similar (Burglin 2011). A cluster is an evolutionarily conserved tandem arrangement in the genome that spatially parallels the order of where the HD-containing genes function in embryo development. This phenomenon is also known as colinearity, and was first demonstrated in *Drosophila* (Nusslein-Volhard and Wieschaus 1980). HDs have since been expanded to various superclasses and classes grouped by sequence similarity and function from different organisms, where some of these HD-containing genes are contained in clusters, while others are dispersed throughout the genome (Gehring et al. 1994a).

While the HD was identified as a functional unit responsible for biological processes through DNA recognition, how HDs regulate these processes can also require other domains or motifs. These associated domains or motifs, including zinc fingers or POU-specific domains, can be located either N-terminal or C-terminal to the HD itself at variable distances from the HD (Burglin 2011). Additionally, HDs can also require cofactors to bind to the *cis*-regulatory element the HD is regulating. The yeast HD, MAT $\alpha$ 2, binds DNA cooperatively with either the MAT $\alpha$ 1 or Mcm1

to regulate mating type switching of the yeast (Herskowitz 1989). Another well-studied example is how HOX proteins recognize their *in vivo* DNA sites. In *Drosophila*, HOX factors require an interaction with the HD Exd for correct anterior and posterior development (Mann et al. 2009).

### **General Homeodomain Characteristics**

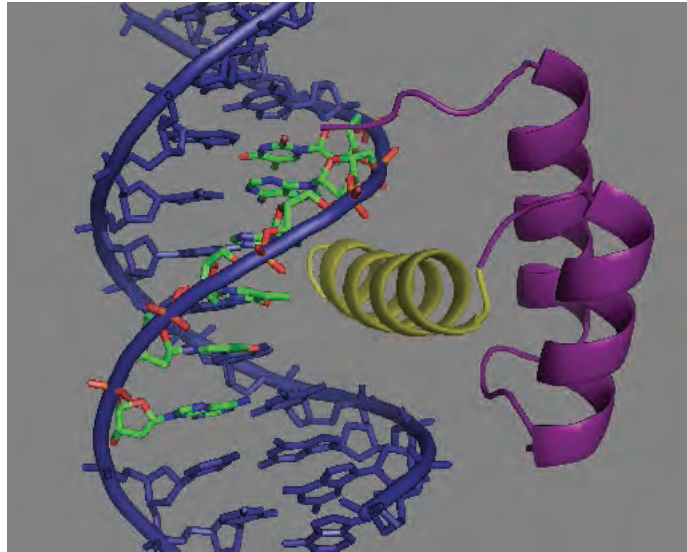
The HD typically consists of a sequence of sixty amino acids and binds to a core six base-pair binding site where an invariant adenine is observed at base 3. The exact binding site length, however, may range from five to eight base pairs depending on the HD. Residues 1-8 are part of an N-terminal arm, 10-22 are part of the first helix, 28-38 part of the second helix, and 43-57 is part of the third helix, also known as the recognition helix. It is the recognition helix and the N-terminal arm that dictates the DNA-binding specificity of the HD, where the N-terminal arm generally directs specificity of the 5' part of the site and the recognition helix directs 3' specificity. Particular to the fold of the HD are 7 positions that are observed to contain the same amino acids more than 95 percent of the time (Gehring et al. 1994a). The HD sequence contains a hydrophobic core of amino acids, which includes: L16, F20, and mostly the invariant W48 and F49. Additionally, an almost invariant N51 and well conserved residues R5 and R53 are involved in direct DNA recognition. The N51 is of particular importance to specifying the adenine at base 3. When base 3 is mutated to N7-deazaadenine to abolish only a single hydrogen bond within the binding site a greater than 100-fold reduction in binding affinity to the HD is observed (Ades and Sauer 1995).

## **Molecular Interactions Between the Homeodomain and DNA**

Original structures of the HD-DNA complex determined by NMR and X-ray crystallography elucidated how the gross structure of the HD interacted with DNA and showed some of the specific residues that interact with the binding site. The HD consist of an N-terminal arm followed by a bundle of three alpha helices with the third helix perpendicular to the first two (Figure 1-1), where the HD bears resemblance to the helix-turn-helix domain found in prokaryotes. To direct DNA-binding specificity, the recognition helix docks in the major groove of DNA, while the N-terminal arm interacts with in the minor groove. Contacts observed in the original structure by NMR of Antp bound to the site TAATGG are residues I47, N50, and M54, which contact the bases. Residues R5 and Y8 of Antp were observed to make contacts with the DNA backbone (Otting et al. 1990).

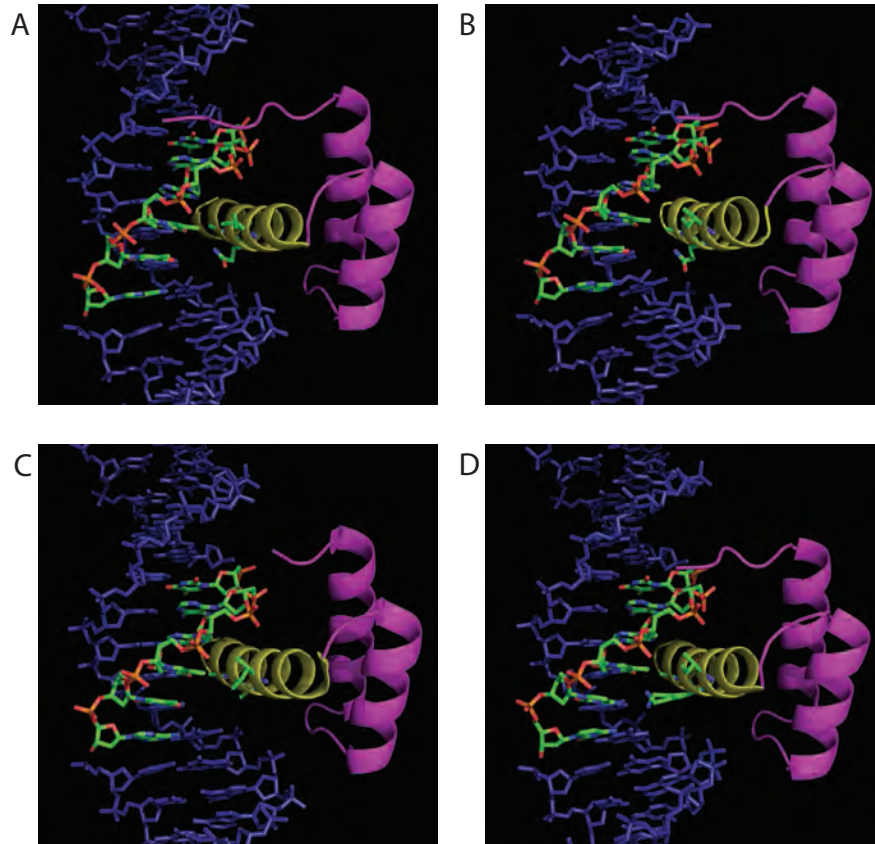
The first X-ray structure of a HD-DNA complex of Engrailed (En) to its cognate site, TAATTA, further identified critical contacts and revealed the more detailed set of side-chain interactions (Figure 1-2A) (Kissinger et al. 1990). The most prominent interaction identified involved in HD-DNA recognition is the invariant N51, where a bidentate hydrogen bond is made with the N7 and N6 of the adenine at base 3. The non sequence-specific interaction made by hydrogen bonding R53 to two phosphate groups of the DNA backbone was also identified. Greater details to interactions by the recognition helix identified by this crystal structure include I47 making a hydrophobic contact with base 4, Q50 with van der Waals interaction with the complement of base 6, and the Q50 is observed to be in a

**Figure 1-1**



**Figure 1-1:** Cartoon representation of the En HD-DNA complex grossly illustrates how the HD and DNA interact (Fraenkel et al. 1998). HD recognition helix (yellow) docks into the major groove of double stranded DNA while the N-terminal arm inserts into the minor groove. The primary strand of the core 6 base pair binding site is highlighted (green) to emphasize the proximity of these residues to the 3' end (bottom) of the recognition sequence. the complement of base 5. Additionally, the recognition helix also makes extensive

**Figure 1-2**



**Figure 1-2:** Cartoons of the multiple En variant-DNA structures show different possible interactions between the HD and DNA. Residue 47, 50, and 54 within the recognition helix are shown and the same coloring scheme is used from figure 1-1. (A) The first wild-type En with its cognate site (TAATTA) structure (1HDD)(Kissinger et al. 1990), (B) higher resolution structure of wild-type En with its cognate site (3HDD)(Fraenkel et al. 1998), (C) Q50A En variant with the wild-type En cognate site (1DUO)(Grant et al. 2000), (D) and Q50K En variant with the site TAATCC (2HDD)(Tucker-Kellogg et al. 1997).



proximity where small changes in DNA conformation would allow intermolecular interaction to occur with sugar-phosphate backbone contacts. The N-terminal arm shows fewer, yet critical base specific interactions, which includes hydrogen bonds of R5 to base 1 and R3 to the complement of base 2.

Since the original studies, numerous structures with greater resolution have furthered the understanding of the intermolecular interactions between the HD and DNA. Multiple structures of En mutants with different residues at position 50 have defined different intermolecular contributions to different binding site. While the wild-type En Q50 shows additional water-mediated contacts to base 4 and 5 to its cognate site (Figure 1-2B) (Fraenkel et al. 1998), a Q50A mutation imparts little overall rearrangement in interactions with the binding site, implying a modest role for Q50 in recognition (Figure 1-2C) (Grant et al. 2000). The structure of En with Q50K complex to the binding site TAATCC, however, demonstrates the importance of residue 50 with a pair of hydrogen bonds from the lysine to the complementary guanines of base 5 and base 6 (Figure 1-2D) (Tucker-Kellogg et al. 1997). These collective structures of En variants demonstrate that the interactions for a given residue in the HD to a binding site are contingent on the residue and base combination present.

Crystal structures of other HDs have further validated critical residues in the HD to interact with its binding site. Within the recognition helix, residues 47 and 54 have also been shown to interact with the bases 4 through 6 (Wolberger et al. 1991; Grant et al. 2000; Hovde et al. 2001), while residue 55 can interact with base 2 (Passner et al. 1999; Piper et al. 1999). Within the N-terminal arm residues 2, 3, 5,

6, and 8 have been observed to interact with base 1 through 3 (Fraenkel et al. 1998; Hovde et al. 2001). Moreover, the N-terminal arm can specify a binding site through the recognition of the minor groove shape as demonstrate by HD-DNA complexes comparing binding to two related DNA sequences that have different minor grove shapes (Joshi et al. 2007). The structures of these HD-DNA complexes taken together with mutational analysis illustrate the specificity determinant sets within the HD that can dictate its binding specificity to different binding sites.

### **Functional Residues for DNA-binding Specificity**

Alongside solved structure of HD-DNA complexes, mutational analysis of HDs has clarified the functional role of a specific residue or groups of residues within the HD (Figure 1-3). These studies have shown that there is rarely a simple one-to-one interaction between a residue and base. Substituting key residues in a HD, either in the recognition helix or the N-terminal arm, can change its binding specificity to recognize a site other than the HD's cognate site. Early mutational analysis focused on the role of residues in the recognition helix, where the S50K mutation in Prd allowed the mutated HD to recognize an alternate promoter (Treisman et al. 1989). Similarly, mutating Q50K within Ftz and Antp allowed these HDs to prefer a different binding site all together, switching from TAATTG to TAATCC (Percival-Smith et al. 1990; Hanes and Brent 1991). Overlapping residues can also affect the specificity of a given base position. Mutating I47N and A54R in En changes its preference from TAATTA to TAACA, however, when either single mutation is made

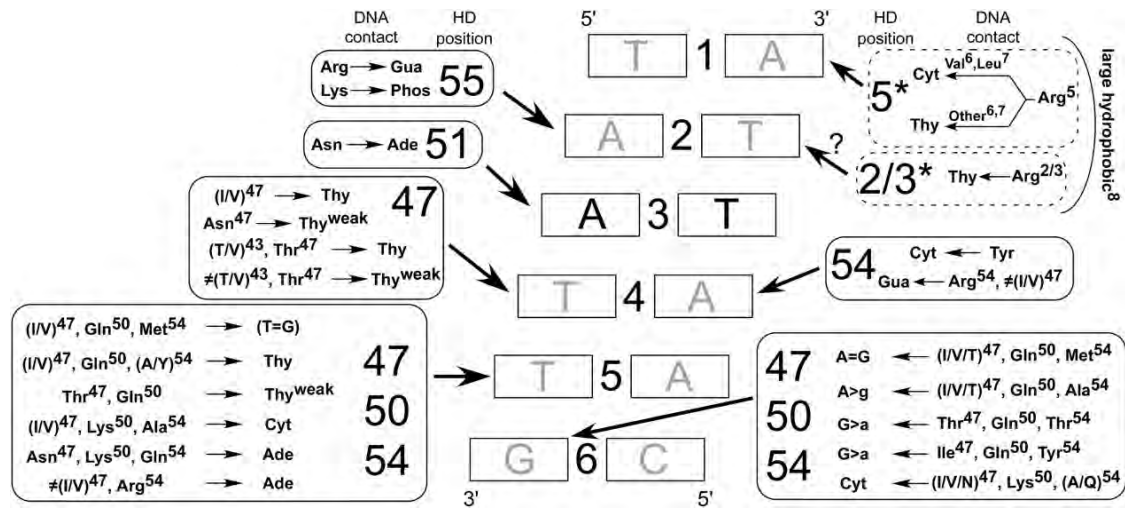
in isolation only a subtle difference from the original binding preference is observed (Noyes et al. 2008a).

The complexity of overlapping residues affecting specificity of a given base position and a residue affecting multiple bases is not strictly limited to the recognition helix. Mutating the residues 3, 6, and 7 of Ubx to that of Abd-b changes the preference of the homeodomain from TAATGG to TTATGG (Ekker et al. 1994). Likewise, mutating residues 6-8 of TTF-1 to those found in Antp changes the preference from CAAGTG to TAAGTG (Damante et al. 1996). Moreover, an A8F in Caup can strengthen the specificity at base 1 and subtly change the specificity at base 2 (Noyes et al. 2008a). A combination of mutations in both the N-terminal arm and recognition helix of R3K and K55R changes the specificity of the En site from TAATTA to TGATTA (Noyes et al. 2008a). A total of R3K, 147N, Q50A, A45R, and K55R can dramatically change the specificity of En to TGACA, illustrating the flexibility in binding site specification of En (Noyes et al. 2008a). Collectively, these studies reveal the overall complexity and interdependence within the HD for what are thought to be the general determinants of specificity (Figure 1-3).

### **Exploring the Recognition Potential of Homeodomains**

With the information present above one would speculate that the HD is a scaffold that is amenable to recognize a broader range of sites. This is, however, not what is observed of naturally occurring HDs that have had their DNA-binding specificity measured. Moreover, some previous attempts at radical specificity

**Figure 1-3**



**Figure 1-3:** Previously published chart that catalog HD specificity determinants (Noyes et al. 2008a).

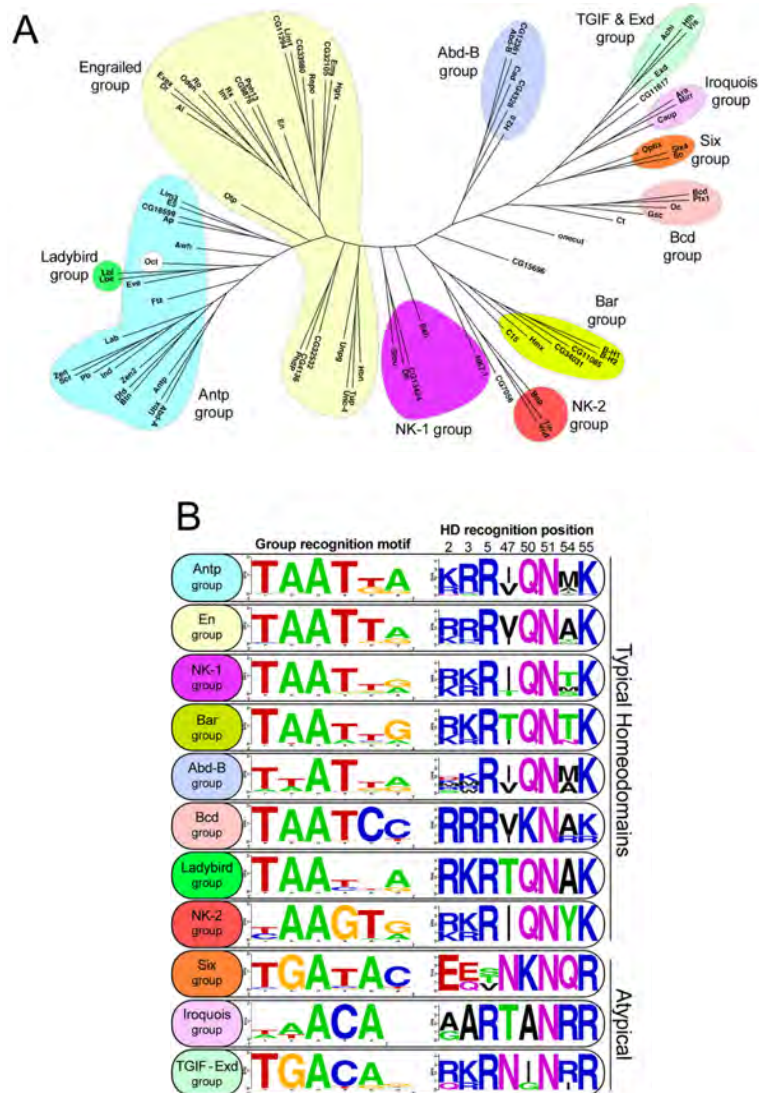
The numbers with the boxes are amino acid positions that most likely influence the sequence preference at a particular base position (solid line, major groove; dotted line, minor groove), where an arrow points from the box of the potential interactions to the base within each described base pair.

alteration have not proven successful, possibly owing to the technology limitations in creating large libraries at the time.

HD specificity has not been observed to deviate far from measured AT-rich sequences (Figure 1-4). Specificities of characterized HDs have been shown to typically recognize a core TAAT site (Gehring et al. 1994b). Recent studies characterizing HD specificity in humans, mice, and flies have measured the comprehensive HD DNA-binding specificities. The HDs measured for humans (146 HDs) (Jolma et al. 2013), mice (168 HDs) (Berger et al. 2008), and flies (84 HDs) (Noyes et al. 2008a) grouped HD specificity into 14, 33, and 11 specificity groups, respectively, where the majority of HDs within these groups are recognizing AT-rich sequences (Figure 1-4).

With the plethora of studies to understand the HD-DNA binding interface, there still appears to be difficulties in reengineering the HDs to recognize a broad range of specific DNA sequences. This is demonstrated in a number of studies: Full randomization of residue 47 and 51 of the POU HD can only give rise to limited differential HD binding of DNA sequences (Pomerantz and Sharp 1994). Reengineering the Mata alpha2 HD through substitutions at residue 50 did not result in any variants with equal affinity to the parent HD or sequence discrimination against its cognate DNA (Mathias et al. 2001). A study that tested 19 En HD combinations of amino acids at residues 50 and 54 against 4 DNA sequences only resulted in moderate changes in DNA-binding specificity. Only one HD variant in this study showed different sequence discrimination than the parent HD, where the affinity of that HD variant was not as strong as the parent HD to its cognate site

**Figure 1-4**



**Figure 1-4:** Previously published clustering of sequence specificity groups based on fly HD sequence specificity appears limited (Noyes et al. 2008a). (A) While fly HDs are clustered in eleven specificity groups, (B) group specificity motifs still appear very similar and limited.

(Connolly et al. 1999). These studies imply that the HD is not amenable to reengineering that would expand the range of DNA sequences a HD could recognize.

Nonetheless, the possible feasibility of the HD being amenable toward global reengineering for recognition of a new DNA sequences was demonstrated when larger combinatorial alterations within the recognition helix were utilized.

Reengineering En to recognize a DNA duplex containing an unnatural nucleotide showed that a HD could be selected to recognize a different DNA sequence other than the HD's cognate site with equivalent affinity and specificity resembling that of the natural HD-DNA interactions by randomizing residues 43-52 and 54 (Simon and Shokat 2004). This particular study implies that a larger combinatorial approach to selecting HD variants may prove successful to broadly reengineer the HD to recognize a diverse range of sites.

Here we challenge the view that the HD can only recognize such limited DNA sequences as demonstrated by previous literature. By doing so, we can examine if the HD scaffold is amenable to recognizing new DNA sequences if a larger, more complex library is utilized or if the HD sequence specificity will somehow still be constrained. Moreover, exploring the recognition potential of the HD will examine the degree to which the HD DNA-binding potential can be expanded and the diversity of protein sequence within the HD that can be obtained. By expanding the DNA-recognition potential of HDs, we can then catalog the novel specificity determinants to further predict HD binding specificity. Additionally, HDs with novel specificity may be utilized in customizable sequence-directed nucleases for targeting specific DNA sequences (see Chapter 3).

## Summary

We assert that HDs may have broader recognition potential than observed previously based on the large specificity analyses of naturally-occurring HDs. The limited recognition diversity observed in naturally occurring HDs is likely a reflection of the limited diversity of residues that are contained within key specificity determinants of the characterized HDs. To test if the HD can recognize a broader range of sequences we attempted to globally reengineer the HD to recognize all TAANN sites. By randomizing a combination of five residues within the recognition helix in En we selected HD variants to all 64 possible 3' binding sites. Our study identified HD variants that preferentially bind to 44 of the 64 possible 3' binding sites, where the novel specificity determinants created a catalog of new specificity determinants to further the understanding of HD-DNA specificity. A subset of these HD variants showed similar affinity and specificity to the naturally occurring En-binding site combination. The specificity determinants were tested to be robust in combination with 5' specificity determinants and thus would be useful to engineer HDs with a combination of 3' and 5' specificity. Our results expand the HD to specify a broader range of sequences than ever previously observed and shows that the HD is indeed a flexible scaffold amenable to broad reengineering.



## **CHAPTER II**

### **EXPLORING THE DNA-RECOGNITION POTENTIAL OF HOMEODOMAINS**

Chapter II has been published previously as:

Stephanie W. Chu, Marcus B. Noyes, Ryan G. Christensen, Brian G. Pierce, Lihua J. Zhu, Zhiping Weng, Gary D. Stormo, and Scot A. Wolfe (2012). Exploring the DNA-recognition potential of homeodomains. *Genome Research* 22, 1889-1898

Marcus B. Noyes performed the selections of HDs. Ryan G. Christensen created the improved prediction model. Brian G. Pierce created the models of interactions.

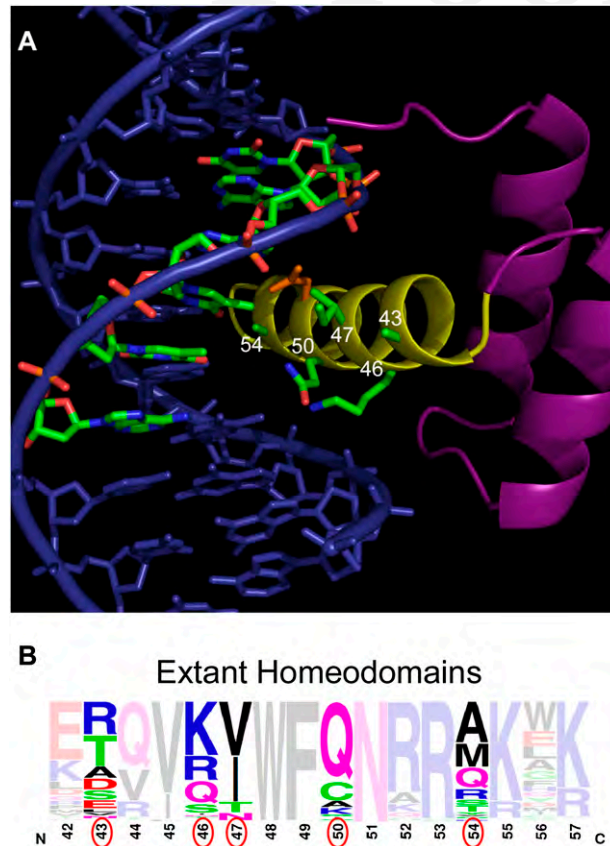
Lihua J. Zhu performed the statistical analysis.

## INTRODUCTION

Homeodomains (HDs) play a prominent role in regulating a multitude of biological processes in eukaryotes ranging from mating type switching in yeast to embryonic patterning in metazoans (Kornberg 1993; Gehring et al. 1994a). Emblematic of their central role in gene regulation, HDs are broadly represented across eukaryotic species; in humans they are the second most common family of DNA-binding domains (Vaquerizas et al. 2009). Consistent with their abundance, HDs display a diverse array of functions in development and cell-type specification, and they can be subdivided into a number of distinct families based on common sequence features and recognition motifs (Burglin 2011). Sequence-specific DNA recognition is central to many aspects of the regulatory function of HDs and as a consequence this characteristic has been extensively studied through genetic, biochemical, and structural analyses (Wolberger et al. 1991; Ades and Sauer 1994; Ekker et al. 1994; Gehring et al. 1994a; Damante et al. 1996; Fraenkel et al. 1998; Grant et al. 2000; Hovde et al. 2001; Babu et al. 2004; Joshi et al. 2007; Rohs et al. 2010; Slattery et al. 2011). HDs are typically composed of a ~60 amino acid motif that folds into a three-helix bundle preceded by an N-terminal arm. Sequence-specific recognition is mediated by the third (recognition) helix docking in the major groove and the N-terminal arm docking in the minor groove (Figure 2-1A) where a HD typically specifies a site of three to eight base pairs.

Many specificity determinants central to sequence-specific DNA recognition by HDs have been defined. A subset of these determinants function semi-autonomously, such that the transfer of a single residue between HDs can result in a

**Figure 2-1**



**Figure 2-1:** Structure of the *engrailed* HD and distribution of HD recognition residues.

A) Structure of the *engrailed* HD-DNA complex (Fraenkel et al. 1998), which serves as the framework for library construction. The numbers (white) on the HD recognition helix (yellow) indicate amino acid positions (green side chains) that were randomized, where the primary strand of the core 6 base pair binding site is highlighted (green) to emphasize the proximity of these residues to the 3' end of the recognition sequence. Asn51 (orange), which is highly conserved within the homeodomain family is shown for reference. B) Frequency logo displaying the diversity of residues (circled in red are the residues randomized in the HD library) at various positions in the N51-containing HDs in the genomes of humans, mice, *D. rerio*, *C. elegans*, *D. melanogaster*, and *S. cerevisiae*.

predictable alteration in specificity. This is demonstrated by seminal studies investigating the role of position 50 in the recognition preference of PRD, BCD and FTZ (Treisman et al. 1989; Percival-Smith et al. 1990; Hanes and Brent 1991). The critical features determining sequence specific recognition by the N-terminal arm remain nebulous and consequently achieving alterations in specificity typically necessitates the substitution of multiple residues between HDs (Ekker et al. 1994; Damante et al. 1996).

Recent comprehensive analysis of HDs specificity in the mouse and fruit fly (194 and 84, respectively) have somewhat clarified the breadth of DNA sequences HDs recognize in natural systems (Berger et al. 2008; Noyes et al. 2008a). While these studies used different approaches for determining DNA-binding specificity, they are in general concordance on the core DNA-binding specificity of homologous HDs. Limited sequence diversity is observed in the residues at the critical recognition helix positions within most eukaryotes (Figure 2-1B), and there is a corresponding paucity in the diversity of preferred recognition sequences observed for the characterized HD population (Berger et al. 2008; Noyes et al. 2008a). This focused sequence preference is similar to many other families of DNA-binding domains (Deppmann et al. 2006; Wei et al. 2010; De Masi et al. 2011), and could be the result of a general constraint of the domain architecture on its recognition potential. Consistent with this conjecture, previous attempts to select HDs with novel specificity have not succeeded in achieving dramatic alterations in recognition potential (Pomerantz and Sharp 1994; Connolly et al. 1999). These attempts, however, allowed variation at only a modest number of recognition positions. Thus,

it remains possible that HDs can recognize a broader range of DNA sequences than is currently observed.

Here we describe radically reengineering the DNA-binding specificity of the *engrailed* homeodomain to clarify the general recognition properties of this family. We systematically selected HD variants from a randomized library against all 64 possible combinations of the target site TAANNN. From these selections we were able to recover HDs that preferentially recognize 44 of the 64 sites, far more than anticipated based on the characterized set of extant HDs. The majority of these HDs harbor distinct combinations of specificity determinants, many of which appear to be uncommon or absent in extant HDs. These determinants expand our understanding of HD recognition, allowing the creation of more explicit recognition models for this family. The potential for this domain to recognize a broader range of DNA sequences raises questions about the fitness barrier that restricts the evolution of more diverse recognition properties for this family in natural systems.

## RESULTS

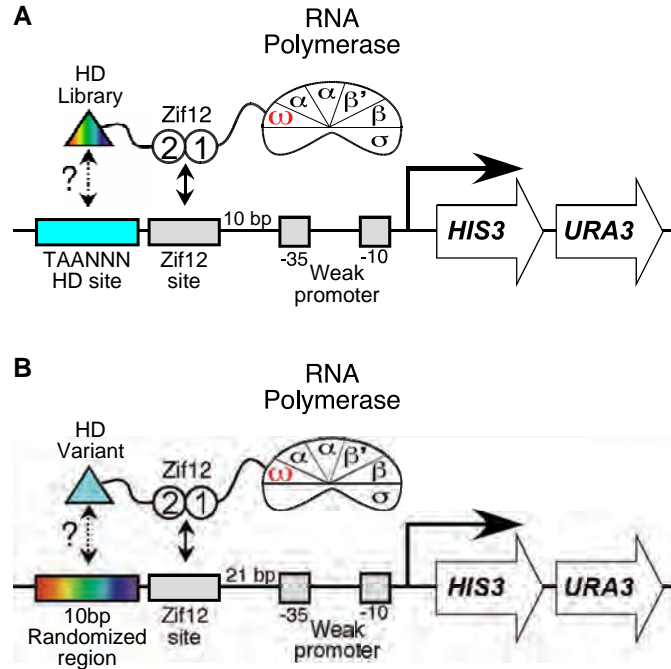
### **Selection of homeodomains with novel DNA-binding specificity.**

To explore the DNA-recognition potential of homeodomains (HDs), we investigated their ability to specify all possible TAANNN sites by selecting compatible HDs from a randomized library. These selections were performed using our bacterial one-hybrid (B1H) system (Noyes et al. 2008a; Noyes et al. 2008b), where the HD library is expressed as a fusion to two zinc fingers that position the library over the preferred target site (Figure 2-2). The *engrailed* (*en*) HD was

chosen as the library backbone because it is amenable to substitutions that change its DNA-binding specificity (Ades and Sauer 1994; Tucker-Kellogg et al. 1997; Noyes et al. 2008a).

Recognition of the 3' region (bases 4, 5, & 6) of the HD binding site is mediated by specificity determinants within the recognition helix. To select HD variants with altered sequence recognition preferences, residues 43, 46, 47, 50, & 54 were fully randomized (Figure 2-1). These positions, which all point toward the major groove in the EN-DNA complex, were chosen based on their potential function as primary or secondary recognition determinants within the 3' region of the target site. Direct base-specific contacts have been observed between residues 47 and 54 and base 4, as well as between residue 50 and bases 5 and 6 (Wolberger et al. 1991; Tucker-Kellogg et al. 1997; Fraenkel et al. 1998; Passner et al. 1999; Piper et al. 1999; Grant et al. 2000; Joshi et al. 2007), where sequence alteration at these positions has a direct influence on specificity (Treisman et al. 1989; Percival-Smith et al. 1990; Hanes and Brent 1991; Damante et al. 1996; Noyes et al. 2008a). Residues at positions 43 and 46 play a more subtle role in recognition (Kissinger et al. 1990; Fraenkel et al. 1998; Mahony et al. 2007; Noyes et al. 2008a). One additional prominent determinant, position 51, is almost exclusively asparagine within the extant HD population, where it specifies adenine at base 3. This position was held constant in our library, in anticipation that our selected HDs could be used to inform a predictive recognition model for extant HDs.

**Figure 2-2**



**Figure 2-2:** The B1H selection system.

(A) Schematic illustrating the components used for the B1H selection of HDs with novel specificity. The HD library (rainbow) is displayed as a C-terminal fusion to omega-Zif12 with a fixed binding site present within the reporter vector (cyan). (B) Schematic illustrating the components used to identify preferred binding sites for each HD variant (cyan). HDs are displayed as fusions to the C-terminus of the omega-Zif12, where compatible binding sites are selected from a 10-base pair randomized library (rainbow) in the reporter vector.

Selections employing the HD library were performed separately against each of the 64 TAANN sites to recover interacting HDs. We observed variability in the selection stringency required to cull the population down to 1000 to 2000 surviving clones for each target site (Figure 2-3). Overall, selections employing the HD library yielded a 20 to 200-fold increase in surviving colonies when compared to a negative control entirely lacking the homeodomain. Sequencing the recovered clones from each target site yielded a catalog of approximately  $4.4 \times 10^4$  HDs (Online Processed Illumina Supplemental Table S3\*), and revealed striking amino acid preferences at some randomized positions within populations recovered from different target sites (Figure 2-4). Some of these preferences were anticipated based on prior studies of HD specificity (Wolberger et al. 1991; Ades and Sauer 1994; Passner et al. 1999; Noyes et al. 2008a), but many appear to represent novel determinants.

### **Analysis of selected homeodomains.**

Prominent HD positions influencing base preference were identified by Mutual Information analysis on the catalog of selected HDs for each target site (Mahony et al. 2007). This analysis identified positions 47, 50 and 54 as strong contributors to 3' specificity, whereas positions 43 and 46 appeared to have little global influence on the 3' site preference (Table 2-1). Significant covariation was observed between residues 47 and 54, and base 4. In addition, a moderate degree of covariation is observed between both of these residue positions and base 5. Moderate covariation is also observed between residue 50 and all of the 3' base positions but is most pronounced with base 6. The most significant relationships



**Figure 2-3**

		POSITION 6			
		A	C	G	T
P O S I T I O N  4 &  5	AA	10	10	10	10
	AC	10	10	10	10
	AG	10	10	10	10
	AT	10	10	10	10
	CA	50	50	10	25
	CC	50	50	25	50
	CG	25	25	25	25
	CT	25	25	25	10
	GA	50	50	50	50
	GC	50	50	10	10
	GG	50	50	50	10
	GT	50	50	50	10
	TA	25	50	25	10
	TC	10	50	10	25
	TG	25	25	25	25
	TT	20	25	25	25

**Figure 2-3:** Stringency used to select HDs for different target sites. Chart indicating the stringency used to select HDs for each of the 64 TAANN sites. The bases present at positions 4, 5 (left of rows) & 6 (above each column) are indicated, where the number in each cell represents the concentration of 3-AT (mM) used for selection against that target site.

**Figure 2-4**



**Figure 2-4:** Logos representing the sequences of the recovered HDs from each target site selection.

Frequency logos representing the top 200 unique HDs sequences recovered for each of the 64 target sites from Illumina sequencing. Red circles indicate the randomized positions in the HD library.

**Table 2-1.** Mutual Information analysis of the selected homeodomain-binding site combinations

	Base Position 4	Base Position 5	Base Position 6
Residue 43	0.06	0.02	0.02
Residue 46	0.08	0.06	0.09
Residue 47	<b>0.71</b>	0.31	0.10
Residue 50	0.31	0.40	<b>0.53</b>
Residue 54	<b>0.77</b>	0.37	0.07

Mutual Information analysis indicates strong (bold) and moderate contributors to 3' specificity from residues 47, 50 and 54, indicating they are the primary determinants that influence specificity at base positions 4, 5 and 6. All values within the table are significant with p-value < 0.001.

identified between HD position and binding site position are consistent with previously published structural and biochemical data (Treisman et al. 1989; Percival-Smith et al. 1990; Hanes and Brent 1991; Wolberger et al. 1991; Damante et al. 1996; Noyes et al. 2008a).

### **Defining the specificity of selected homeodomains.**

In an attempt to distinguish selected HD variants that can preferentially bind to each of the 64 TAANN sites from those that can merely associate favorably with a target site, we determined the DNA-binding specificity for 151 HD variants (Figure 2-5, and Online Processed Illumina Supplemental Table S6\*). HDs variants were chosen for analysis based on their overlap with the consensus sequence recovered in each selected population or the presence of combinations of recognition residues that were deemed interesting (Figure 2-4 and Online Processed Illumina Supplemental Table S3\*). For example, in anticipation of identifying a HD variant that specifies TAACGG, we characterized a clone containing residues R47, E50, and R54 that reflects the predominant consensus sequence recovered for this target site. Preferential DNA-binding specificity for each HD was determined using the B1H system (Noyes et al. 2008a) where the entire population of hundreds to thousands of recovered binding sites was sequenced to construct a recognition motif (Figure 2-5).

Based on this analysis, we are able to identify HD variants that preferentially bind to or are compatible with 44 out of the 64 target sites (Figure 2-6), which



Figure 2 5.1

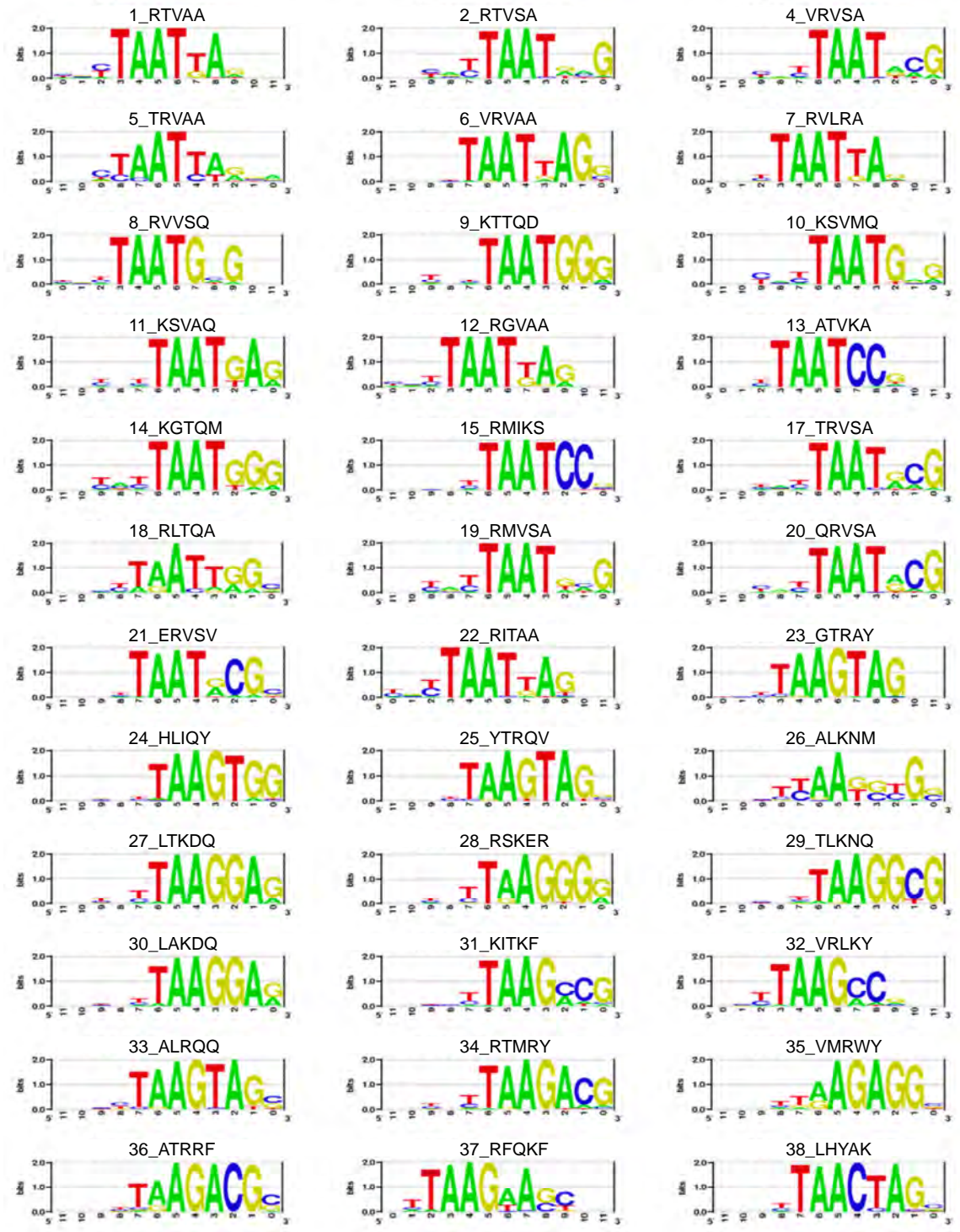


Figure 2-5.2

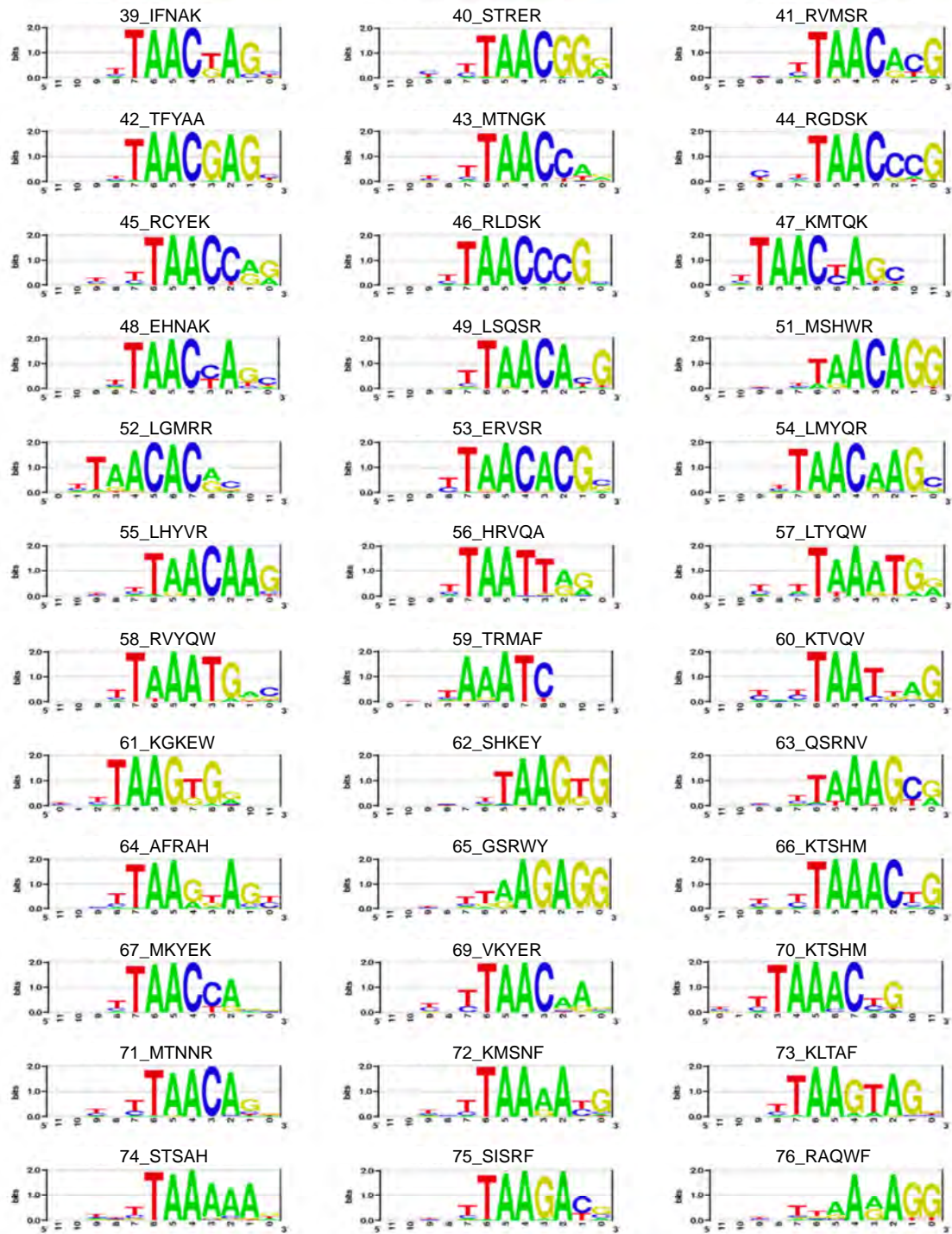




Figure 2-5.3

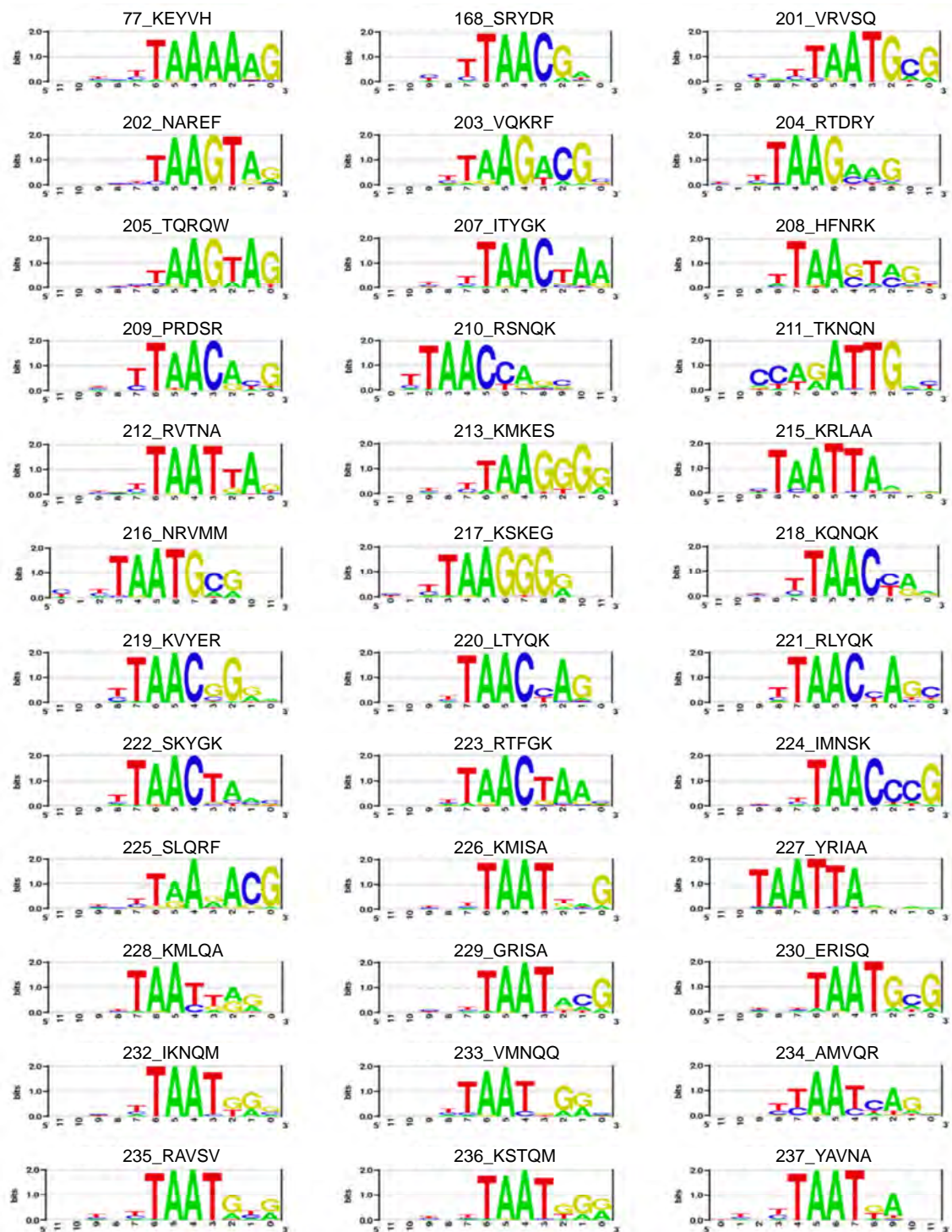
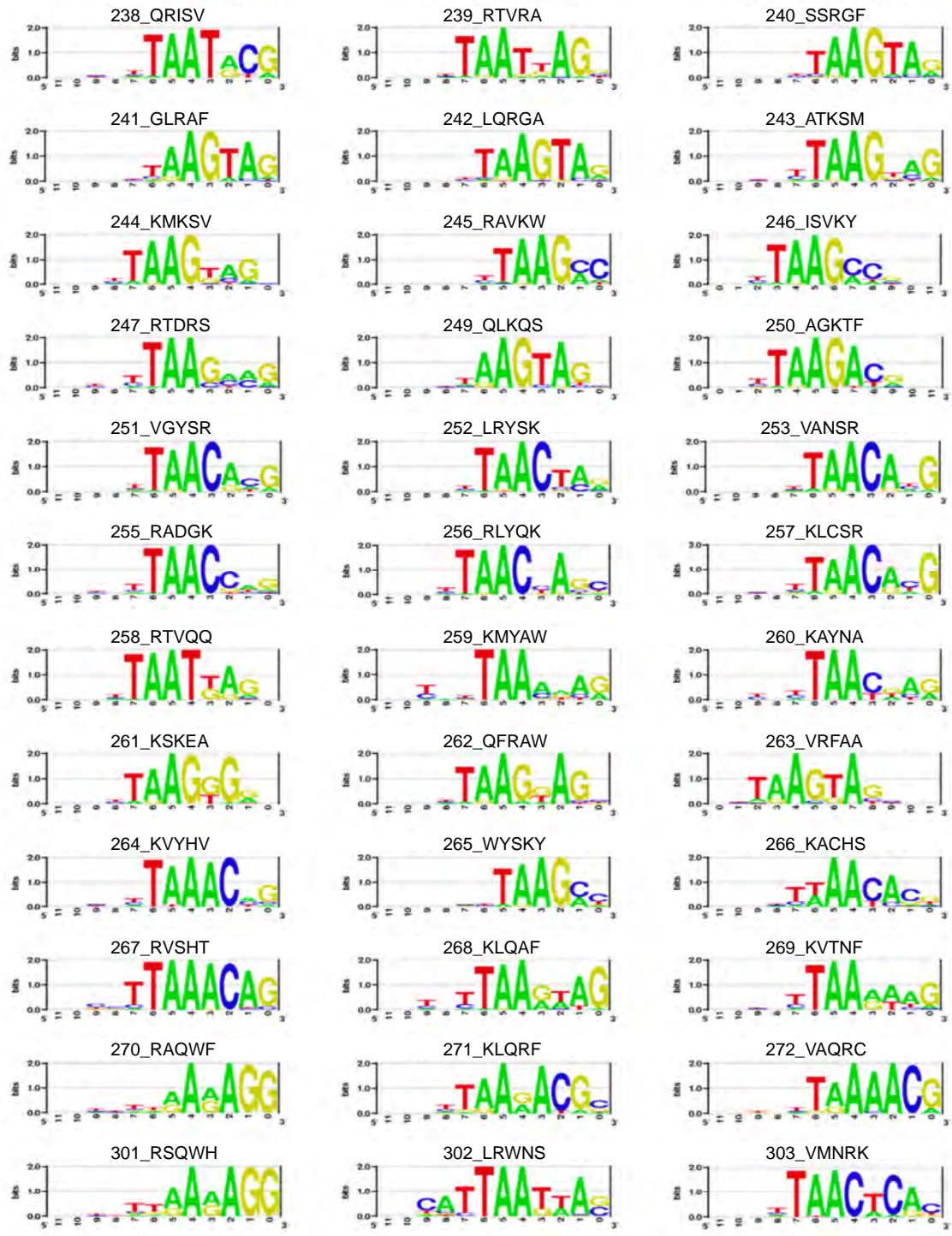
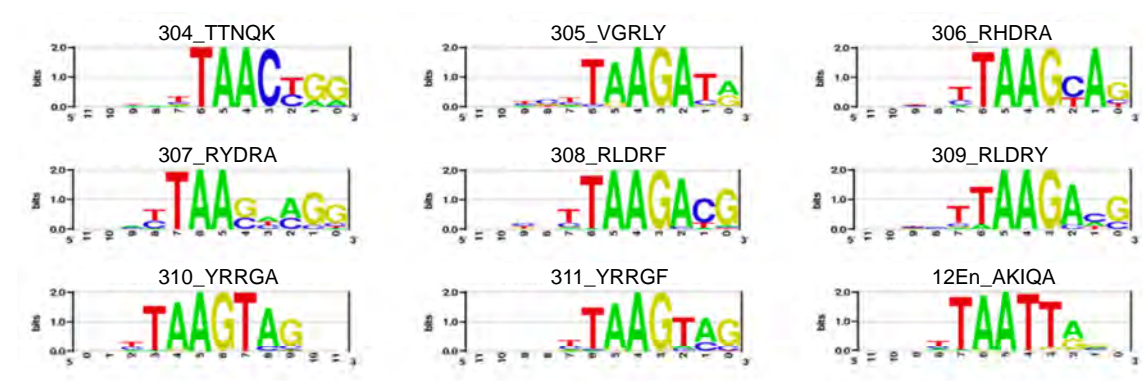


Figure 2-5.4





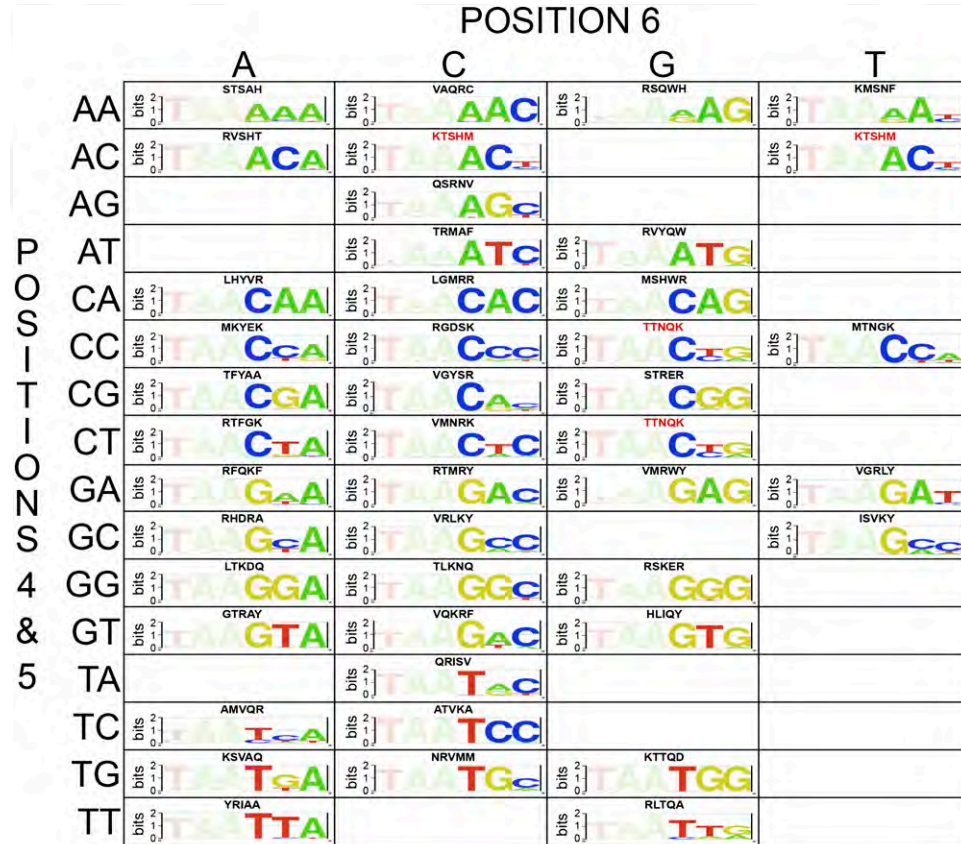
**Figure 2-5.5**



**Figure 2-5:** DNA-binding specificity of selected HD variants.

The calculated recognition motifs (bit scale) determined for each HD variant using the randomized 10-base pair library. The clone ID numbers and the amino acids that are present at the randomized recognition positions (43,46,47,50 & 54) are indicated above each motif.

**Figure 2-6**



**Figure 2-6:** Selected HDs with favorable recognition preferences for each target site.

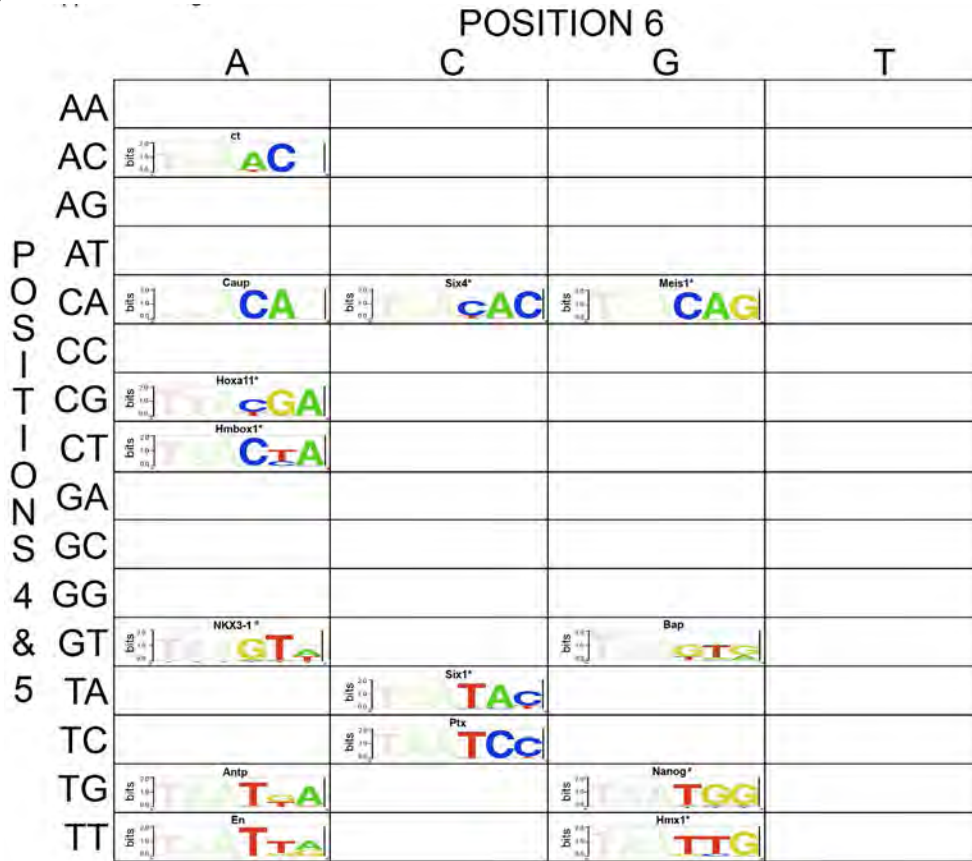
A grid illustrating the selected HD variants that preferentially recognize or are compatible with particular 3' binding site sequences. The amino acids that are present at the randomized recognition positions (43,46,47,50 & 54) are indicated above each motif. Sequences in red indicated those that are present in more than one grid position (i.e. are compatible with 2 different sites). Empty boxes indicated the absence of quality HD recognizing these sequences.

represents a sizeable expansion of the 3' specificities observed in characterized extant HDs (Figure 2-7). Our analysis of specificities further clarifies the significant association of specificity determinants with certain sequence preferences (Appendix Table A-1) and validates many novel specificity determinants (Figure 2-8 and Appendix Table A-2). Although, this analysis expands the number of primary determinants that can dictate recognition preferences, it is not possible to codify DNA recognition as a set of independent determinants because of the overlapping influence of neighboring determinants. Moreover, specifying some sequence features, such as T at base 6, appears challenging in any sequence context with this HD backbone and randomization scheme.

### **Sequence discrimination by homeodomain variants.**

We determined the affinity and specificity of a subset of HD variants for different binding sites *in vitro* using electrophoretic mobility shift assays. For this analysis, a subset of seven HDs were chosen that span members with both well-defined and novel specificity determinants (Table 2). In all cases, the apparent equilibrium dissociation constant of each HD for its cognate site was similar to the affinity of Engrailed for its cognate site (Figure 2-9). Cold competition assays were employed to determine the degree of discrimination of each HD variant between its cognate site and the parent Engrailed binding site (Figure 2-10). The difference in the free energy of binding the cognate and parent site ranged from 0.8 to 2.2 kcal/mol, where binding the cognate site was always favored (Table 2-2). The degree of discrimination determined for En between its preferred site, TAATTA, and

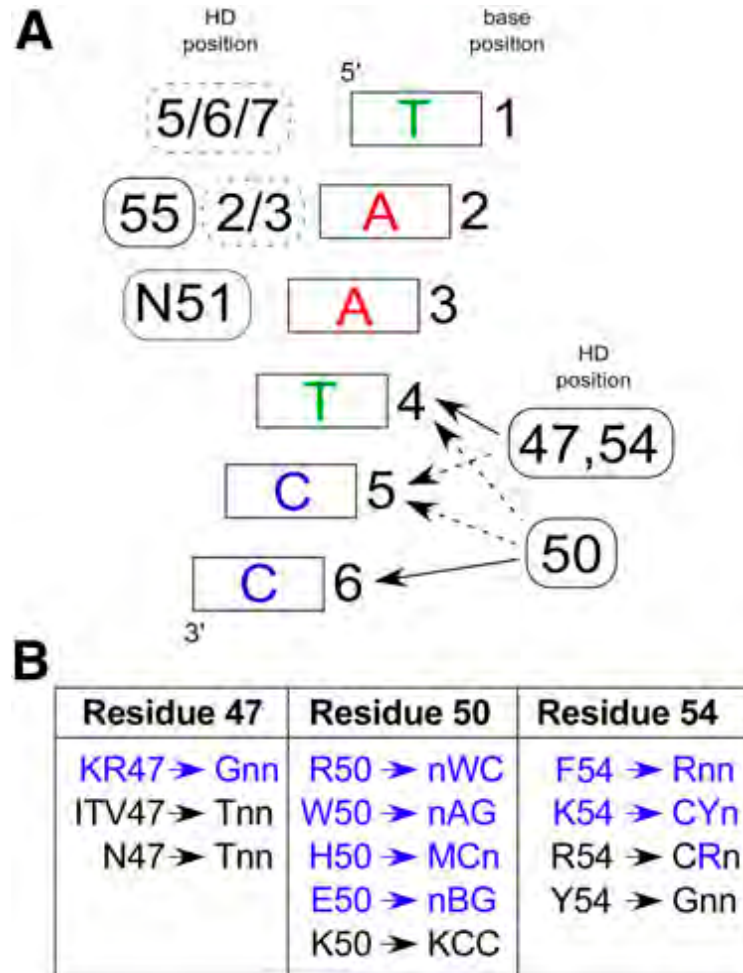
**Figure 2-7**



**Figure 2-7:** Diversity in the specificity of extant HDs.

A grid illustrating the different 3' specificities found in previously measured extant HDs from Noyes M.B and colleagues (Noyes et al. 2008a), Berger M.F. and colleagues (as denoted by \*) (Berger et al. 2008), Steadman D.J and colleagues (as denoted by @) (Steadman et al. 2000), and Jauch R. and colleagues (as denoted by #) (Jauch et al. 2008).

**Figure 2-8**



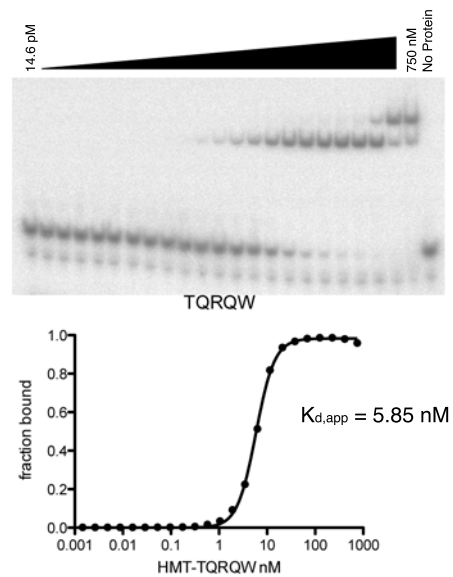
**Figure 2-8:** Robust specificity determinants observed in the selected HDs. (A) Canonical recognition pattern for HD-DNA interaction. At the 5' end of the binding site (bases 1, 2 and 3), positions on the recognition helix (solid boxes) and the N-terminal arm (dashed boxes) contribute to specificity, where the position(s) of the contributing determinants are indicated to the left of the base pair. At the 3' end of the binding site (bases 4, 5 and 6), homeodomain specificity is primarily defined by positions 47, 50 & 54, where these determinants have overlapping regions of influence. Solid arrows indicate primary positions of interaction and dotted arrows indicate secondary influences on specificity. (B) New specificity determinants (blue) and previously described specificity determinants (black) for HDs containing the conserved N51 are broken down by position and trends in base preference within the three basepairs at the 3' end of the target site. Note: there are exceptions within our characterized HDs to these specificity preferences, likely reflecting the overlapping influence of these determinants.

**Table 2-2.** Equilibrium dissociation constants of homeodomain variants

HD variant (Cognate site)	$K_{d,app}$ <sup>a</sup> (nM)	$h$ <sup>b</sup>	$K_{c,app}$ <sup>c</sup> (nM), Cognate site	$K_{c,app}$ <sup>c</sup> (nM), <i>engrailed</i> site <sup>d</sup>	Relative $\Delta\Delta G$ Affinity (kcal/mol )	
ATVKA (taaTCC)	4.40 ± 2.09	1.51 ± 0.19	3.17 ± 0.51	41.87 ± 4.25	13.22	1.52
HLIQY (taaGTG)	1.52 ± 0.08	1.57 ± 0.09	1.04 ± 0.11	16.64 ± 0.61	16.06	1.64
ERVSR (taaCAC)	19.09 ± 4.56	2.04 ± 0.11	14.00 ± 4.15	66.37 ± 22.40	4.74	0.91
TRMAF (taaATC)	4.03 ± 1.00	1.61 ± 0.22	1.74 ± 0.37	6.78 ± 1.65	3.90	0.80
TQRQW (taaGTA)	3.71 ± 1.31	1.99 ± 0.22	4.87 ± 0.21	193.72 ± 9.63	39.75	2.17
RSNQK (taaCCA)	9.83 ± 1.18	1.75 ± 0.12	8.92 ± 1.30	37.13 ± 7.33	4.16	0.85
LAKDQ (taaGGA)	5.69 ± 1.91	1.61 ± 0.21	3.50 ± 2.62	85.23 ± 26.52	24.37	1.89
Engrailed AKIQA (taaTTA)	2.34 ± 0.15	1.44 ± 0.08	0.74 ± 0.18	15.93 ± 4.73*	21.59**	1.81

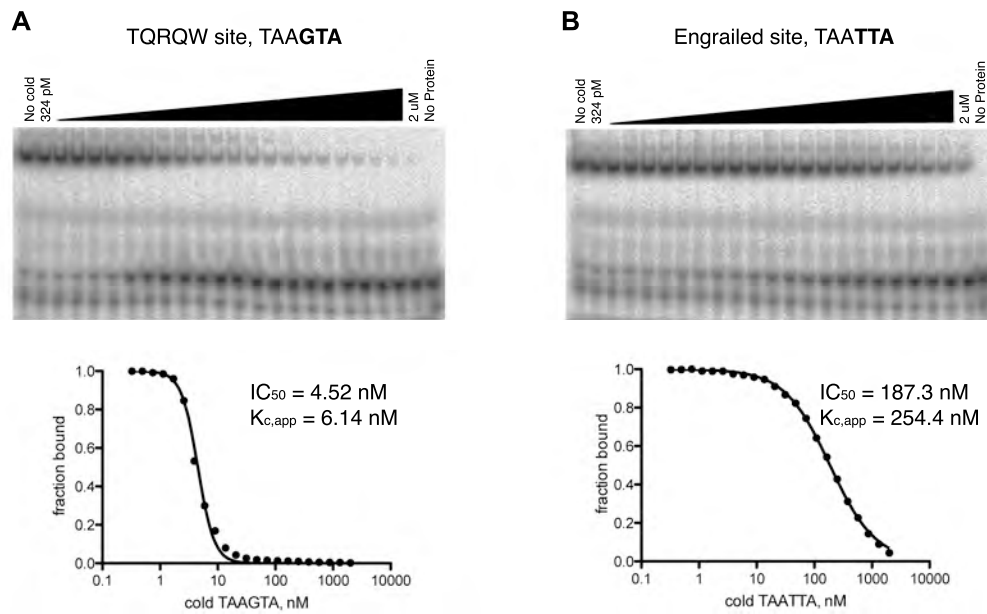
<sup>a</sup> Apparent equilibrium dissociation constant as determined by EMSA. <sup>b</sup> Hill coefficient ( $h$ ) as determined by EMSA. <sup>c</sup> Apparent equilibrium dissociation constant as determined by cold competition with indicated sequence. <sup>d</sup> Relative affinity ( $K_{c,app}$  *engrailed* site/ $K_{c,app}$  Cognate site). \* The  $K_{c,app}$  measured for the Engrailed HD is with the TAATCC site. \*\* The relative affinity for Engrailed ( $K_{c,app}$  TAATCC site/ $K_{c,app}$  Cognate site) is similar to that which was previously reported (Ades and Sauer 1994).

**Figure 2-9**



**Figure 2-9:** Determination of the dissociation constant for each HD variant. Apparent equilibrium dissociation constant as measured by EMSA for the HD variant TQRQW.

**Figure 2-10**



**Figure 2-10:** Determination of the dissociation constant for different binding sites through cold competition.

Apparent equilibrium dissociation constants were measured by cold competition for the HD variant TQRQW. The degree of competition achieved by titration of a cold competitor duplex containing (A) its preferred binding site, TAAGTA, or (B) the *engrailed* binding site, TAATTA, was measured by the decrease in complex formation with the labeled preferred binding site as a function of increasing concentration of the competitor.

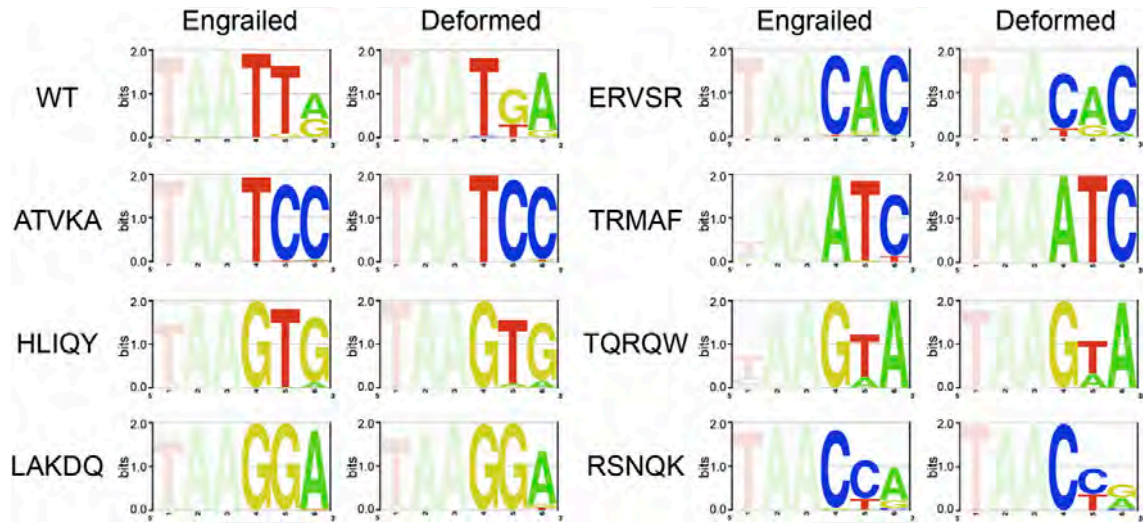


TAATCC (22-fold), which served as our internal control, was nearly identical to the difference previously reported by Sauer and colleagues (Ades and Sauer 1994). The TQRQW HD variant (selected HD variants are identified by the 5 amino acids selected at the randomized positions) has the greatest discrimination against the Engrailed site, displaying a 40-fold preference, while the TRMAF HD variant displays a modest 4-fold preference for its target sequence. Thus, our selected HDs display a consistent preference for their identified cognate site outside the B1H system.

### **Robust behavior of new specificity determinants.**

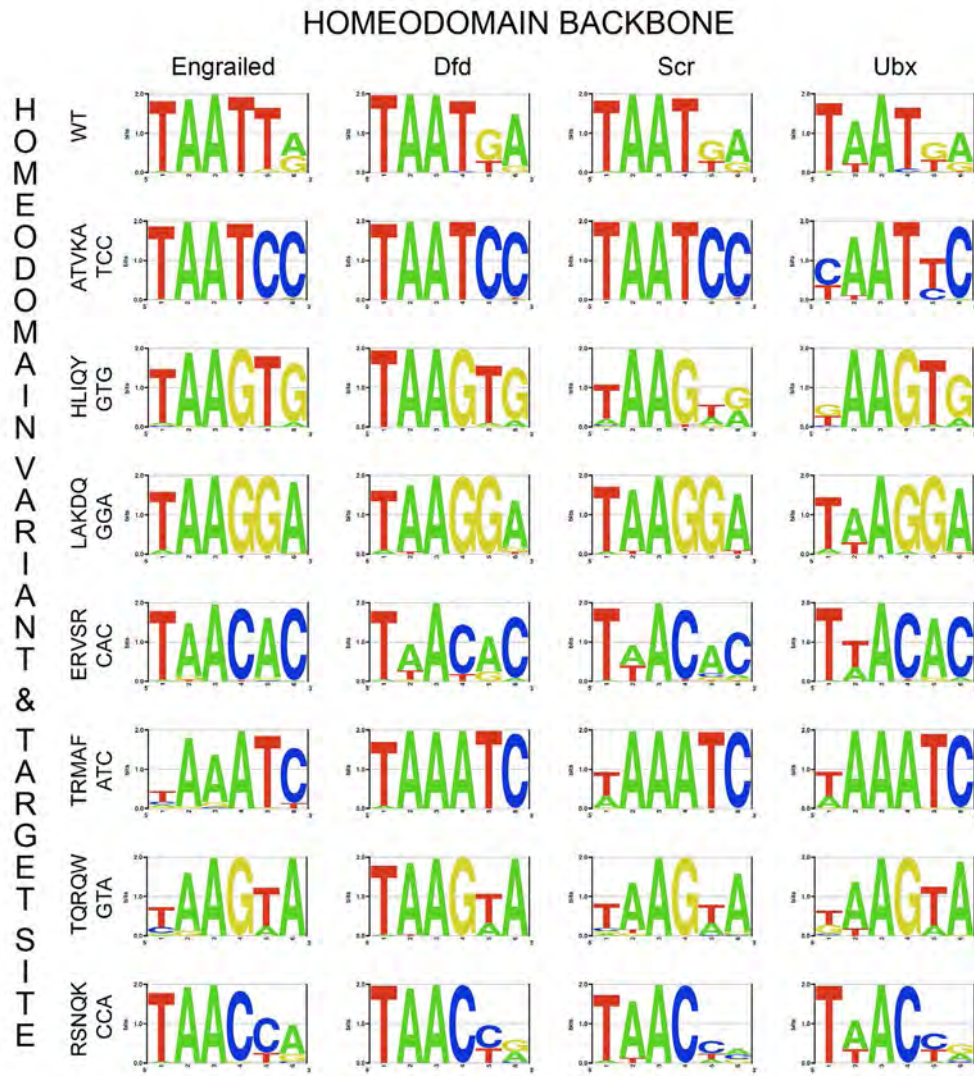
To determine if the newly observed specificity determinants are able to define similar DNA sequence preferences in the context of other HD backbones we grafted the 5 key residues, residues 43, 46, 47, 50, and 54, from each of the seven HD variants within the sample set into three other *D. melanogaster* HD backbones: Dfd, Scr, and Ubx. These HDs share 53%, 51%, and 46% identity with Engrailed, respectively. We then determined the DNA-binding specificity of all these variants using the B1H system (Figure 2-11 and Figure 2-12). In almost every instance the grafted residues altered the DNA-binding specificity of each Hox factor in a predictable manner, in agreement with the previously defined DNA-binding specificity in the Engrailed backbone. In a few instances, such as HLIQY, the introduction of these residues into the Hox backbone slightly altered 5' sequence preference. This alteration may indicate weak indirect effects of these altered

**Figure 2-11**



**Figure 2-11:** Robust function of the new specificity determinants. Grafting key residues (43, 46, 47, 50 & 54) selected from the Engrailed library into the HD backbone of the Hox factor Deformed transforms its sequence preference to resemble the corresponding selected HD mutant.

**Figure 2-12**



**Figure 2-12:** Robust function of these New specificity determinants. Grafting key residues (43, 46, 47, 50 & 54) selected from the *engrailed* HD library into the HD backbone of the Hox factor Dfd, Scr or Ubx results in a factor with a similar binding preference to that observed when the key residues are present in Engrailed.

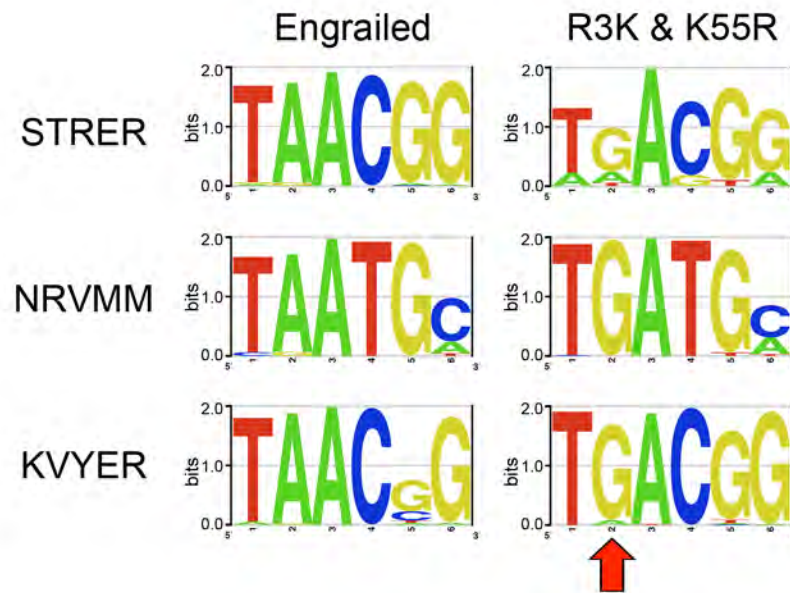
determinants on the 5' base preference, potentially through interactions with residues 51 and 55, which can influence 5' specificity.

We also examined the influence of different 5' specificity determinants on the 3' specificity of our selected HDs. Previous studies have shown that residues 3 and 55 influence the specificity at base 2, where the presence of K3 and R55 will preferentially recognize G over A (Passner et al. 1999; Piper et al. 1999; Noyes et al. 2008a). We introduced the mutations R3K and K55R into the Engrailed backbone for three HD variants (STRER, KVYER, and NRVMM) and determined their DNA-binding specificity (Figure 2-13). In all cases we observe a shift in specificity from A to G at position 2 without substantial alteration in base preference at the other recognition positions. The robust behavior of our new specificity determinants suggests that they will serve as useful parameters for the prediction of DNA-binding specificity in extant HDs.

### **Computational models of the interactions mediating sequence-specific DNA recognition.**

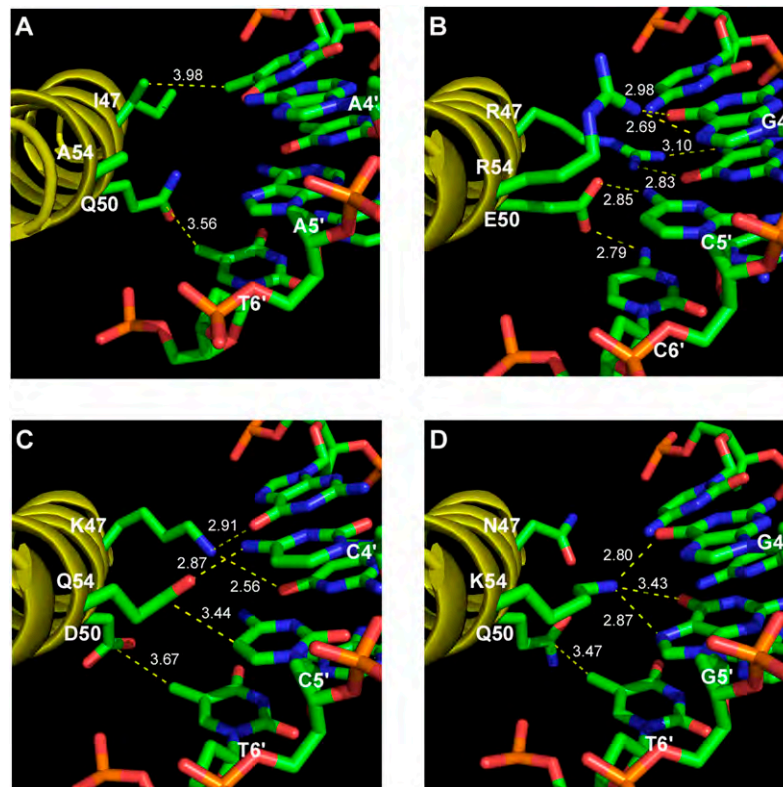
We utilized the Rosetta molecular modeling package, which has recently undergone significant revision for protein-DNA complexes (Yanover and Bradley 2011), to predict the base-specific interactions between our sample set of seven HDs and their cognate sites. These structural calculations used a high resolution Engrailed-DNA co-crystal complex as a starting model (Grant et al. 2000). In a number of instances, the calculated structural models yielded determinant – base interactions that are consistent with the correlated sequence preferences observed

**Figure 2-13**



**Figure 2-13: Supplemental Figure 9.** New specificity determinants function with 5' specificity alterations. Mutating the 5' specificity determinants R3K and K55R specifically alters 5' binding preference from TAANN to TGANN while the 3' binding preference remains unchanged.

**Figure 2-14**

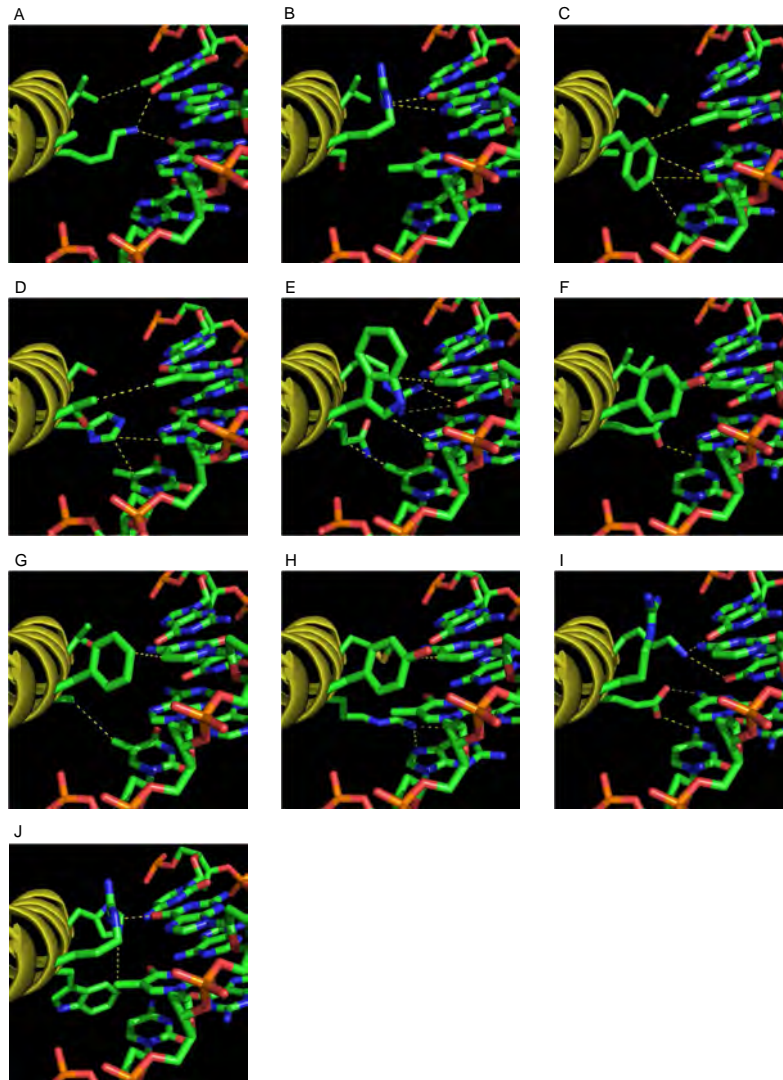


**Figure 2-14:** Modeling of HD variants.

(A) Cocystal structure of Engrailed bound to TAATTA (Fraenkel et al. 1998). (B) Model of HD variant STRER bound to its cognate site taaCGG. (C) Model of HD variant LAKDQ bound to its cognate site taaGGA. (D) Model of HD variant RSNQK bound to its cognate site taaCCA. Dotted lines indicate interactions between the protein and DNA (either hydrogen bonds or van der Waals interactions) where the numerical values indicate the distance in angstroms.



**Figure 2-15**



**Figure 2-15:** Additional modeling of HD variants.

(A) Model of HD variant ATVKA bound to its cognate site taaTCC. (B) Model of HD variant ERVSR bound to its cognate site taaCAC. (C) Model of HD variant TRMAF bound to its cognate site taaATC. (D) Model of HD variant RVSHT bound to its cognate site taaACA. (E) Model of HD variant TQRQW bound to its cognate site taaGTA. (F) Model of HD variant HLIQY bound to its cognate site taaGTG. (G) Model of HD variant KLTAf bound to its cognate site taaGTA. (H) Model of HD variant RTMRY bound to its cognate site taaGAC. (I) Model of HD variant RSKER bound to its cognate site taaGGC. (J) Model of HD variant MSHWR bound to its cognate site taaCAG. Dotted lines indicate interactions of less than 4 Å between the protein and DNA (either hydrogen bonds or van der Waals interactions).

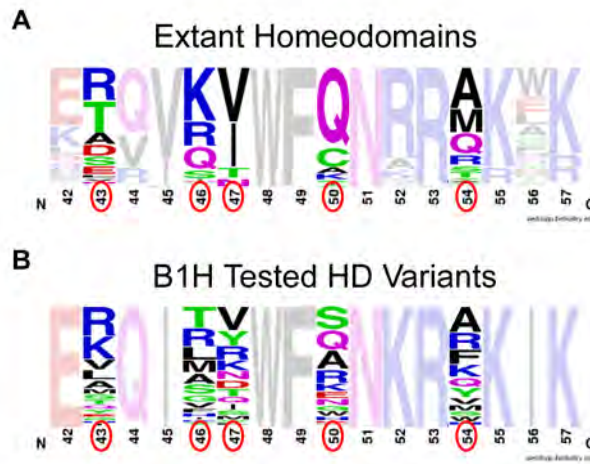
within our dataset of selected HDs allowing the potential roles of these determinants to be inferred (Figure 2-14 and Figure 2-15). For example, K47 in the LAKDQ – TAAGGA structural model positions the primary amine of this lysine between the O6 carbonyls of G4 and G5, mimicking the observed interaction of K50 with a pair of guanines on the complementary strand in the Q50K En – DNA structure (Tucker-Kellogg et al. 1997).

### **Improved predictive models of homeodomain specificity.**

Previous efforts to predict the DNA-binding specificity of HDs based on their amino acid sequence have focused on nearest neighbor estimates of specificity (Noyes et al. 2008a; Alleyne et al. 2009). We have recently shown that when high quality alignments of recognition motifs can be obtained, improved recognition models of HD specificity can be achieved using Random Forest based methods (Christensen et al. 2012). This recognition model, which is trained on the existing data for extant HDs, is a poor predictor of DNA-binding specificity for our selected HDs (MSE = 0.053; Appendix Table A-3). This deficit in predictive accuracy was expected given the increased diversity of recognition residues that are present in our selected HDs (Figure 2-16). Reassuringly, we found that a new recognition model trained only on the selected HDs performed reasonably well in the prediction of the extant HD set (MSE = 0.025; Supplemental Table S9), suggesting that much of the recognition repertoire that is present in the extant set is found in our selected HDs (Figure 2-17). In a 10-fold cross validation analysis, a joint recognition model between the selected and extant HDs provides excellent accuracy in the prediction



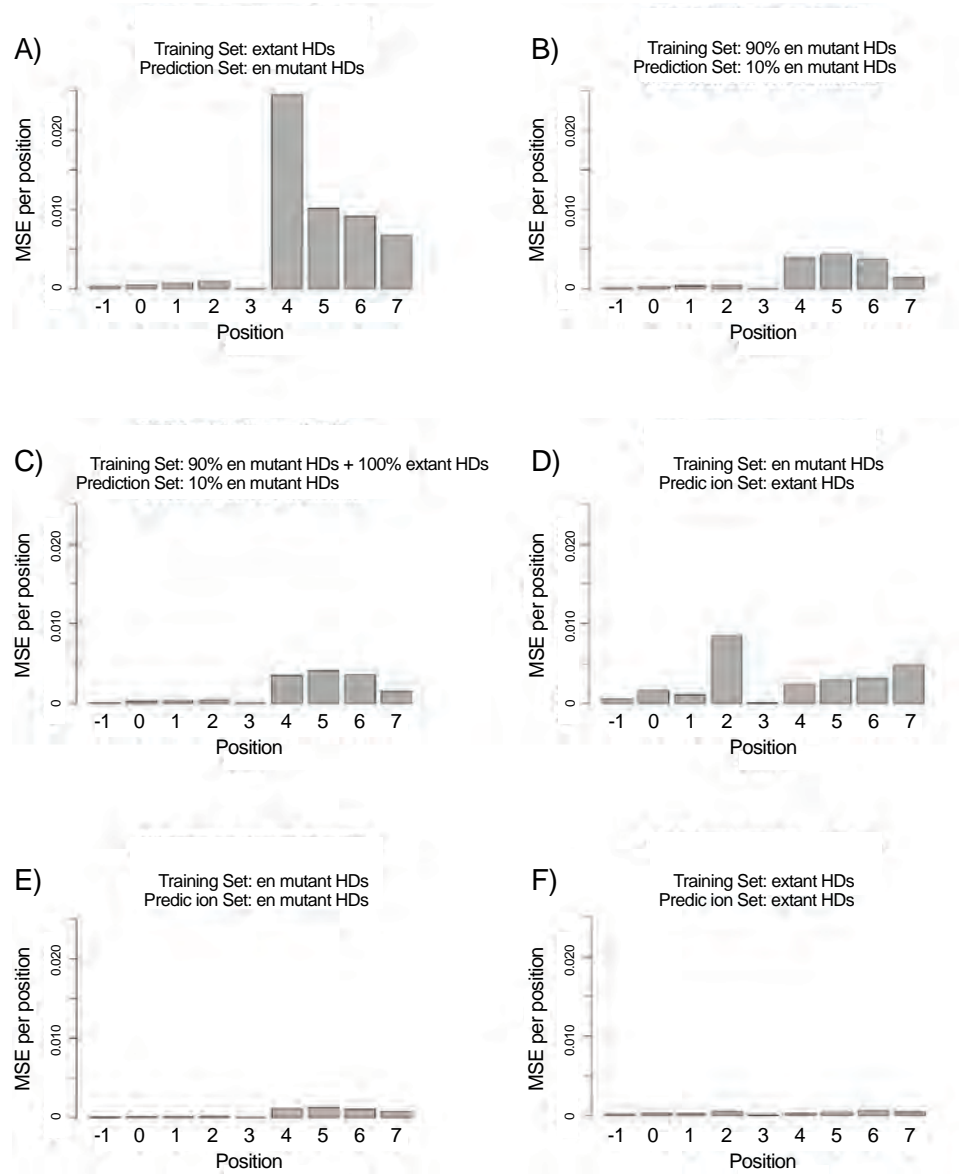
**Figure 2-16**



**Figure 2-16:** Limited diversity at the key recognition positions is observed in extant HDs.

(A) Frequency logo displaying the diversity of residues (circled in red are the residues randomized in the HD library) at various positions in the N51-containing HDs of humans, mice, *D. rerio*, *C. elegans*, *D. melanogaster*, and *S. cerevisiae*. (B) Frequency logo representing the diversity of residues found in our selected HDs that were characterized using the ZF10 library.

**Figure 2-17**



**Figure 2-17:** MSE contribution per position in refined RF recognition models, where the panels correspond to the Trials in Appendix Table A-3.

of HD specificity within our mutant set (MSE = 0.014; Appendix Table A-3). To facilitate the prediction of HD specificity, we have constructed a website ([stormo.wustl.edu/PreMoTF.v2](http://stormo.wustl.edu/PreMoTF.v2)) that incorporates our improved recognition model. Users can enter the amino acid sequence of a protein containing one or more HDs, and the algorithm will extract each HD sequence and generate a predicted recognition motif and representative Position Frequency Matrix (PFM). When tested on mouse HDs the predicted PFMs were very similar to those obtained by analysis of PBM data using BEEML-PBM (Pabo and Sauer 1992). Using this model we have also populated a page that displays predicted recognition motifs for the majority of the human HDs to facilitate the use of this data in constructing transcription regulatory networks within the human genome (Appendix Table A-4).

## **DISCUSSION**

In this study we performed an unbiased assessment of the breadth of sequences that HDs can specify by selecting variants of Engrailed that would preferentially recognize each of the 64 possible TAANN binding sites. Using our selection system, we recovered HDs that preferentially recognized 44 of these sites (Figure 2-6); a dramatic increase in the diversity of described recognition sequences. Many of these new sequence preferences are mediated by novel 3' specificity determinants that are functional when incorporated into independent HD scaffolds (Figure 2-11, Figure 2-12, and Figure 2-13).

Consistent with prior studies on HDs, Mutual Information analysis demonstrates critical overlapping roles for the residues at positions 47, 50, and 54

for 3' base recognition. The overlap between these determinants may represent either direct or indirect effects, however at the level of individual subsites one determinant typically dominates base preference at a specific subsite position. For example, while strong covariation is observed between residues 47 and 54, and base 4 (Table 1), K54 is highly preferred for recognition of CYN subsites whereas the recovered residue at position 47 is more variable. The presence of a positively charged residue at positions 43 or 46 is anticorrelated over the entire dataset (Appendix Table A-5) suggesting that these residues tune the overall affinity of the HD by adjusting electrostatic interactions with the phosphodiester backbone. These and other positions may also be responsible for more subtle sequence preferences that have been observed in Protein Binding Microarray analysis of HD specificity (Berger et al. 2008) that potentially lead to discrimination of TFs between different binding sites of moderate affinity (Badis et al. 2009).

The diverse and potentially independent assortment of specificity determinants within our dataset provides a foundation for constructing more accurate predictive models for 3' DNA-recognition by HDs. While significant prior effort has been expended on characterizing HD recognition, the functionality of specific determinants at critical recognition positions has remained poorly defined and as a consequence past predictive models of HD-DNA recognition have relied on nearest-neighbor type analyses (Noyes et al. 2008a; Alleyne et al. 2009). These models perform poorly when trying to predict the specificity of our selected HDs, which likely results from a lack of amino acid diversity at the key determinant positions within their training sets (Figure 2-1). In the context of our improved

predictive models, we can predict 3' specificity of a representative set of extant HDs with reasonable accuracy (Appendix Table A-4) and a predictive model combining all of the available data provides superior performance in predicting HD specificity. Thus, selection-based interrogation of HD recognition can inform predictive models, much as it has for Cys<sub>2</sub>His<sub>2</sub> zinc finger proteins (Benos et al. 2002; Koshland 2002; Liu and Stormo 2008; Persikov et al. 2009; Persikov and Singh 2011).

Our ability to select HDs with radically different specificity from characterized extant HDs, where novel sets of specificity determinants are employed, raises questions as to why extant HDs appear to be constrained in their diversity at the key recognition positions? Naively, we expect nature to exploit the full recognition potential of this domain to make a variety of orthogonal regulators for the independent function in transcriptional regulatory networks. This characteristic is observed in the largest family of DNA-binding domains, Cys<sub>2</sub>His<sub>2</sub> zinc fingers (Emerson and Thomas 2009), where comparison of zinc finger proteins across the mouse and human genomes indicates that this family is rapidly evolving (Myers et al. 2010), which is presumably creating factors with novel specificities. This diversity in ZFP recognition potential is even manifest within the human population, where differences in the fingers present in PRDM9 and their resulting specificity leads to difference in the location of meiotic recombination hotspots in individuals (Baudat et al. 2010). In this regard ZFPs appear to be an outlier, as most other well-characterized families of DNA-binding domains (Deppmann et al. 2006; Wei et al. 2010; De Masi et al. 2011) -- like HDs -- display limited diversity in their core recognition motifs and the recognition residues that they employ. It is possible

that the recognition potential of these other families of DNA-binding domains are similarly constrained. For HDs, the source of the selective pressure limiting the employed diversity of recognition residues is unclear, but understanding its origin would provide insight into the fitness barriers that influence the evolution of novel transcriptional regulatory networks in organisms.

In many instances HDs function as complexes with other DNA-binding domains to exert their gene regulatory function (Mann et al. 2009). This aspect of recognition is critical for the biological function of many of these factors, where complex formation can alter that recognition preference of the component HDs. The most thoroughly characterized example of the influence of partner association on recognition is the Hox-Pbx heterodimer, where minor groove features play critical roles in defining sequence preference for this complex (Joshi et al. 2007; Slattery et al. 2011). In general, the role of residues within and neighboring the N-terminal arm in DNA recognition remain poorly defined, although there is evidence that sequence preference may be driven by complementarity to DNA sequence-dependent minor groove width (Slattery et al. 2011; Jinek et al. 2013). We have demonstrated that some of our selected HDs can tolerate changes that alter 5' sequence recognition, but the degree of crosstalk between the recognition residues in the 5' and 3' segments of the binding site remains poorly defined. A selection-based analysis of the recognition potential of the N-terminal arm could help to clarify the roles of individual positions in minor groove recognition.

Our archive might present an opportunity to employ these domains as components of artificial transcription factors or endonucleases. The area of

engineered DNA-binding domains has primarily been the purview of ZFPs (Urnov et al. 2010), however, efforts to engineer ZFPs to recognize a wide variety of target sites using public archives have been most successful for guanine-rich binding sites (Ramirez et al. 2008; Zhu et al. 2011). HDs provide potential utility in the recognition of A-T rich sequences, and in the context of zinc finger-HD chimeras (Pomerantz et al. 1995; Rivera et al. 1996) may have utility in expanding the sequences that be efficiently targeted by zinc finger-based artificial nucleases.

## **MATERIAL & METHODS**

**Construction of the homeodomain (HD) library.** A pB1H2ω2-12En (Noyes et al. 2008a) (pB1H2ω2-12En(SB)) construct was created with the following modifications to the original *engrailed* (*en*) sequence: restriction sites SacI & BamHI were installed for use with cassette mutagenesis of the recognition helix through introduction of a synonymous mutation at L38 and a T60G mutation, respectively (Appendix Table A-6). The randomized recognition helix was cloned into the SacI and BamHI sites of pB1H2ω2-12En(SB) by the direct ligation of the following phosphorylated and annealed three oligonucleotide: EN K55 library, EN Library 5p comp, and EN Library 3p comp (Appendix Table A-6). Following transformation into electrocompetent XL1Blue cells, the library was plated on 20 150mm 2xYT plates containing 100ug/ml carbenicillin and incubated at 37°C overnight. The recovered library size was  $1.3 \times 10^8$  where the theoretical library size,  $3 \times 10^7$ , was over sampled 3-4 fold.

**Design of the target binding sites for the selection of HDs.** The 64 target sites (GGCCG**nnnTTAGCTGGGCG**GGACG) for use with the HD Library selections were cloned between the NotI and EcoRI site in pH3U3 (Noyes et al. 2008b). The bold nnnTTA element is the reverse complement of the 6bp HD target site TAANNN, where the NNN represents each of the 64 possible 3bp combinations. The bold TGGGCG element is the Zif12 binding site, which is positioned 10bp upstream the -35 box.

**Bacterial-One Hybrid (B1H) selections with the HD library.** Each HD library/TAANNN selection in the B1H system was performed basically as previously described (Noyes et al. 2008b). For each selection at least  $1 \times 10^8$  dual transformants (of HD expression vector and binding site reporter vector into the selection strain) were plated on NM media supplemented with 1uM IPTG and 200uM uracil. The stringency of each selection was adjusted such that 1000-2000 colonies were recovered (Figure 2-3). About 24 colonies were initially sequenced to confirm the success of the HD selections. Subsequently, recovered HD library members were identified via Illumina sequencing. Surviving colonies from each selection were pooled and prepared for sequencing as previously described (Gupta et al. 2010). HD clones were amplified using a forward primer (CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTATGCTTGCCCTGTCGAGTCC) and reverse primer (CTTAATGCGCCGCTACAGGGC), where the forward primer incorporated the Illumina P2-adaptor sequence (bold). Each PCR product was then digested with either BamHI or XbaI for the ligation of barcoded P1 adaptors (Appendix Table A-7 & A-8) prior to Illumina library generation and sequencing.



**Mutual Information (MI) and other statistical data analysis.** The catalog of approximately 44,000 selected HDs identified by Illumina sequencing for the 64 target sites was used to calculate MI between the randomized positions within the HD and base positions 4, 5, and 6 in the DNA target site as previously described (Mahony et al. 2007). Significance was determined by calculating the MI for a set of randomly associated selected recognition helices to the 64 target sites performed one thousand times followed by a non-parametric test used to derive a null distribution where a p-value < 0.001 for each MI value was considered significant.

The two-sided Fisher Exact Test was applied to assess significant association between the positive charge status at position 43 and that at position 46 for HDs recovered for each of the 64 binding sites and all binding sites combined. The odds ratio and its 95% of confidence interval were computed for each triplet and combined using the `fisher_test` function based on conditional maximum likelihood estimation. These statistical analyses were performed using R, a system for statistical computation and graphics (Ihaka and Gentleman 1996). To adjust for multiple comparisons for the 64 binding sites, p-values were adjusted using B-H method (Eneameh et al. 2013a) where sites with adjusted p-value < 0.05 were considered significant.

**B1H selections of HD variants with the ZF10 library.** All HD variants characterized from the HD library selections were sequences that were directly isolated from colonies on the selection plates, either from direct isolation of individual clones or the reconstruction of variants identified by Illumina sequencing through the ligation of phosphorylated and annealed oligonucleotides into

pB1H2 $\omega$ 2-12En (Appendix Table A-9). Each ZF10 library/HD variant selection was performed as previously described (Noyes et al. 2008a) except that all selections were plated on NM media supplemented with 5mM 3-AT, 1uM IPTG, and 200uM uracil. Recovered ZF10 library members were identified via Illumina sequencing as previously described (Gupta et al. 2010) except that the initial PCR product was digested with either BamHI or NcoI for the ligation of barcoded P1 adaptors (Appendix Table A-7 and A-10). Overrepresented sequence motifs were identified using MEME (Bailey and Elkan 1994) from the top 1000 most frequently occurring unique sequences within the Illumina dataset except for the grafted HDs where the top 500 most frequently occurring unique sequences were used. Additional sequences were included in cases where they had the same of reads as the one-thousandth (or five-hundredth) sequence in the set. The input parameters used for MEME were zero or one motif per sequence (zoops), 4 bases as the width minimum, 10 bases as the width maximum, while all other parameters retained the program default settings. Recognition motifs for each HD were then constructed as previously described (Zhu et al. 2011) by weighting the number of reads for each sequence that comprise the most significant motif identified by MEME, where the number of sequences input for motif discovery and incorporated into each motif is reported in Supplementary Table 6

**Expression and purification of proteins.** Each HD variant was expressed in Rosetta2(DE3)pLysS cells as C-terminal fusions to a purification tag sequence consisting of a His-6 tag, maltose binding protein (MBP), and Tev protease cleavage site. Cells were lysed by sonication. Protein was purified from the lysates using

Amylose Resin (New England Biolabs) and then was eluted from the Amylose Resin in binding buffer without BSA and IGEPAL CA-630 (25mM NaCl, 10mM Tris-HCl pH 7.5, 0.1mM EDTA, 1mM DTT, and 5% glycerol) supplemented with 40mM Maltose. Protein concentrations were determined by absorbance at 280 nm. Single use aliquots of protein were stored at -80 prior to use.

**Preparation of binding sites for electrophoretic mobility shift assays (EMSAs).**

Duplex binding sites were prepared by annealing the top oligonucleotide (GGGCAGNNNNNNGGACG) and bottom oligonucleotide (GGCGTCCNNNNNNCTGC) (Invitrogen) for a given binding site in annealing buffer (10mM Tris-HCl, 50mM NaCl, and 1mM EDTA) to the final concentration of 40uM dsDNA, where the N<sub>6</sub> represents the 6bp-binding site used in a given EMSA. Initial single stranded oligonucleotide concentrations were determined by absorbance at 260nm. For detection, annealed oligonucleotides were radiolabeled with alpha-<sup>32</sup>P dCTP and Klenow (exo-) (New England Biolabs) followed by a MicroSpin G-25 column (GE Healthcare) purification.

**Determination of apparent dissociation constant via EMSAs.** Varying concentrations of a given purified HD variant were equilibrated with 40pM of labeled oligonucleotide in binding buffer (25mM NaCl, 10mM Tris-HCl pH 7.5, 0.1mM EDTA, 1mM DTT, 5% glycerol, 0.1mg/ml BSA, and 0.1% IGEPAL CA-630) for 4 hours at room temperature. Samples were loaded onto a 5% polyacrylamide gel without loading dye in 0.5X TBE buffer while running at 300V at 4°C. Gels were run for 40 minutes following loading. Gels were dried and then exposed on phosphoimaging plates for 8-72 hours. Plates were imaged using a Typhoon FLA

9000, and quantified using ImageGauge V4.22. The apparent equilibrium dissociation constants ( $K_{d,app}$ ) were determined using the modified Hill equation:

$$Y = m \left( \frac{[P_t]^h}{[K_{d,app}] + [P_t]^h} \right)$$

where Y is the fraction of bound DNA as determined by the ratio of the bound DNA band to the total (free + bound) bands,  $m$  is a normalization factor that represents  $Y_{max}$ ,  $[P]_t$  is the total protein concentration, and  $h$  is the Hill coefficient.

#### **Determination of apparent dissociation constant via competition binding**

**assays.** Competition assays were performed under the conditions described for the determination of apparent dissociation constant via EMSA except that varying concentrations of an unlabeled-annealed oligonucleotide were added to a subsaturating (70-90%) amount of a given purified HD variant and 40pM of labeled oligonucleotide prior to equilibration. The concentration of DNA that disrupts 50% of the bound labeled complex ( $IC_{50}$ ) was determined using a simplified sigmoidal dose-response curve (Ryder et al. 2008):

$$Y = \left( \frac{1}{1 + (IC_{50}/[C])^h} \right)$$

where Y is the fraction of bound DNA, C is the concentration of unlabeled competitor, and h is the Hill coefficient. The  $IC_{50}$  is then converted into the apparent equilibrium dissociation constant for the competitor ( $K_{c,app}$ ) using the Lin and Riggs equation (Lin and Riggs 1972):

$$K_{c,app} = \frac{2[K_{d,app}]IC_{50}}{2[P] - [R] - 2[K_{d,app}]}$$

where  $P$  is the purified HD variant concentration,  $R$  is the concentration of the labeled oligonucleotide, and  $K_{d,app}$  is the apparent equilibrium dissociation constant of the HD for the labeled oligonucleotide as measured by EMSA.

**Computational modeling of HD-DNA complexes.** Modeling of mutant homeodomain structures was performed with RosettaDNA, using the recently described flexible DNA protocol and scoring function (Yanover and Bradley 2011) (RosettaDNA executable and accompanying parameter sets kindly provided by Philip Bradley at the Fred Hutchinson Cancer Research Center). Starting with the structure of the DNA-bound engrailed Q50A homeodomain (Grant et al. 2000), 20 models were generated by RosettaDNA for each DNA-bound mutant homeodomain. Each model was minimized with flexible DNA backbone and bases, and side chain packing was performed for residues adjacent to the DNA major groove (residues 31, 43-44, 46-51, 53-55, 57-58 in the crystal structure). Extended side chain rotamer sets were used for buried residues having 15 neighbors within 10 Å (“-ex1 -ex2 -ex1aro::level 6 -extrachi\_cutoff 15”), while extra DNA rotamers were used to sample base flexibility (“-exdna::level 2”). DNA backbone flexibility was specified for the 6 base pair DNA target site plus 2 base pairs flanking each side of the site. For each mutant, the 20 models from RosettaDNA were rescored using DDNA, a knowledge-based energy potential developed to predict protein/DNA structures and binding affinities (Zhao et al. 2010), and the top DDNA score was used to select a structural model reflecting the anticipated interactions at the HD-DNA interface.

**Random Forest (RF) Predictive Modeling.** Protein and Position Frequency Matrix (PFM) alignments and relative scaling of the PFMs used as inputs for the construction of a RF model were performed as previously described (Christensen et al. 2012). RF regression was performed as described using the previously identified determinant positions (3, 6, 19, 47, 50, 54 and 55) identified from the adjusted Mutual Information assessment of the 264 characterized extant HDs described in our previous study (Christensen et al. 2012). Models to test the utility of the extant HD specificity data from 246 mouse and fruit fly HDs (Thomas and Chiang 2006; Berger et al. 2008; Noyes et al. 2008a; Noyes et al. 2008b) and the selected HDs in this study were trained as noted in Supplemental Table S9 where the evaluation incorporated 10-fold cross validation when the training set and prediction set overlapped. The reported MSE values reflect the MSE per motif parameter in the predicted motif (Christensen et al. 2012).

## **DATA ACCESS**

Illumina Data for the selected and characterized HDs has been deposited with GEO (GSE35806). A website ([stormo.wustl.edu/PreMoTF.v2](http://stormo.wustl.edu/PreMoTF.v2)) provides user access to the predictive model of HD specificity and predictions for all of the annotated HDs in the human genome.

**\* Online Processed Illumina Supplemental Tables** can be found at:

<http://genome.cshlp.org/content/22/10/1889/suppl/DC1>

**CHAPTER III**  
**INTRODUCTION TO GENOMIC TARGETING**

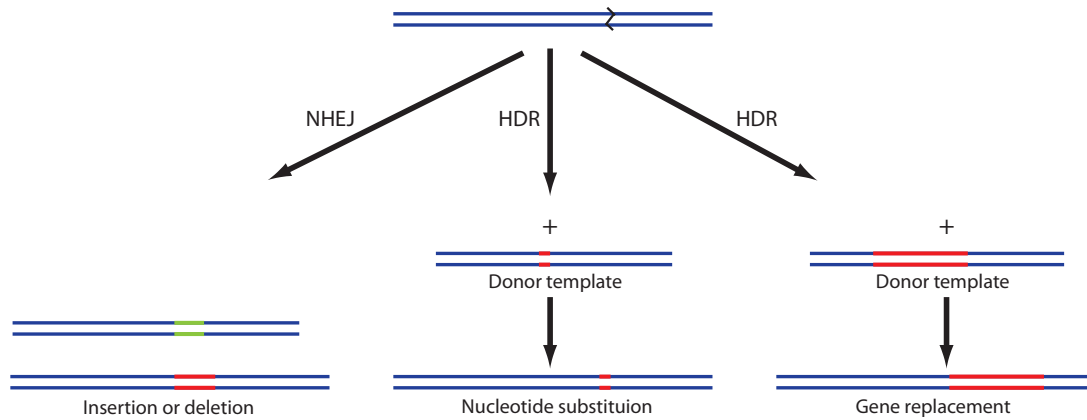
## **Advancing biology, biotechnology, and medicine through targeted genome editing and targeted gene regulation**

Precise targeted genome editing is a strategy that allows for the controlled modification of an organism's genome to change it for the needs of biology, medicine, and biotechnology. Targeted genome editing is induced by artificial nucleases that direct (Figure 3-1): 1) non-specific nucleotide insertions or deletions at a specific location, 2) a specific template-derived substitution to change a particular nucleotide, or 3) extensive template-derived alterations such as gene replacement or gene fusion for tagging that spans a large portion of a gene. The advent of tools for genome editing has proven to be effective in a variety of organisms, where the toolbox of engineered nuclease platforms include ZFNs (Carroll 2011), TALENs (Joung and Sander 2013), and most recently, Cas9/CRISPR (Jinek et al. 2012).

Genome editing can further both the fields of basic biology and biotechnology. In basic biology genome editing tool facilitate the interrogation of gene function in organisms that were previously less tractable or not amenable to other reverse genetics techniques. While making directed genetic changes have been established in yeast, bacteria, and mice, such techniques had not been established in most eukaryotes. For example, while the nematode, *C. elegans*, has been studied for the past forty years to allow for a vast increase of knowledge in biology (Brenner 1974), gene function has been typically studied through forward genetic techniques that induce random mutations. After random mutagenesis, to identify the genotype responsible for a given phenotype, tedious classical and



**Figure 3-1**



**Figure 3-1:** Possible modes of repair after a DSB is created in the genome by customizable sequence-directed endonucleases  
Artificial nucleases inducing a DSB results in three possible repair outcomes: 1) non-specific insertions or deletions at a specific location, 2) specific substitution to change a particular nucleotide, or 3) extensive alterations such as gene replacement or gene fusion for tagging that spans a large portion of a gene.

molecular genetic techniques must be performed. While some reverse genetic techniques are available, such as transient gene knockdown by RNA interference, they do not allow for germ line transmission that propagate progeny with heritable mutations (Zhuang and Hunter 2012). Only recently has gene editing been demonstrated in *C. elegans* by TALENs, ZFNs, and the Cas9/CRISPR system that induce site-specific mutations resulting in heritable germ line transmission (Wood et al. 2011; Friedland et al. 2013) as well as other organisms.

In biology, ZFN have been shown to function in fly, zebrafish, frog, mouse, rat, sea urchin, and hamster (Carroll 2011). While TALENs have been shown to function in fly, zebrafish, frog, and cricket (Joung and Sander 2013). Even the Cas9/CRISPR system has shown to function in fly (Gratz et al. 2013), zebrafish (Hwang et al. 2013), and mice (Hwang et al. 2013). Moreover, genome editing in biotechnology of plants and livestock can enable improvements in food production and biofuels by increasing yield and robustness, decreasing pesticide use, and increasing efficiency of creating genetic modifications in livestock with long reproductive cycles. Similar to the advantages for biology, direct manipulation by genome editing is less time consuming than other techniques typically used in biotechnology such as genetic crossing and random mutagenesis. To date, various organism of agriculture relevance have been genetically manipulated by ZFNs (tobacco, maize, and pig) and TALENs (pig, cow, silkworm, and rice).

Genome editing impacts medicine in both disease modeling and therapeutics. Target genome manipulation in somatic cell lines can further disease modeling, such as in mammalian cell lines. Furthermore, the therapeutic potential of genome

editing is significant. Instead of treating symptoms of a given genetic disease, the genetic defect that the disease arises from can be corrected, thereby curing the disease all together. ZFNs have targeted genes in humans for therapeutic value and TALEN hold similar potential (Perez-Pinera et al. 2012; Wirt and Porteus 2012). For example, ZFNs have enabled gene correction in the IL2Rgamma gene providing a potential treatment of X-SCID (Urnov et al. 2005) as well as disrupt the CCR5 gene to prevent the entry of the HIV virus into CD4+ T-cells (Holt et al. 2010), the later of which is currently in phase II clinical trials (Sb-728). So far, many uses in cell lines have been demonstrated, while therapeutics appear more challenging.

In addition to gene editing, targeted gene regulation has important implication in medicine and biology. Precise control of the regulation of gene expression is a quality necessary for the use of artificial transcription factors as therapeutic agent. Targeted gene regulation in organisms and cell lines can provide a means to control cellular processes for their study. Thus tools to edit the genome and target gene regulation are vital to furthering many fields of science.

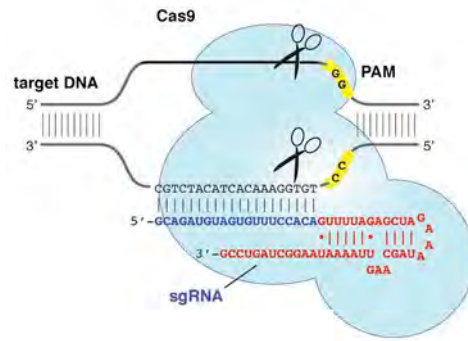
### **Tools to target specific genomic sites**

To create tools to manipulate the outcome of a cellular process through regulating gene expression or by editing a genomic location, the function of a particular protein or enzyme must be directed to a given genomic site precisely and predictably. Current tools using a DBD or, more recently, a small guide RNA (sgRNA) can direct an enzyme, such as an endonuclease, to edit the genome or a effector domain, such as an activation or repression domain, to regulate gene

expression. The sgRNA can be designed and synthesized to direct the endonuclease to a specific DNA address and thus is potentially easier to reprogram than DBDs. The sgRNA is an artificial component of the bacterial Cas9/CRISPR system that targets the Cas9 endonuclease to a specific genomic sequence (Figure 3-2)(Jinek et al. 2012; Jinek et al. 2013; Qi et al. 2013) science, doudna 2013 elife, lim wa 2013 cell). Since the specificity of Cas9/CRISPR system can be simply modified by changing the sgRNA, it may be the method for quick reverse genetics in basic biology where off-target effects is of less importance since this system appears to have high off-target effects (Fu et al. 2013) While this technology is gaining traction, it is in its infancy and requires much further exploration, thus details of this system will not be further discussed in this chapter. To date, DBDs have been the most utilized method for directing chimeric proteins to a particular genomic target.

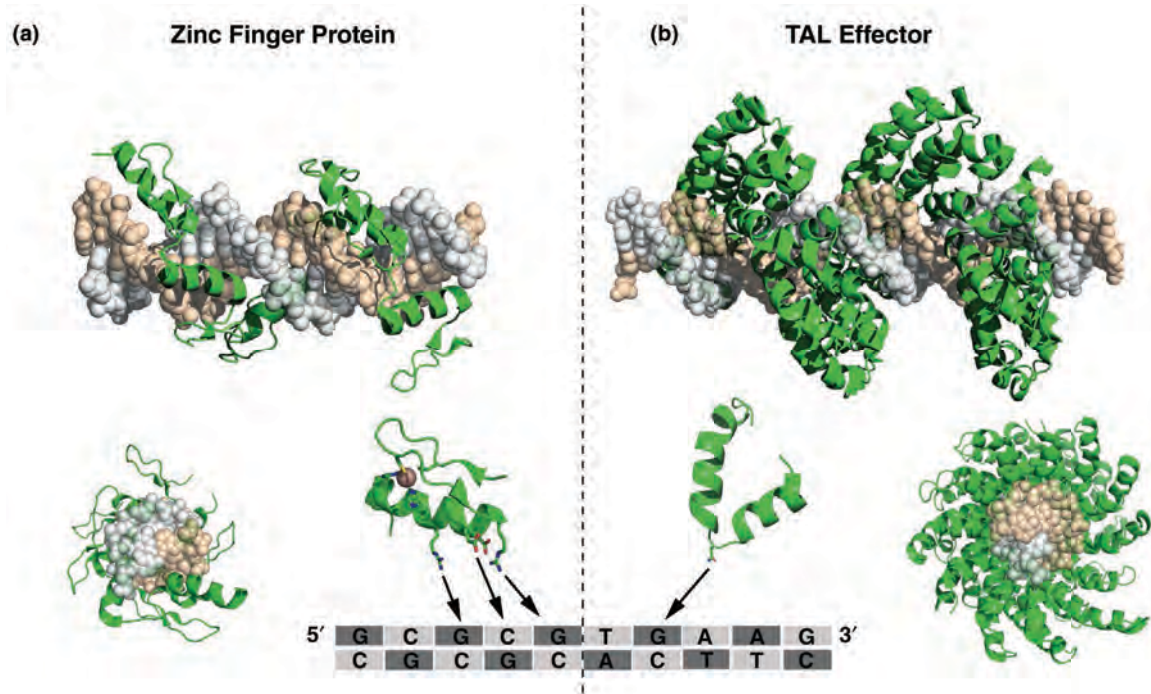
The two most commonly utilized DBDs are ZFs and transcription activator-like effectors (TALEs), as these two domains can be programmed to recognize a variety of different DNA sequences (Figure 3-3). Each of these DBDs has been incorporated within artificial transcription factors to alter gene expression or customizable sequence-directed endonucleases for gene editing. For the past two decades ZFs have been extensively studied and reengineered to recognize a broad range of sites, where one ZF module recognizes a three to four base pair site (Klug 2010). ZFs, however, appear more suited to recognize guanine rich sequences and can display context-dependent modularity (Ramirez et al. 2008; Zhu et al. 2011). As TALEs have been rapidly characterized over the past four years, they have been found to be more modular in nature as compared to ZFs, as one TALE

**Figure 3-2**



**Figure 3-2:** Cas9/CRISPR utilizes the sgRNA to direct site-specific DNA cleavage. A single strand of synthetic RNA (the sgRNA) directs the bacterial Cas9 endonuclease subunit (blue) to a target DNA site to induce DSB to the DNA, where the protospacer adjacent motif (PAM) next to the complementary region of the target DNA is also necessary.

**Figure 3-3**



**Figure 3-3:** Structure of ZFP and TALE interacting with DNA and cartoon illustrating their mode of specificity (Perez-Pinera et al. 2012)  
(A) ZFP (green) contains six ZFs where each ZF consists of 30 amino acids and recognizes three base pairs of DNA (IP47). (B) TALE (green) contains 22 repeats where each repeat consists of 34 amino acids and recognize one nucleotide (3UGM).

module binds one nucleotide with minimal context dependent effects (Bogdanove and Voytas 2011). While TALEs are now being used more frequently as targeting domains for nucleases or regulators their DNA-binding properties have been less extensively studied than ZFs. For precise genomic targeting to occur, DBDs with a balance between affinity and specificity are necessary (Ptashne 1992), and it is currently not clear what artificial nuclease platform will be the most precise for gene therapy applications.

### **Targeted gene regulation by artificial transcription factors**

Multiple routes to direct the regulation of gene expression, either repression or activation, utilizing the DBDs discussed above have been published. To activate gene expression, ZFs and TALEs have been fused to different activation domains, including VP16, VP64, and p65. In order to repress gene expression, ZFs and TALEs have been fused to the KRAB domain or simply used to interfere with transcription.

The initial study demonstrating that a DBD could regulate site-specific gene expression showed that a tandem array of ZF modules, typically referred to as a zinc-finger protein (ZFP), could be fused to VP16 to activate reporter expression, while at the same time the ZFP alone was able to block transcription *in vivo* in cell culture (Choo et al. 1994). Since then, a multitude of other studies have demonstrated that DBDs can regulate gene expression in mammalian cells. Gene activation by ZFPs can be regulated not only through their fusion to VP16 (Choo et al. 1994; Liu et al. 2001) but also other activation domains such as p65 (Liu et al. 2001) or VP64 (Beerli et al. 1998). Similarly, TALEs have been fused the p65 or

VP64 activation domains to activate gene expression in mammalian cells (Zhang et al. 2011; Joung and Sander 2013). Several of these studies have also demonstrated that chromosomal location may impact *in vivo* gene activation and that targeting DNase hypersensitive sites may give greater activation *in vivo* (Liu et al. 2001; Rebar et al. 2002; Maeder et al. 2013). Moreover, ZFPs fused to VP16 have also been shown to direct gene activation in mice (Rebar et al. 2002). Alternatively, DBDs also have utility to repress transcription as demonstrated by the fusion of a ZFP to a KRAB domain, where using such a chimeric protein results in the inhibition of gene expression (Choo et al. 1994). While DBDs used to direct gene expression show promise as possible therapeutic and tools for basic biology, they have yet to demonstrate as much utility as customizable sequence-directed endonucleases.

### **Genome editing by customizable sequence-directed endonucleases**

Customizable sequence-directed endonucleases (hereafter referred to as artificial nucleases) are engineered nucleases used to direct site-specific DNA cleavage. They must requisitely function to: 1) target a specific DNA location, and 2) create a break in the DNA. As a result, most artificial nuclease consists of two parts, a DBD to recognize a DNA target and a nuclease domain to cleave DNA. Two exceptions are the Cas9/CRISPR system, as described above, and meganucleases. For meganucleases, the two requisite functional parts are not spatially separate and are encompassed in one large molecule (Silva et al. 2011). They, however, lack flexibility to recognize a broad range of sites and will not be further discussed. The two major types of artificial nucleases that are currently used are ZFNs and TALENs,

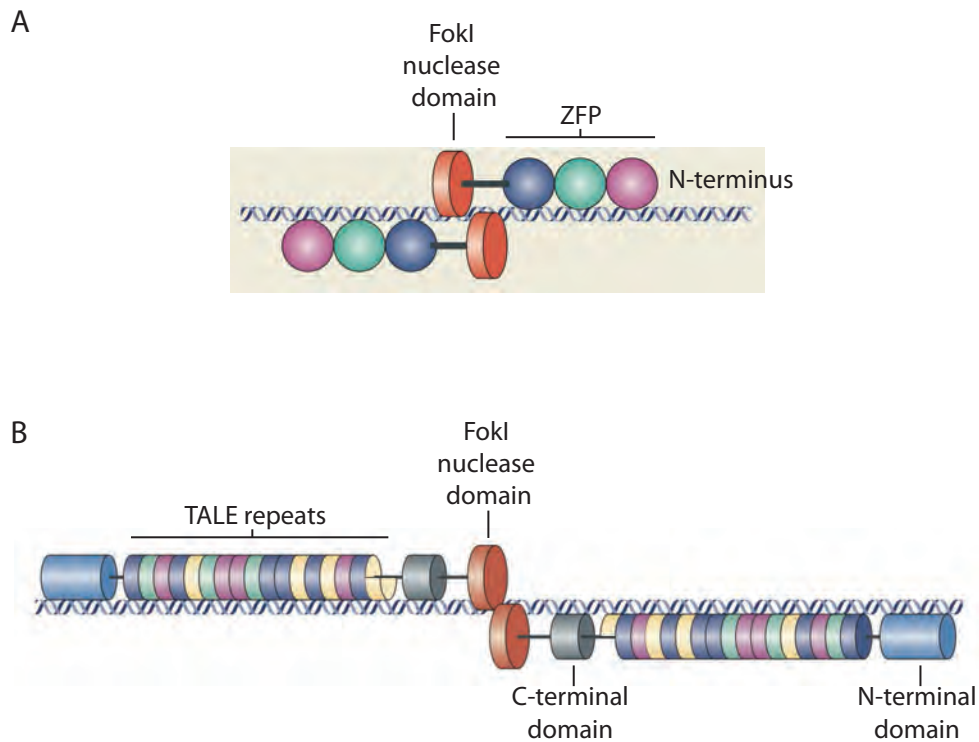


where they consist of a ZFP or TALE domains, respectively fused to the FokI nuclease domain (Figure 3-2) (Carroll 2011; Perez-Pinera et al. 2012).

The original artificial nuclease created the general architecture of the first ZFN in 1996, where the FokI domain, belonging to the type II restriction endonuclease FokI, is fused to the C-terminus of the ZFP (Kim et al. 1996). This basic scaffold is still currently in use. For ZFNs and TALENs to function, two monomers of the FokI nuclease domain must be in close enough proximity for the domain to dimerize in order to cleave DNA (Figure 3-4). A pair of ZFNs requires 6 bps between two monomer sites while a pair of TALENs requires 16 bps between two sites. Current artificial nucleases use engineered obligate heterodimeric versions of the FokI nuclease domain to decrease off-target activity by precluding homodimer formation via a single monomer (Miller et al. 2007; Szczeppek et al. 2007; Doyon et al. 2011).

Artificial nucleases induce site-specific DSBs that are then repaired by endogenous cellular mechanisms (Figure 3-1). It is the repair of the DSB, by either nonhomologous end-joining (NHEJ) or homology-directed repair (HDR), that leads to gene editing. The NHEJ repair pathway rejoins two broken ends together, and when this repair is imprecise, it can lead to the generation of small insertions or deletions (lesions) at the repair site (Wyman and Kanaar 2006). In this targeted mutagenesis, a lesion creating a frame shift mutation in a coding exon can thus create a truncated nonfunctional gene product or gene product destined for nonsense-mediated decay. HDR is a template-directed repair, where the template can be exogenously supplied to either change or replace a targeted sequence

**Figure 3-4**



**Figure 3-4:** Cartoon representation of the general architecture of ZFNs and TALENs (Joung and Sander 2013)

(A) Representation of a pair of ZFNs, where the ZFP of each ZFN contains three ZF (circle) and the FokI nuclease domain (orange disc) is C-terminally fused to the ZFP.

(B) Representation of a pair of TALENs containing 16.5 repeats (multicolored disc) and the necessary N-terminal and C-terminal domain of the TALE where the FokI nuclease domains is fused to the C-terminus of the TALE.

(Porteus and Baltimore 2003; Wyman and Kanaar 2006). Recently, inducing single-stranded nick by an artificial programmable nickase, where the cleavage activity of one FokI monomer is inactivated, has shown to restrict the DNA repair pathway to HDR, thereby decreasing the frequency of unwanted indels (Kim et al. 2012; Ramirez et al. 2012; Wang et al. 2012). While nickases show promise, increasing their gene correction frequency, which is lower than achieved by a DSB, is needed to show further utility.

ZFNs are the artificial nucleases most thoroughly studied, however, TALENs are rising rapidly in utility by building off the foundational work on ZFNs. Regardless, ZFNs, TALENs, and CRISPRs each have their advantages and disadvantages (Table 1). Characteristics inherent to the properties of the DBD for each nuclease system allows each system is best suited for a particular target or function. The established methods for ZFN assembly and thorough studies of ZFNs allows for them to be widely utilized in a variety of organisms. Several studies have characterized the off-target effects of ZFNs to understand their *in vivo* precision, however, similar studies have not been performed with TALENs (Gupta et al. 2011; Pattanayak et al. 2011). However, the ZFNs modularity is limited in that each finger can influence the specificity of the neighboring triplet, thus they are best suited to recognize triplets of GNNs (Ramirez et al. 2008). Moreover, each ZFP is typically able to specify from nine to twelve bps, where three to four ZFNs, respectively, are joined together through canonical linkers. TALENs complement these deficiencies of ZFNs in that TALEs can recognize up to sixteen bps and are more modular than ZFNs (Reyon et al. 2012a). TALENs have also been shown to be less cytotoxic than ZFNs

**Table 3-1.** The advantages and disadvantage of the different types of customizable sequence-directed endonucleases.

	<b>Advantages</b>	<b>Disadvantages</b>
<b>ZFNs</b>	<ul style="list-style-type: none"> <li>- Small DBD, 30 amino acids per 3 bp recognition</li> <li>- Off-target effects well defined</li> <li>- Safely used in clinical trials</li> </ul>	<ul style="list-style-type: none"> <li>- DBD has context dependency of GNN</li> <li>- Typical ZFP can recognize 12 bps</li> <li>- Less activity than TALENs</li> <li>- More cytotoxic than TALENs</li> </ul>
<b>TALENs</b>	<ul style="list-style-type: none"> <li>- Can theoretically recognize any site</li> <li>- Each TALE can recognize up to 16 bps or more</li> <li>- Higher activity than ZFNs</li> <li>- Less cytotoxic than ZFNs</li> </ul>	<ul style="list-style-type: none"> <li>- Has 5' base requirement of T</li> <li>- Large DBD, 34 amino acid per 1 bp recognition</li> <li>- Off-target less established than ZFNs</li> </ul>
<b>Cas9/CRISPR system</b>	<ul style="list-style-type: none"> <li>- Ease of target modification</li> <li>- Can theoretically recognize any site</li> </ul>	<ul style="list-style-type: none"> <li>- Limitations poorly defined due to recent development</li> <li>- High off-target rates</li> </ul>

and also have greater activity (Reyon et al. 2012b). Nonetheless, the established methodology of ZFNs in the artificial nuclease field has defined their off-target effect better than other nucleases and their safety has been demonstrated, thus far, in clinical trials. Moreover, the small size of the ZF over TALE, which needs to include its N-terminal and C-terminal domain, is an advantage for lentiviral delivery (Holkers et al. 2013). While Cas9/CRISPR system is new to the field they hold advantages over both TALENs and ZFNs. The ability of the Cas9/CRISPR system to target its site through a synthetic RNA allows quick synthesis of new targets and theoretically can target any given site. However, both the Cas9 endonuclease and RNA must be introduced into the target organism. Moreover, the system has been shown to have high off-target rates that may limit their use to organisms with quick generation times for outcrossing unintentional off-target mutations created (Fu et al. 2013).

Since initial studies demonstrating *in vivo* function of ZFNs in fruit flies and mammalian cells (Porteus and Baltimore 2003; Carroll et al. 2010) ZFNs and TALENs have facilitated targeted mutagenesis and gene replacement in a variety of organisms at different paces (Carroll 2011; Joung and Sander 2013). For example, targeted mutagenesis by ZFNs in zebrafish was first demonstrated five years ago (Doyon et al. 2008; Meng et al. 2008), however, only in the past year has homology mediated repair been demonstrated in zebrafish with exogenous donor DNAs utilizing TALENs (Bedell et al. 2012; Zu et al. 2013). The discrepancy in whether an artificial nuclease can function in a given organism maybe due to different cellular machinery of different organisms but is also dependent on the particular specificity

of the given DBD used. Limitations of artificial nucleases still remain that include the ability of DBDs to target a greater range of sequences and off-target effects of DSBs, where by they can create cytotoxicity to a cell or organism. More stringent DBD to increase the specificity of DNA-binding will alleviate such detrimental effects of artificial nucleases. Additionally, identifying new DBDs to complement ZFs and TALEs binding specificity can expand the sequences artificial nucleases can target.

### **Previous gene targeting utilizing homeodomains**

The original chimeric nuclease utilizing a HD-FokI fusion was created by the Chandrasegaran lab several years prior to the creation of ZFNs, however, it resulted in moderate non-specific cutting and such a chimeric molecule has not been revisited since (Kim and Chandrasegaran 1994). Attempts to further the utility of HDs resulted in the engineering of the ZFHD, where it was subsequently used as an artificial transcription factor to recognize a specific DNA site to drive gene activation for potential use as a therapeutic (Pomerantz et al. 1995; Magari et al. 1997). This, however, also has not resulted in further development of utilizing HDs in chimeric proteins. These experiments utilizing HDs were performed prior to our lab's expansion of HD specificity. Thus, the ability of the HD to target a broad range of site-specific genomic regions were limited by the inability of the HDs to be engineered to recognize a variety of different target sites.

### **Summary**

To test if HDs can be incorporated as the DBD in customizable sequence-directed nucleases to direct sequence specific gene editing, we developed a more stringent ZFHD chimeric framework than the original ZFHD. Thus, HDs can be used to complement the utility of ZFs to target sequence-directed nucleases. We set out to create a functional HD-containing nuclease, the nZFHD, where the nuclease domain is fused to the N-terminus of ZF. We incorporate the use of the HD modules our lab previously created to demonstrate the feasibility of utilizing HDs as a new DBD in artificial nucleases. To create an nZFHD, we optimized the linkers between the ZF and the HD, as well as the linker between the nuclease and the ZFHD. The functionality of this platform was demonstrated by creating targeted lesions in zebrafish embryos. Thus, nZFHDs can direct targeting in a complex genome.

**CHAPTER IV**

**UTILIZING ENGINEERED HOMEODOMAINS IN CUSTOMIZABLE SEQUENCE-  
SPECIFIC NUCLEASES**



## INTRODUCTION

Customizable sequence-directed nucleases, such as zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and, most recently, the Cas9/CRISPR system, are important tools to further biology, biotechnology, and medicine. These tools induce site specific DSBs, which enables repair by nonhomologous end-joining (NHEJ) or homology-directed repair (HDR) to precisely or imprecisely modify the target of interest (Carroll 2011; Jinek et al. 2012; Joung and Sander 2013). These nucleases have been used in a wide variety of organisms, including flies (Carroll et al. 2010), zebrafish (Meng et al. 2008), plants (Osborn et al. 2013), livestock (Carlson et al. 2012), cell culture, and humans (Porteus and Baltimore 2003; Jinek et al. 2013). While artificial nucleases have been studied over the past twenty years they are still somewhat constrained in the targets they can specify due to required recognition features for the DBD (ZFNs or TALENs) or guide RNA and PAM sequence (Cas9/CRISPR system).

HDs prefer to recognize AT-rich sites and thus complement the ZF preference for G-rich binding. Since HD variants from our prior study were selected in the context of a ZFHD fusion, HDs have demonstrated function as ZFHDs. While the original ZFHD was published in 1995 (Pomerantz et al. 1995), this construct has shown limited use in literature (Magari et al. 1997); since then further utility of this construct has not been demonstrated. While the ZFHD has demonstrated functionality as a DBD, it recognizes binding sites with limited stringency using a linker of GGRR between the ZF and HD. Furthermore, the linker used in previous studies by our lab utilizing the ZFHD construct with the linker TGTGR has been

shown to recognize sequences with different spacings between the ZF and HD.

Optimization of the linker for specificity and activity will improve the functionality of the ZFHD construct as a DBD (Noyes et al. 2008a; Chu et al. 2012).

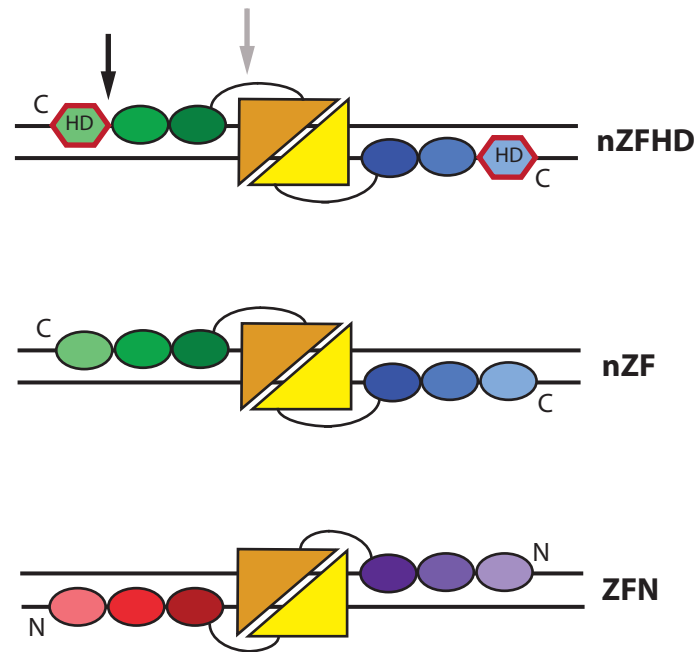
To incorporate a new DBD into customizable sequence-directed nuclease, we set out to utilize HDs, within the ZFHD construct, in artificial nucleases (termed the nZFHD). Here we create a functional nZFHD, where we optimize the linker between the ZF and HD to increase stringency and activity. In the commonly used customizable sequence-directed nuclease, the ZFN, the *FokI* nuclease domain is fused to the C-terminus of the zinc finger (ZF) of a ZFN (Figure 4-1). To fuse the ZFHD to the *FokI* nuclease domain, a different fusion point is necessary to create a nuclease incorporating ZFHDs. To this end, we identified a linker between the N-terminus of the ZFHD and the C-terminus of the nuclease domain to create a functional nZFHD. We subsequently show that this architecture is functional *in vivo*, in zebrafish, to create indels at a given target site.

## RESULTS

### Optimize linker selected between ZF and HD for specific DNA recognition

To create functional nZFHDs we first optimized the linker that fused the ZF to the HD to robustly recognize specific binding sites of various spaces between the ZF and HD. ZFHD modules to be used for specific genomic modification would require precise DNA-binding specificity (Figure 4-1 and Figure 4-2A). The linker previously utilized in our ZFHD (TGTGR) constructs did not constrain the ZF and HD to bind in

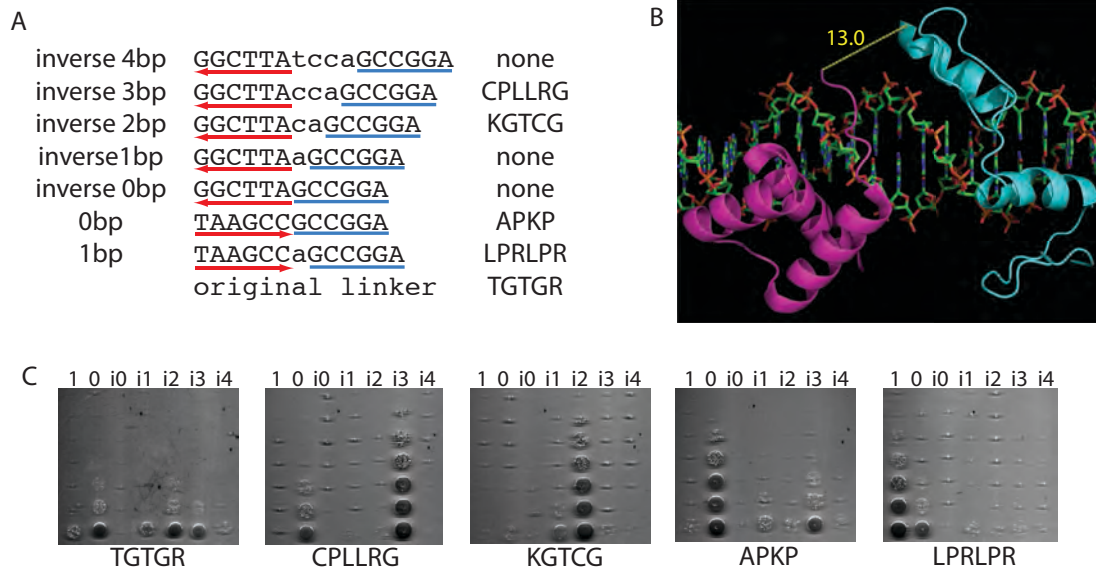
**Figure 4-1**



**Figure 4-1:** Schematic of nZFHD, nZF, and ZFN.

In the typical ZFN construct the FokI nuclease domain (triangles) is fused to the C-terminus of the ZFs (ovals). In an nZF the nuclease domain is fused to N-terminus of the ZFs. In an nZFHD the nuclease domain is fused similarly as the nZF, however, a HD (hexagon) follows the C-terminus of the two ZFs. The arrows represent the two linkers we optimize: one between the ZFs and the HD (black) and the other between the nuclease domain and ZF (grey). The N and C denotes the N and C-terminus of the DBDs.

**Figure 4-2**



**Figure 4-2: Optimization of the linker between the ZF and the HD**

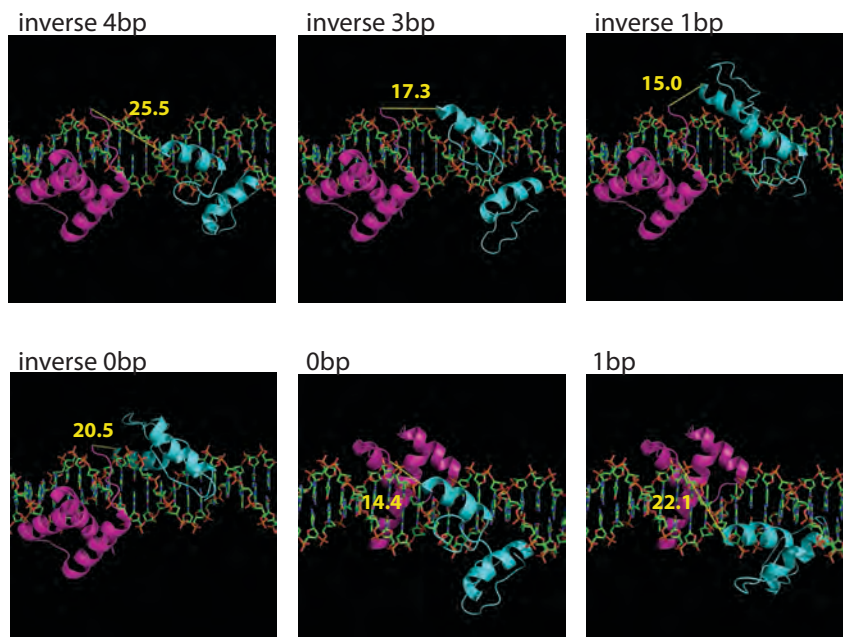
(A) The target DNA sequences containing different spacings and orientations of the ZF and HD binding sites, which we attempted to identify linkers to specifically recognize these spaces. The red arrow pointing to the left designates that the primary strand the HD binds to is on the bottom strand; while the red arrow pointing to the right designated that the primary strand binds to the displayed top strand. The blue line designates the ZF binding site. To the right of the binding site list the optimized linker identified. (B) The model created by superimposition of the HD (magenta) and ZF (turquoise) with an inverse 2 base pair binding site with an estimated 13 angstrom distance between the ZF and HD (yellow line). (C) Activity assays on selective media showing how the selected linkers compare alongside the original linker, TGTGR, and the different binding sites.

specific orientations relative to each other, which leads to an overall reduction in specificity (Figure 4-2C) (Noyes et al. 2008a). Moreover, the original linker (GGRR) fusing the ZF and HD together used by other labs had not been tested for relative activity in our system (Pomerantz et al. 1995; Magari et al. 1997)(Pomerantz et al. 1995, Ariad).

Molecular models created by structural superimposition of the ZF and HD at different spacings and orientations between the ZF and HD binding site on a DNA template allowed an estimation of the distance between the C-terminus of the zinc finger and the N-terminus of the HD. Based on this modeling, we estimated (where the approximation of an amino acid spanning maximally 3 angstroms was used) that the longest linker library (6 amino acids) could possibly span inverse four base pairs through one base pair of spacing (Figure 4-2B and Figure 4-3), where an inverse ZFHD site refers to the ZF and HD binding to its expected sequence on opposite strands. Using the models created, the range of the different linker lengths were estimated to span 13 angstroms for an inverse two base pairs between the ZF and the HD for the shortest observed length and 25.5 angstroms for an inverse four base pair between the ZF and HD for the longest length (Figure 4-3).

We calculated that linker libraries spanning one through six amino acids between the ZF and HD could be exhaustively or nearly exhaustively searched in the BIH system when each amino acid was encoded as NNS; the largest library consisting of the 6 amino acids linker containing  $1 \times 10^9$  possible members. The B1H system, which was also previously used to characterize the specificity of all homeodomains in *D. melanogaster* (Noyes et al. 2008a), allows for very large

**Figure 4-3**



**Figure 4-3:** Models with spacings between the ZF and HD  
Models created by superimposition with the various spacings we attempted to identify linkers for with the estimated distance between the ZF and the HD (yellow).

libraries, up to  $1 \times 10^9$  members in diversity, to be conveniently built into the system. Resulting growth on selective media implies an interaction between a given DNA-binding site and protein pair. BIH selections for each of the seven binding sites, from inverse 4bp through 1bp, were performed with each of the six libraries (Figure 4-2A). Based on its constrained length, the one amino acid library yielded negligible growth with any of the seven binding sites. Thus this library also served as a negative control throughout our experiments. Based on the results with this library, we deemed selections yielding over two hundred colonies as successful. For the inverse 4bp, inverse 1bp, and inverse 0bp binding site spacings, no viable linkers were identified. Constructs with high activity were identified for binding site spacings of inverse 3bp, inverse 2bp, 0bp, and 1bp. Individual surviving members were sequenced from successful selections to identify functional library members for each spacing between the ZF and HD (Table 4-1). Individual linkers deemed to represent the consensus of selected linkers for a particular binding site were characterized in comparison with the original linker, TGTGR, in the BIH system with the ZF10 randomized binding site library (Noyes et al. 2008a) (Figure 4-4). Based on this analysis, linkers with improved activity and specificity were identified for binding site spacings of inverse 3bp (CPLLRG), inverse 2bp (KGTCCG), 0bp (APKP), and 1bp (LPRLPR) as demonstrated by the recovered binding site from the ZF10 library and an activity analysis in comparison with the original linker (Figure 4-2A, 4-2C, and Figure 4-4).

#### **Linker identified between the nuclease and ZF to create a functional nZFHD**

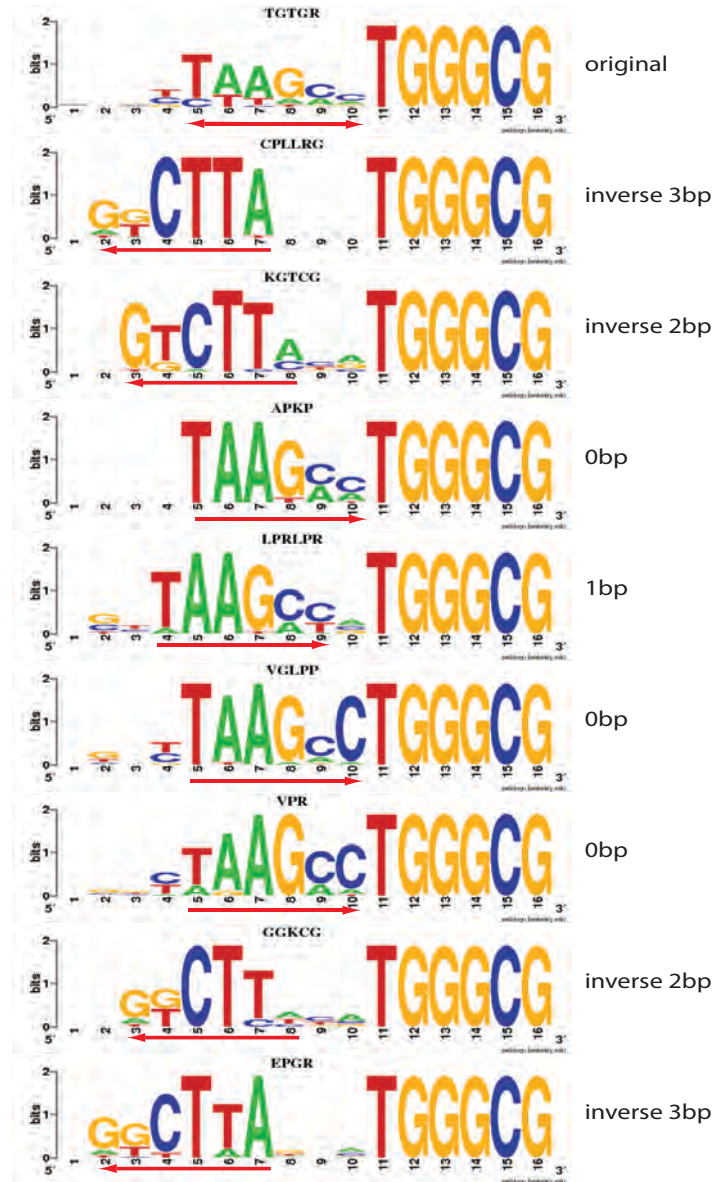
**Table 4-1. Linkers identified between the ZF and HD from B1H selections**

Spacing between the ZF and HD:	inverse 4bp	inverse 3bp	inverse 2bp	inverse 1bp	inverse 0bp	0bp	1bp
Number of Amino Acid in the linker library:							
2aa		FS NG FS SG SS GR FG					
3aa		MNT LQP MPS EPS ENT EPS FDR FNL VNL SPS NNP FNT FNL EAS				QPK VPR LPK QKR EQR SPR EPR QPR QPR QRR EKR LPR QYR SPR MPR LPR	
4aa		APEW DSNR LPGR VPNR DPDW AWRP APSW SWRP LPGR DPGR DPDR DPNP DPDR DPSR ASAG	ANGK KPGV RPGW EAGR ERYP EKYP VPGR RPGV RPGV ARNP MKYP VRNP VPGR SRFP VPGK			LPKP LPKP GPKP FAIS DPKP DPSR GPKP NKSG APRP RPKP CPKP MPKP RVQT WATP ELMA	



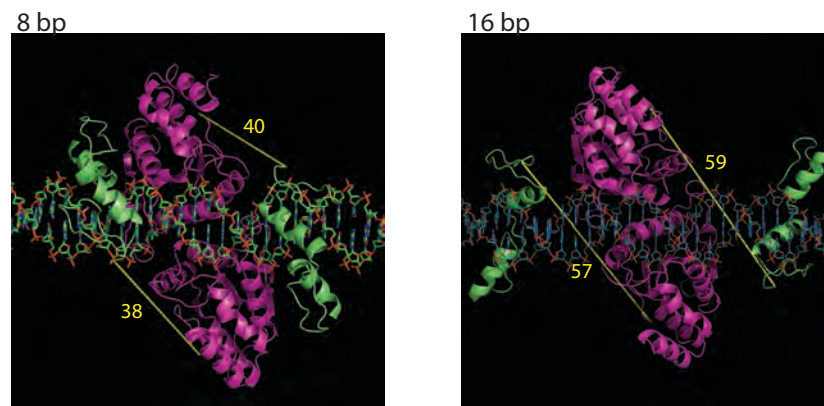
	EYNP	ERNP		LPNN
	SPNY	VPGH		FKLF
	APDT			LPNN
	EPGR			LPNN
	APGR			LPNN
	SPER			APRN
	SPQQ			LPKP
	SPHW			SPKP
	EPNR			APKP
	DVGR			
5aa	TRPAG	GGKCA	VGLPP	
	FPTNS	AGKCA	VGLPE	
	APWTG	AGKCA	GGRPH	
	SPIRS	RGTCA	VGLPA	
	APSTE	GGKCG	VGLPE	
	APSTV	RGQCG		
	DPAAG	KGRCG		
	FRLPG	KGTCG		
	APTTV	GGLCA		
	CDWFA	GGACA		
	FWRPG	GGKCG		
	MRRPG	AGLCG		
	LDWMG	KGACG		
	SRPVG	AGTCG		
		KGSCG		
		AGRCH		
		EGKCA		
		RGSCG		
		GGHCG		
6aa	CPLLRG	PCGECG	YCGRSG	LPRLPA
	CPLLRG	APRLGP	HESPGQ	LPRVKK
	CPILRG	PDGACA	DLGPLK	LPKVKK
	CPALRG	PAGSCA	SRPGWK	LPRLRR
	CPLLRG	PHGSCR	SCWPGK	LPRPRR
	ALRGQG	PNGACV	ESGTWK	LPKVKK
	CPLLRG	PRGECG		LPRLPR
	CPLLRG	PCGECR		LPRLPR
	CPLLRG	PAGXCK		LPRLPT
	CPMLRG	PNGVCL		SPRLDG
	EARARG	PAGSCR		APRNWG
	TPEWRS	GPSPLP		LPPIAHG
	EAHRRG	PGGSCA		APRLSG
	SPQWRL	PNGECQ		EPRVLP
		PGGSCG		

**Figure 4-4**



**Figure 4-4:** Stringency of selected linkers between the ZF and HD  
Motifs of ZF10 selections for selected linker between the ZF and HD showing the binding specificity for a particular linker (right).

**Figure 4-5**

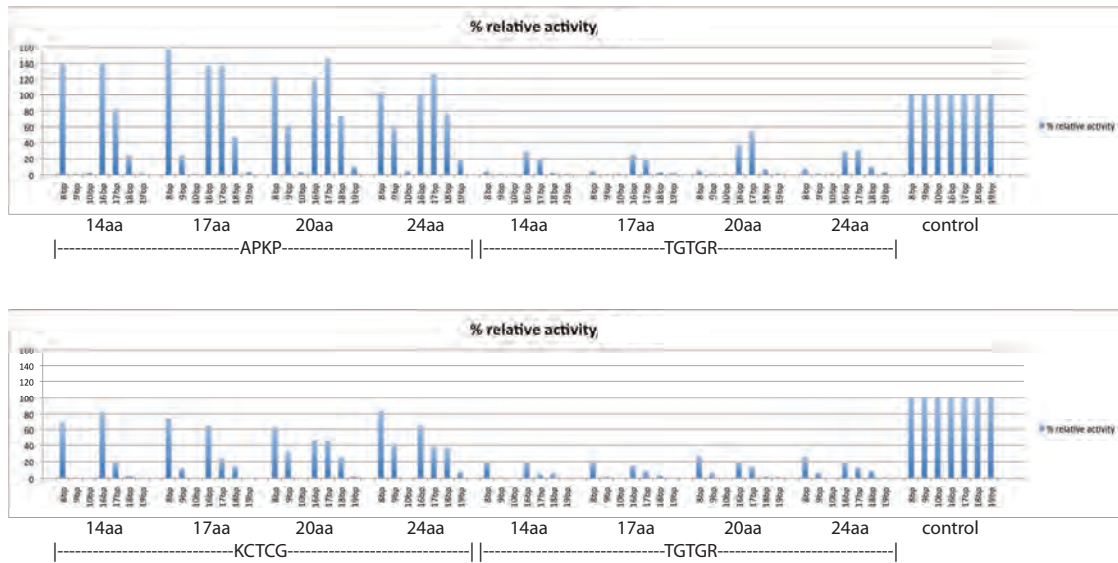


**Figure 4-5:** Models estimating the distance between fusing the nuclease to the N-terminus of the ZF  
The models created by superimposition to estimated the distance between the ZF and nuclease with 8 base pairs and 16 base pairs between the two ZF monomers for an N-terminal fusion to the ZFs.

Using superimposition, the necessary space between the two ZFHD monomer sites and the length of linker between the nuclease from the ZF were estimated to create a functional nZFHD (Figure 4-5). We oriented two ZFPs, Zif268, over double-stranded DNA with the N-terminus of each ZFP facing towards each other using superimposition (Elrod-Erickson et al. 1996). The *FokI* nuclease domain dimer, centered between the two ZFPs, was superimposed over the active site of the BamHI crystal structure as previously describe (Wah et al. 1998). A previous study examined the fusion of *FokI* nuclease domain to the C- terminus of the HD, but this displayed non specific activity (Kim and Chandrasegaran 1994), thus we chose to fuse the nuclease to N-terminus of the ZFHD to create a more specific chimeric nuclease. Our molecular models allowed estimation of the distance between the N-terminus of the zinc finger and the C-terminus of *FokI* in various orientations and spacings. The most favorable spacing appeared to be 8bp and 16bp between the ZF and *FokI* binding sites (40 and 60 angstroms apart, respectively), which provided a starting point of possible spaces to test between the two monomer sites.

Linker lengths of 14 through 24 amino acids, composed of alanine, serine, and glycine, were fused to optimized ZFHDs, with either the APKP or KCTCG linker between the ZF and HD, to test for activity with spacing of 8, 9, 10, 16, 17, 18, or 19 base pairs between the two monomer sites (Figure 4-6 and Table 4-2). These nZFHD constructs with various spacings were tested using a yeast-based chromosomal reporter assay (Ryan et al. 1998). Consistent results for the two different linkers between the ZF and HD identified the shortest linker tested, 14 amino acids, to be

**Figure 4-6**



**Figure 4-6:** Yeast activity assay showing relative activity for nZFHDs. Each nZFHD either contains the APKP or KCTCG linker for nZFHD binding sites of 0 base pair or inverse 2 base pair, respectively, alongside the original linker, TGTGR. Various linker lengths between the nuclease domain and the ZFHD were tested against different spaces between the two ZFHD binding sites as compared to an internal ZFN control.

**Table 4-2.** Sequences tested in the yeast reporter assay

Linker lengths test between nuclease and ZF

14aa ASGGGSGSSGSGGA  
17aa ASGGGSGASGSGSGGGA  
20aa ASGGGSGASGSGAGSSGGGA  
23aa ASGGGSGAGSGSGAGSGSGSGGA

Spaces between 2 ZF-0bp-HD sites (bolded), the same space is used between 2 ZF-inverse2bp-HD sites

8bp **TAAGCCTGGGCGGCGCTCACCGCCCAGGCTTA**  
9bp **TAAGCCTGGGCGGCGCATCACCGCCCAGGCTTA**  
10bp **TAAGCCTGGGCGGCGCATCAaCCGCCCAGGCTTA**  
16bp **TAAGCCTGGGCGGCGCATCAgcgcatcCCGCCCAGGCTTA**  
17bp **TAAGCCTGGGCGGCGCATCatgcgcatcCCGCCCAGGCTTA**  
18bp **TAAGCCTGGGCGGCGCATCAttgcgcatcCCGCCCAGGCTTA**  
19bp **TAAGCCTGGGCGGCGCATCAttgcgtcatcCCGCCCAGGCTTA**

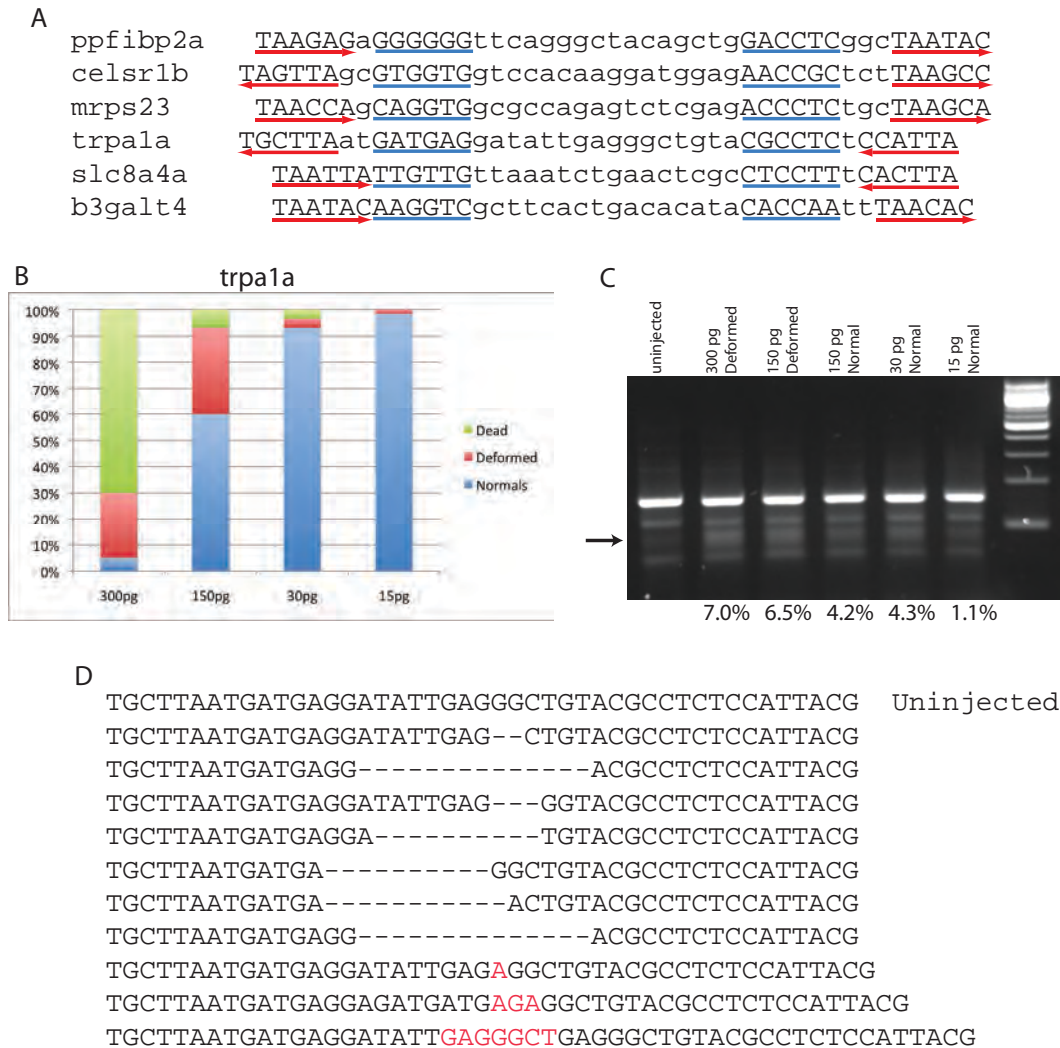
the most stringent in activity to the 8 and 16 base pair spacing, with high activity comparable to an internal ZFN positive control.

### **Gene disruption in zebrafish created by an nZFHD**

To validate that the nZFHD platform is functional *in vivo* we sought to target six genes in zebrafish to create insertions or deletions (indels) within the targeted region. Each target site contained 16 base pairs between the two nZFHD monomer binding sites. The ZFHD sites tested a range of different spacings (inverse 3bp, inverse 2bp, 0bp or 1bp) and employed the linkers characterized earlier that are specific for a particular spacing (Figure 4-7A and Table 4-3). ZF and HD modules used in the construction of each nZFHD are derived from our previously published archives (Zhu et al. 2011; Gupta et al. 2012). The 14 amino acid linker between the nuclease and the ZFHD was used in all nZFHDs.

mRNA for each pair of nZFHDs was transcribed and injected into one-cell stage embryos. In *trpa1a*, lesions were inferred by the measurement of toxicity using a dose response curve, where the ratio of normal morphology, deformed morphology, or dead embryo at 24 h.p.f. was calculated (Meng et al. 2008). The other five targets did not yield dose response curves indicative of possible lesions. Of the embryos injected with nZFHDs targeting *trpa1a*, genomic DNA was isolated from pools of embryos with either visually deformed or normal morphology. Each pool of DNA was followed by T7E1 digestion of the target locus then ran on a gel to visualize mismatches (indels) at the target created by injecting the embryos with nZFHDs. At this locus, as the dose of nZFHD increased, the frequency of lesions also

**Figure 4-7**



**Figure 4-7: Utilizing nZFHDs in zebrafish**

(A) The six different zebrafish gene targets nZFHDs were tested on. (B) Dose response curves for varying doses of *trpa1a* mRNA showing phenotypes associated with increasing doses of injected nZFHDs (C) T7E1 digestion of pooled embryos of different dosages showing potential lesions created by the *trpa1a* pair of nZFHDs. Percentages below each lane are the lesion rates measured for each pool of embryos (D) Lesions observed in morphologically normal embryos injected with 150pg of both nZFHD mRNAs. Dashes or red letters indicate the positions of deletions or insertions, respectively.



**Table 4-3.** nZFHD constructs used for zebrafish. Amino acid sequences of the recognition helix of the ZFs or residues 43, 46, 47, 50, and 54 of the HDs used for the nZFHD, and the space for the between the ZF and HD in the target gene for a given nZFHD.

Left 5p ZFHD				
Gene Name:	finger 1	finger 2	space	homeodomain
ppfibp2a	RSDHLTR	RSDHLTR	1bp	VMRWY
celsrlb	RSDALTR	RSDALTR	inverse 2bp	LHYAK
mrps23	RSDALTR	RSDNLSE	1bp	MKYEK
trpala	RSDNLTR	LSFNLTR	inverse 2bp	RHDRA
slc8a4a	RSDLKA	RSDALRK	0bp	KRLAA
b3galt4	DRSALAR	RSDNLQ	0bp	QRISV
Right 3p ZFHD				
Gene Name:	finger 1	finger 2	space	homeodomain
ppfibp2a	DRSALAR	RSDNLTR	inverse 3bp	QRISV
celsrlb	YRQSLTR	RSDDLTR	inverse 3bp	VRLLKY
mrps23	LAHHLTR	RSDNLTR	inverse 3bp	RHDRA
trpala	RSDDLTR	RSDNLTR	1bp	KTTQD
slc8a4a	RSDNLTR	RSDNLQ	1bp	HLIQY
b3galt4	RSDALTR	RSDALRK	inverse 2bp	LGMRR

increased as observed in the T7E1 assay, where embryos injected with the range of 15 pg of mRNA to 300 pg mRNA resulting in 1.1 percent to 6.7 percent lesion frequency, respectively (Figure 4-7B and 4-7C).

Types of mutations generated by nZFHD at 150 pg, normal embryos, were identified by cloning the treated target site into the vector of the LacZalpha blue-white assay (Zhu et al. 2013). A frame-shift of a created lesion in the embryo is identified by white colonies amongst a background of blue wild-type colonies in this assay. Thus, we identified short indels created around the nZFHD binding site indicating that the nZFHDs did create a double-stranded break at the target region to allow for imprecise repair at the target loci (Figure 4-7D).

## **DISCUSSION**

In this study we were able to create sequence-directed genomic lesions in zebrafish by using the ZFHD nuclease platform that we engineered. This achievement required the optimization of the linker joining the ZF and HD modules and the creation of a functional N-terminal fusion of the *FokI* nuclease domain to this DNA-binding platform. Guided by molecular modeling of the ZF and HD modules on idealized B-DNA, we designed linker libraries to span different spacings and orientation between the ZF and HD modules. We selected linkers that defined preferences for particular binding relationships between the ZF and HD modules and these linkers also displayed higher activity within the B1H system implying that they may also have improved affinity. Additionally, we identified a linker between the *FokI* nuclease domain and ZFHD to create a functional nZFHD at a particular

spacings between the monomer binding sites. Moreover, the selected linkers between the ZF and HD also demonstrated higher activity in a yeast based activity assay as compared to the original linker joining these modules.

The four different linkers identified for the four different orientations between the ZF and HD shows that a linker joining two DBD for a fixed spacing between two DNA binding sites can be identified. While linkers of approximate spacing have been identified and utilized in ZFs and ZFNs in previous studies (Moore et al. 2001; Soldner et al. 2011), artificial linkers joining two DBDs with good stringency for a specific spacing and orientation between two DBDs as we have shown here have not been previously published. This study implies that it is possible to identify linkers between other DBDs to create a larger toolbox for what can be recognized by chimeric DBD combinations to be used in artificial nucleases or transcription factors. The selections performed here can also be used for other DBDs, such as ZFs, to expand their flexibility of specificity. Moreover, the linker between the nuclease and ZFHD can potentially be further optimized to function with higher stringency of spacing between the two ZFHD binding sites.

Using HDs in ZFHDs complement the number of sequences that can be recognized by ZFs because HDs prefer to recognize AT-rich sites while ZFs recognize G-rich sites. By increasing the number of different orientations between a ZF and HD by three more from the original ZFHD construct, we have increased the number of targetable 6 bp binding site from four percent to twenty percent of all possible 6 bps sites. Taking into account that the 5' of the HD can recognize TGA, TAA, TTA, CAA, and CGA (triplets that ZFs can not target), this is twenty percent is in addition

to the possible 6 bp sites that can be recognized by our single fingers ZF modules (18 percent) (Zhu et al. 2011). Additionally the current set of single finger modules from our lab cannot recognize the five 5' triplet sites recognized by HDs, which highlights the complementarity between these module sets.

While only one of the six nZFHDs we created for zebrafish resulted in lesions, their failure to function may not be due to the architecture of the nZFHDs themselves. Properties inherent to *in vivo* genomic DNA can affect nuclease targeting of endogenous sequences, such as chromatin architecture and DNA methylation, which can hinder the access to or recognition of the target site (Reyon et al. 2012a; Valton et al. 2012). To evaluate if the five nZFHD are not functional due to properties inherent to the *in vivo* system, they may be tested in a system outside of the zebrafish, such as the yeast-based nuclease assay (Zhu et al. 2011). Moreover, the specificity of each ZFHD created for zebrafish can be tested in the B1H system to test the true specificity of each ZFHD once it has been assembled. This has been performed for ZFs in ZFNs where assembling finger modules may result in unexpected specificity within the entire array (Gupta et al. 2010; Zhu et al. 2011).

By selecting for stringent linkers between the ZF and HD and identifying a functional linker between the nuclease and ZFHD we successfully used HDs to create site-specific lesions in the complex genome of the zebrafish. By doing so, the expanded sites ZFHDs can recognize will complement the limitations of ZFNs, where both ZFs and ZFHDs have advantages over TALENs due to their small size to be incorporated into gene delivery vectors (Holkers et al. 2013). The advantage over CRISPRs by nZFHDs and ZFNs of being functionally encompassed by one molecule

may further provide advantageous to simplify delivery of a nuclease to a given system. Moreover, studies have CRISPRs have shown they have high off-target effects (Fu et al. 2013). While this study demonstrates that HDs can be used as DBDs in ZFHD, it is possible that HDs can be developed as stand alone DBDs, either as individual HDs or tandem HDs fused to an effector domain, for use in gene regulation or genome editing. Further engineering to expand the HD binding specificity at either the 5' or 3' specificity (Chu et al. 2012) will also expand the utility of the ZFHD. Moreover, ZFHDs have additional utility to be used as DBD for artificial transcription factor. Thus ZFHDs have broad utility for genome editing and targeted gene regulation in organisms of biology, biotechnology, and therapeutics.

## **MATERIAL AND METHODS**

### **Superimposition modeling of ZFHDs**

ZFHDs were built by superimposing the ZFs, finger 1 and 2 of zif268 (Elrod-Erickson et al. 1996), and HD, Msx1 (Hovde et al. 2001) over the respective DNA binding sites, using Pymol over B-form DNA created by X3DNA (Lu and Olson 2003), with the different spaces between the ZF and HD sites. Measurements were then estimated using Pymol's measurement function.

### **BIH-linker selection**

Linker libraries of 1 through 6 amino acids between the last histidine of the two ZFs in a Zif268 finger 1 and 2 backbone and the glutamate at the beginning of the engrailed HD was encode as NNS for each amino acid in the p1352-omega-UV2. The recognition helix of ZF backbone variant was QKGHLTR for finger 1 and DRSDLTR

for finger 2 while the HD was VRLKY at positions 43, 46, 47, 50, and 54 in the previous published Engrailed variant (Noyes et al. 2008a). For each linker library selection with each binding site was oversample 3 times, with the exception of the 6 amino acid linker library, which was covered maximally to  $2 \times 10^8$  combinations due to the co-transformations efficiency of the US0 selection strain. Selections were plated on NM minimal medium selective plates lacking uracil and containing 25mM 3-AT as the competitor and grown at 37° for 60 to 120 hours. Up to 24 individual colonies for each successful selection were sequenced.

#### **BIH-binding site selection using the ZF10 library**

Selections characterizing the DNA-binding specificity of individual ZFHD linker clones were performed as previously described (Noyes et al. 2008a) except that all selections were plated on NM minimal medium selective plates with 5mM 3-AT, 1mM IPTG, and 200mM uracil then grown at 37° for 24 to 32 hours. 24 individual colonies for each selection were sequenced. The overrepresented sequence motif was determined with MEME (Bailey and Elkan 1995) and sequence logs created by Weblogo (Crooks et al. 2004).

#### **BIH-based activity assay**

Activity assay were performed as previously described (Noyes et al. 2008b). 10-fold serial dilutions were grown on NM minimal medium selective plates containing 10mM 3-AT, 1mM IPTG, and 200mM uracil then grown at 37° for 36 hours.

#### **Yeast-based nuclease assay**

The Mel1-based yeast activity assay (Ryan et al. 1998) was performed from the integration of the nZFHD target site to be tested with the ZFN positive control

through the modified ySSA vector. nZFHDs were cloned into pYLeu containing a wild-type *FokI* nuclease domain and a modified assay was performed as previously described (Gupta et al. 2012).

### **Zebrafish husbandry**

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

### **nZFHD mRNA injections and lesion analysis**

For targeting sites in zebrafish, the left 5' nZFHD and right 3' nZFHD was cloned into pCS2 vectors containing the DD and RR obligate heterodimer versions of the FokI nuclease domain, respectively (Szczeppek et al. 2007). pCS2-nZFHD constructs were linearized with NotI. The mRNA was transcribed, purified and injected as previously described (Zhu et al. 2013). Pools of 20 injected embryos were collected at 24 h.p.f. for a given dosage and phenotype to be assayed for lesions and lesion rate was calculated as previously described (Zhu et al. 2013).

### **LacZalpha blue-white assay**

To identify the types of lesions created in zebrafish, the targeted genomic regions were cloned in pBluescript-KS(-) vector and assayed for indels as previously described (Zhu et al. 2013).

**CHAPTER V**  
**DESCRIPTION OF METHODS**



## Bacterial-One Hybrid System For Selections

The Bacterial-One Hybrid (B1H) system can be utilized to determine the DNA-binding specificity of a DNA-binding domain (DBD) or identify a DBD that binds to a particular DNA sequence (Noyes et al. 2008b). The B1H system consists of a plasmid with the DBD fused to the omega-subunit of RNA polymerase (bait) and a second plasmid with the target site upstream of two reporter genes, *HIS3* and *URA3* (prey). To identify DNA-binding specificity of a DBD the library members are contained within the prey in the form of a randomized binding site library, while to identify a DBD specific for a target site the library members are contained within the bait in the form of a randomized DBD library. Both the bait and the prey plasmids are transformed into a bacterial selection strain with bacterial homologs, *hisB* and *pyrF*, of the reporter genes contained on the prey plasmid deleted. Plasmid concentration in the transformation of the selection strain are titrated to minimize the opportunity that more than one library member will be transformed into a cell. If the bait interacts with the prey, the recruited RNA-polymerase will activate the transcription of the reporter genes to allow for growth on selective media. Colonies that grow on the selective media can then be isolated and sequenced to identify the library member(s) at either the bait or prey level.

The advantages of the B1H system include: 1) It is a quick method that only requires a single round of selection and does not require protein purification. This is unlike other techniques such as protein-binding microarray where protein purification is necessary or SELEX where several rounds of enrichment and protein purification are necessary; 2) It allows for a large number of library members (1 x

10<sup>9</sup>) to be searched since the transformation efficiency of bacterial cells are higher than other cell types such as yeast hybrid systems; 3) Library members can be at either the target site level or the DBD level; 4) The genome acts as competitor DNA to prevent less specific DNA-protein interactions from being captured.

The disadvantages of the B1H system include: 1) It is necessary for the interaction of the DBD with target to be in the dynamic range of the system to recover a library member, as low affinity binders may not be identified or low affinity interaction must be fused to another DBD to increase overall affinity. 2) While this system allows for a large number of library members, the searchable library size is limited by the transformation efficiency of the selection strain. 3) Often different conditions (stringency of selective media) may need to be optimized to allow for growth of colonies to identify interactions.

### **Identifying Target Sites or DBDs From the B1H System**

Sanger sequencing or Illumina sequencing is used to identify the selected library members recovered from the B1H system. Regions of the plasmid containing the randomized region (either DBD or target site) are PCR amplified for both types of sequencing. For Sanger sequencing, individual colonies are used for PCR amplification, where multiple colonies are screened. For Illumina sequencing, the plasmids from an entire selection plate of surviving colonies are pooled together and the randomized region is PCR amplified. Each pool amplification is prepared with a barcoded adapter to differentiate individual selection plates. Pooled barcoded amplicons can then be submitted for an Illumina sequencing run.

Overrepresented target motifs for the target sites were identified using MEME followed by alignment and visualization by WEBLOGO on a bit scale representing information content. DBDs identified were displayed through WEBLOGO as frequency logos to visualize overrepresented amino acids.

Sanger sequencing can identify long stretches of sequences, which was initially necessary for our HD library. Since the HD library has residues 43, 46, 47, 50, and 54 randomized, which spans 36 base pairs in length, Illumina sequencing at the time of the experiment did not span the necessary length. The readout length of 36 bases pairs would not cover all randomized residues in addition to the barcode of the adaptor at that time. Nonetheless, Illumina sequencing allows for the upwards of 30 million read, which can allow for over 50 different barcoded selections (for each different target site) to be run in one Illumina flowcell lane to read out greater than  $5 \times 10^5$  reads per barcoded sample. Sanger sequencing necessitates individual colonies be isolated, thus can be cumbersome if one is to look at a large number of library members. Since the initial sequencing of HD library member, Illumina sequencing has been expanded to 75-100 base pairs, thus allowing for Illumina sequencing of HD library members.

Using WEBLOGO, the frequency of the amino acids identified in the randomized region of the HD variant for each target site is displayed as frequency logo (Crooks et al. 2004). While frequency logos are used to identify amino acids that are overrepresented in the selections for HD variants, the logos can be misleading. Each amino acid is not equally represented in the HD library. (Each randomized residue position was randomized as NNS.) For example, Phe occurs

once while Ser occurs thrice in NNS, thus if a residue has a higher frequency in the logo it does not necessarily equate that the residue was enriched during the selection. Nonetheless, taking into account the randomization scheme, frequency logos are useful to quickly visualize the highest occurring residues at a given position.

Sequences for DNA-binding specificity are identified by MEME (Bailey et al. 2009) where MEME identifies overrepresented sequences from all sequences submitted. For our data, the ZOOPS model was used, as it assumes that either zero or one motif occurs per each sequence. While this is a simple tool, the output of MEME is only as good as the selection performed, thus selections should be performed at a selection stringency that recovers a range of motifs with different activities. Moreover, each base is treated as an independent contribution to the binding motif, thus interdependence between bases cannot be captured.

### **Mutual Information Analysis of Amino Acid-Base Interactions**

Mutual Information (MI) identifies covariation between a DBD's recognition residue and a base within the recognition sequence which is a hallmark of a residue's influence on DNA recognition at that position. MI compares the probability of the cooccurrence of a base and an amino acid to the independent occurrence of the base and the amino acid. Particularly in DBDs that lack prior information, such as DNA-protein cocrystal structures or other DNA-protein binding experiments, MI analysis can provide information for which residues may influence the specificity of a given base. The MI calculations can then be used to direct structure and function mutational analysis to ultimately be used in protein

engineering of binding specificity. MI calculations are dependent on the number of samples, thus a small population size (<200 members) will yield high background noise (Mahony et al. 2007). Moreover, MI can also identify evolutionarily linked residues and indirect or subtle effects, thus mutational analysis guided by MI calculation may not always identify large changes, if any, in binding specificity.

### **Electrophoretic Mobility Shift Assays & Competition Binding Assays to Determine Equilibrium Dissociation Constants**

Gel mobility shifts are used to visualize protein-DNA interactions by resolving the differential mobility of free DNA from protein-bound DNA, where electrophoretic mobility shift assays (EMSAs) and competition binding assays are two different types of gel shifts used. The equilibrium dissociation constant ( $K_d$ ) can be calculated from EMSAs when the bound DNA and free DNA are quantified with different protein concentrations mixed with a labeled DNA that is well below the  $K_d$ . EMSA is the most established and widely used method to determine  $K_d$ . To properly calculate a  $K_d$ , the protein concentrations titrated for an EMSA needs to be determined to cover two orders of magnitude over and under the constant. In addition, the binding transition is where the greatest number of data points should occur. To fulfill both these requirements, optimization is necessary. In addition, EMSAs can be used to look at cooperative binding of DNA-protein interactions. While it is a sensitive assay, weak interactions may not be detected due to high off rates of a given interaction. Moreover, very strong interactions cannot be identified, as a minimal amount of labeled DNA needs to be detected and quantified.

Competition binding assays consists of titrating in various concentrations of unlabeled DNA of the same or mutated sequence into subsaturating protein-DNA complex where the DNA of the starting complex is labeled. The free and bound DNA as compared to the competitor concentration can then be used to calculate the equilibrium dissociation constant of the competitor ( $K_c$ ). The advantages of competition binding assay includes the lack of need to label the DNA competitor, thus many different DNA sequences can be tested to compare affinities against each other. Since there is no lower limit of DNA, weak interactions can be measured unlike EMSA. However, each unlabeled sequence to be tested is performed as an individual assay.

### **Superimposition to Estimate Distance Spanning Two DBDs**

Superimposition by structurally aligning solved crystal structures on idealized B-form DNA allow for gross estimation of distance between positions on different DBDs aligned over the DNA. This is a quick method that estimates the shortest possible distances between two DBD while at the same time avoiding possible steric interference between the two DBDs. This is opposed to more precise energy minimization methods that require more technical expertise and accurate force fields to model the complexes, which are still under development (Liu and Bradley 2012) Estimations are dependent on crystal structure quality and idealized B-form DNA may not be representative of what a built chimeric protein will recognize, as a DNA *in vivo* is expected to display sequence-dependent conformational differences. Estimations are then used as a guide of distance to

allow for empirical determination of the length of linker needed to join two DBDs. The estimation, however, can be quite different from final determined linker length.

### **Bacterial-One Hybrid Activity Assay**

Utilizing the B1H system described above, individual bait and prey combinations are transformed into the selection strain. Each combination can then be compared for growth rates on selective media to compare the relative DBD-target site activity to each other. While there is a correlation of activity to affinity for the B1H system, this assay, however, does not allow for absolute quantification of affinity between the different bait and prey combinations (Noyes et al. 2008b). Nonetheless, the comparison of activity does not require the need for labeled DNA or protein purification, unlike gel mobility shift assays.

### **Yeast-Based Nuclease Assay**

The yeast-based nuclease assay allows for readout of nuclease function to a target site. This assay is a chromosomal reporter system where the target site is integrated between an alpha-galactosidase gene (*MEL1*). Expression vector(s) containing the nuclease(s) is transformed into the strain containing the integrated target site. If the nuclease(s) is able to target the site to create a double-stranded break, it allows for *in vivo* yeast machinery to resect the intervening target to repair the *MEL1* reporter gene via single-strand annealing. The restored *MEL1* gene can then be assayed in liquid culture by spectrophotometry, where a higher nuclease activity gives higher measured readout (Doyon et al. 2008).

This assay allow nucleases to be tested in a controlled system where different target sites are integrated into yeast at the same location within the yeast genome. Thus different combinations of nuclease and target sites can be easily tested and compared against each other. The advantage of first testing nucleases in this system is that the readout of nuclease activity to a target site can be compared to each other. For *in vivo* targets that are first tested in this system, it allows for the nuclease and target site combination to be tested without having to take into account confounding factors inherent to an *in vivo* systems (such as local chromatin structure and DNA methylation). Just because a nuclease(s) demonstrates activity in the yeast system, however, it does not always correlate to function in other organisms, such as zebrafish.

### **Nuclease Treatment of Zebrafish**

To introduce artificial nucleases to zebrafish to create targeted genomic lesions, mRNA of the nucleases are injected into single-cell stage embryos. The mRNA is translated *in vivo* by the zebrafish that will then potentially allow the nucleases to generate lesions (insertions or deletions) at the target site. The optimal dose of mRNA that is likely to result in lesions is empirically determined by injecting various concentrations to identify a dose that yields 30 percent deformed embryo morphology (and the remainder normal morphology). Thus embryos at the desired dose, both normal and deformed, are then isolated for lesion identification, although this is not an absolute indication that lesions have been create at the target site (Meng et al. 2008).



Nuclease treatment is a reverse genetic technique, thus it has the advantage over forward genetic techniques to direct targeted mutagenesis and does not require massive screening followed by cumbersome identification of the genetic variant. Moreover, nuclease treatment has advantages over other reverse genetic techniques in zebrafish of morpholinos, TILLING, and retroviral/transposon-mediated mutagenesis. Nuclease treatment creates permanent lesion that can create founders, unlike morpholinos that only transiently knocks down gene expression. TILLING requires the identification of a mutagenic event from a large library of mutants, where these mutants have many mutations in addition to the mutation in the gene of interest. Random insertions of retroviral/transposon-mediate mutagenesis does not allow for controlled directed mutagenesis that is often desired in a genetic technique. Overall, nuclease treatment has the advantage to create heterozygous carriers with minimal off-target effects.

With the advantages over many reverse genetic techniques, artificial nuclease treatment has several limitations. The targets of artificial nuclease are limited by the ability of the DBD (whether it be HDs, ZFs, TALENs, or CRISPRs) to target a site, although there are current efforts to expand targeting by DBDs. Even if a DBD is theoretically available for a particular site, different nucleases have different success rates due to the DBD properties (affinity and specificity), general nuclease architecture (specificity and stringency), and local genomic effects (chromatin structure and DNA methylation). Nonetheless, artificial nucleases are being more commonly used as a genetic technique in many fields.

## **LacZalpha Blue-White Assay for Lesion Identification**

To identify the types of lesions created by the nuclease, the region that was targeted by the nuclease is PCR amplified from pooled 24 h.p.f. embryos and cloned into the LacZalpha gene. The amplified product is designed to be short (60-90 base pairs) and in-frame with the LacZalpha gene to have minimal disruption on the function of the LacZ peptide. If a lesion, as a deletion or insertion, is present, it will disrupt the reading frame of LacZ resulting in a non-functional product. A functional product results in blue colonies plated on X-gal and IPTG, while, a non-functional product, indicative of a lesion, will result in white colonies due to inactive beta-galactosidase. The white colonies are then isolated and sequence to identify the exact sequence present (Zhu et al. 2011).

This assay allows for visual assessment of lesions, instead of shotgun cloning, where wild-type sequences are also sequenced, thus it decreases unnecessary sequencing. Since a triplet insertion or deletion does not result in a frame shift, these types of lesions will not be detected with this assay. While this method has a quick sample preparation to identify lesion types present as compared to Illumina sequencing, it requires that each pooled sample be prepared, cloned, screened, then multiple colonies be sequenced, thus assessing lesions in numerous pools can be time consuming.

**CHAPTER VI**  
**GENERAL DISCUSSION**

Here we were able to dramatically expand HD 3' specificity from previously observed specificity in naturally occurring HDs. This raises further questions on the evolutionary implications of the expansion. Moreover, we show that new HD variants that we identified can be used in the new type of artificial nuclease we engineered, the nZFHD. By optimizing the linker between the ZF and HD as well as the nuclease and ZFHD, we were able to ultimately create site-directed lesions in zebrafish. Thus, we introduce the nZFHD as an additional tool to create site-specific lesions in a complex genome.

### **Expanding HD Sequence Specificity**

By attempting to engineer the HD to recognize all 64 3' triplet sites (TAANNN) by fully randomizing positions in the recognition helix we were able to dramatically increase the range of sequences HDs can preferentially recognize. We were able to expand the 3' end of the HD binding site from 14 to the 64 triplet sites to 44 of the possible 64 sites. This was accomplished by searching a larger randomized library of HDs than had been previously described for novel recognition properties. Some of the recovered variants contained amino acids not previously observed in naturally-occurring HDs at residues 43, 46, 47, 50, and 54 that appear to promote novel recognition properties. These selected HD variants display similar binding affinity and specificity to that of the parent HD for its cognate site. Moreover, amino acid combinations that specify a particular 3' sequence are able to function in alternative HD backbones or in combination with other 5' specificity determinants.

The success of our experiments likely originates from our ability to exhaustively randomize five recognition helix residues of the HD that we deemed important for 3' specificity. Each randomized residue was anticipated to contribute to different aspects of specificity based on prior literature studies, and this expectation was largely confirmed by our MI analysis of selected clones from each target site. The inability of prior studies to identify HD variants with novel HD specificity (Pomerantz and Sharp 1994; Connolly et al. 1999; Mathias et al. 2001) may be due to their more limited variation at a subset of these recognition positions, as many of our HDs with new recognition properties contain a diverse set of amino acids

The similarity in binding affinities of the identified HD variants for their cognate sites to the En parent for its cognate site is expected as the HD variants were selected under similar stringencies (where we have observed a loose correlation between affinity and stringency in the B1H system (Noyes et al. 2008b). Moreover, since the strongest contributor to affinity, N51 (Ades and Sauer 1995), was held constant in our study, this result is not unexpected. The favorable affinities of the HD variants demonstrates that when the HD is fused to zinc fingers in the B1H system, sufficient dynamic range remains within the system to select variants with recognition properties similar to the parent HD. It is plausible that HD variants with different binding affinities can also be selected. This may require that different HD backbones be used and the affinity of the zinc fingers, which are fused to the HD in the B1H system be tuned to account for changes in affinity. Grafting the novel specificity determinants on more dramatically different HD backbones, such

as an atypical HD (Noyes et al. 2008a), will allow further exploration of the interdependence between our selected specificity determinants and other backbone residues.

### **Limitations of HD Recognition Potential**

For 20 sites, out of the 64 TAANNN interrogated, we could not identify a HD variant with preferential binding. In particular, we found difficulty in specifying thymidine at base 6. This may be a result of the lack of inherent flexibility within the HD scaffold to specify this particular base pair. Residue 50 dictates specificity for base 6 by contacting the complementary base within this base pair. The favorable interaction of adenine in the complementary base 6 position to residue 50 could be limiting the recognition potential.

While we chose residues to randomize in our library that we deemed to be major determinants of 3' specificity, other residues in the HD are in contact with the phosphate backbone of the binding site, which include residues 8, 25, 31, 44, 53, and 57 (Fraenkel et al. 1998; Passner et al. 1999; Grant et al. 2000). These contacts likely contribute to binding affinity, but they may also influence specificity in unanticipated ways by indirect readout of the target site. Residues within the HD, those that are not observed to make contacts to the binding site, particularly residues in the recognition helix, may also contribute to orienting the major specificity determinants to specify a particular site. Mutations to these residues could broaden HD specificity and additional libraries with different combinations of

randomized residues in the HD could allow for further expansion of 3' specificity of the HD.

### **Evolutionary Implications of Expanding HD Specificity**

The broad spectrum of HD specificity observed in this study raises the questions of why naturally occurring eukaryotic HDs do not fully exploit their recognition potential. It is particularly striking when the diversity of DNA-binding specificity of HDs is compared to ZFPs, where ZFPs appear to be rapidly evolving with regards to recognition potential (Myers et al. 2010). This characteristic of ZFPs appears as an outlier compared to other DBD families such as bHLHs, bZIPs, and ETS (Wei et al. 2010; De Masi et al. 2011). Thus extant HDs have limited measured specificity (Berger et al. 2008; Noyes et al. 2008a; Jolma et al. 2013) similar to that observed for other DBD families.

HDs are essential for the early development of embryos and bind to many sites throughout the genome (Mann et al. 2009). Thus, they can be viewed as highly connected nodes in a network. Could it be that HDs evolve slower (with less diversity of interaction) because each member is so highly connected and vital to embryogenesis? This view parallels that of the high conservation of RNA-binding specificity (Ray et al. 2013). (A high degree of connectedness is also consistent with the essential nature of many HDs.) This view of restricted evolution is controversial as essential hubs have also been demonstrated to have more diverse specificity and than non-essential hubs (Song and Singh 2013). Moreover, since HDs typically recognize a small hexamer binding site, they can recognize potentially millions of

sites in a genome with high affinity therefore have evolved to be highly necessary since they recognize so many sites in a genome. If a larger recognition site is utilized then it occurs less in a genome, therefore less essential and allowed to increase in diversity.

The DNA recognition of HDs also function with a cofactor that extends or modifies the HD's recognition potential. For example, in Hox-Pbx heterodimer complex results in more specific binding when the two HD complexes to bind to DNA (Joshi et al. 2007). Thus, to understand how a binding partner of a HD can constrain the evolution of the HD's specificity, directed evolution of the HD specificity can be performed with a given cofactor. This may identify how DNA-binding specificity evolves when a TF binding partner is involved in its specificity. Furthermore, the limited HD specificity brings about the question of how their CRMs evolve compared to CRM regulated by a TF with greater diversity of binding, such as ZFPs.

While we are able to measure the specificity of HDs in our study it does not recapitulate the true dynamics of *in vivo* binding. To further understand the nuances of DNA binding by HDs and their evolution, network interaction studies of HDs in an organism or several organisms, such as that performed in *C.elegans* TFs (Reece-Hoyes et al. 2013), would be useful to understand the broader picture of HD evolution. Such a study will build a clearer picture to why we were able to select for such expanded specificity of a DBD and how that relates to neofunctionalization of HDs, or if it does at all. Could it be that HDs have DNA bispecificity (a term coined for secondary specificity) not observed in previously measured HDs therefore only



allowing for 14 of the 64 possible triplet site to be measured as measured for other TFs (Nakagawa et al. 2013). Further studies measuring HD specificity, such as HT protein binding microarrays to more broadly define extant HD specificity will also allow a more complete assessment of the extent to which HD DNA-binding specificity is limited in natural systems as protein binding microarray can capture DNA bispecificity.

### **Future Directions for Broadening HD Specificity**

Expanding HDs to recognize a greater range of target sites demonstrates that the HD backbone is amenable to further changes in specificity and thus further expansion of specificity at the 5' and 3' end may be possible. Identifying new specificities through the creation of a library of residues affecting 5' specificity may greatly increase the number of sites HDs can recognize. A combination of residues that influence 5' specificity - 2, 3, 6-8, and 55 (Ekker et al. 1994; Damante et al. 1996; Noyes et al. 2008a) (Figure 1-2) - could be randomized to create a HD library identifying novel 5' specificities. Additionally, by understanding 5' recognition through reengineering, it would allow for more sophisticated predictive models to be created as we have done with the 3' specificity of HDs (Chu et al. 2012). Moreover, while residue 51 contains an almost invariant asparagine, other HDs are known to contain other residues at this position, such as the Lag1 HD (Noyes et al. 2008a). Changing this residue would allow exploration of the potential for recognition of other bases at position 3. However, the lack of studies for understanding the affinity of residue L51 in Lag1 may necessitate further

exploration before utilizing it as a HD backbone for reengineering. New 5' specificities identified will likely be compatible with 3' specificity to greatly expand the diversity of hexamer binding by the HD.

### **HDs in artificial nucleases**

By attempting to optimize the linker between the ZF and HD from the original ZFHD construct used in the B1H system that expanded the HD binding specificity we were able to identify more stringent and higher activity DNA recognition by the ZFHD modules. Identifying a linker to join the *FokI* nuclease domain to the N-terminus of the new ZFHD allowed us to create an active nZFHD, which ultimately created sequence-directed genomic lesions in zebrafish.

The four different orientations and spacings that the ZF and the HD can bind relative to each other increase the number of sites the ZFHD can target. While we created six pairs of nZFHD to target six different sites in zebrafish, we only identified lesions in one of the six targets. The lesion rate for this target shows that nZFHD can be as efficient as ZFNs with the ability to create similar types of small insertions and deletions at the target site (Gupta et al. 2012).

The ability to incorporate HDs in combination with ZFs broadens their joint targeting capacity as DBDs as ZFs are better at recognizing guanine rich sequence and HD are better at recognizing AT-rich sites. Moreover, by increasing the number of different sites a ZFHD can target from the original ZFHD construct site, we have increased the number of targetable 6 bp binding site from four percent to twenty percent of all possible 6 bps sites. Taking into account that the 5' of the HD can

recognize TGA, TAA, TTA, CAA, and CGA (triplets that ZFs can not target), this is twenty percent is in addition to the possible 6 bp sites that can be recognized by our single fingers ZF modules (18 percent) (Zhu et al. 2011). Additionally, the inability of the current set of single finger modules from our lab to recognize the five 5' triplet sites recognized by HDs highlights the complementarity between these module sets.

### **Limitations of HDs in artificial nucleases**

Our current results show that while we have created functional nZFHDs, our success rate is limited, where only one of the six nZFHD used resulted in lesions. Here we have built the ZFHD by combining the expected DNA-binding sequence of the ZF with the expected DNA-binding sequence of the HD. We could further validate the DNA-binding specificity of each constructed ZFHD chimera as the fused domain (containing our selected linkers) to identify the actual sequence each ZFHD recognizes to test if modularity of the ZF and HD is preserved. The assembly of what were deemed modular ZFs in ZFNs has resulted in unexpected specificity for the final ZFP (Zhu et al. 2011). Thus, the specificity of each ZFHD created for nZFHDs can be tested in the B1H system to measure if the expected specificity of each assembled ZFHD has been preserved.

Nonetheless, the nZFHDs failure to function may not be due to the architecture of the nZFHDs themselves. Further exploration of the cause of the nZFHD's inability to create lesions at the other five target sites may identify if it is the inability of the nZFHD to function properly or if it is factors not inherent to the

nZFHD. Properties inherent to *in vivo* genomic DNA can affect nuclease targeting of endogenous sequences, such as chromatin architecture and DNA methylation, which can hinder the access to or recognition of the target site (Reyon et al. 2012a; Valton et al. 2012). To evaluate if the five nZFHD are not functional due to properties inherent to the *in vivo* system, they can be tested in a system outside of the zebrafish, such as the yeast-based nuclease assay.

What nZFHDs can target is only as good as what the HD (and ZF) can recognize. Increasing the number of target sites that the HD can recognize via engineering its 5' and 3' specificity will also increase the possible number of targets a ZFHD can specify. This assumes that the modularity of the 5' and 3' of the HD specificity determinants is preserved if new specificity determinant at either the 5' or 3' are identified.

Moreover, the linker between the nuclease and the ZFHD can possibly be improved by selecting for a more stringent linker with higher activity for a particular spacing, as we have done with the linker between the ZF and the HD. This, however, will be more technically challenging as the length of the linker between the nuclease and ZFHD is much longer than the linker between the ZF and HD. In addition, these experiments are more complicated due to the need to gauge nuclease activity. A possible system to test for nuclease activity is an *E.coli* based system that has been utilized for directed evolution of the *FokI* nuclease domain to create a higher activity ZFN.(Guo et al. 2010). This system utilizes a toxic reporter plasmid that is destroyed upon nuclease activity to allow for the evolved nuclease to be recovered and identified. Identifying a linker between the nuclease domain and

ZFHD to give the nZFHD higher stringency and activity could provide a versatile nZFHD.

### **Future Directions of HD Variants**

Demonstrating that HDs can be utilized in customizable sequence-directed nucleases provides the impetus for further developing different architectures of chimeric DBDs with ZFs and HDs or even HDs alone. Alternative architectures will allow for even greater flexibility in the toolkit to direct sequence specific activity of a protein or enzyme, which will complement ZFs. Alternative architectures include: fusing the nuclease directly to the C-terminus of the HD, fusing the ZF and HD as HDZF to then connect the nuclease to either termini of the HDZF, or even designing tandem HDs as chimeric nucleases (or artificial transcription factors). Then a linker could be identified to connect the nuclease to create a functional chimeric nuclease using a yeast based assay, as we have done, or utilizing the *E.coli* based assay mentioned above (Guo et al. 2010). Increasing the possible architecture of HDs will increase the utility of ZFs, by broadening the frameworks in which ZFs can be employed with HDs for targeted DNA recognition.

Fine-tuning the affinity of the HD may aid in decreasing off-target effects of artificial nucleases or artificial proteins, as this has been explored in a limited number of cases in ZFNs (Gupta et al. 2011; Pattanayak et al. 2011). Residues within the HD (8, 25, 31, 44, 53, 57) have been observed to make several backbone contacts implying that these interactions do not contribute to specificity but rather affinity (Zhu et al. 2011) (Fraenkel et al. 1998; Passner et al. 1999). Mutating these

residues to abolish one or more of these interactions may be used to fine-tune the affinity of the HD.

HDs can be fused to other DBDs to increase the complex sequences they can specify. For example, ZFs have been fused to the leucine zipper to create a heterodimeric functional unit (Wolfe et al. 2003). Thus, the combinations of chimeric DBDs utilizing HDs are endless where HDs can be seen as parts that can be added to the toolbox to build chimeric transcription factors or chimeric nucleases.

Methods to modulate active DNA-binding of the HD within an *in vivo* system hold potential through the phosphorylation of the HD. The DNA-binding activity of HDs can be modulated by phosphorylation at residue 7, where either a decrease or increase in affinity is observed after phosphorylation, depending of the particular binding site (Kapiloff et al. 1991). However, sequences adjacent to the core-binding motif can influence the affect of the phosphorylation, thus further exploration to understand the specifics of this phenomenon may be necessary. Nonetheless, the utility of phosphorylation in the HD maybe useful as a switch for turning DNA recognition on or off.

### **Overall Utility of HDs as DBD**

Broadening what HDs can specify to then utilize the new variants in chimeric nucleases complements the limitation of ZFs in ZFNs. While the field of artificial nucleases is quickly growing with TALENs and CRISPRs, nZFHDs (and ZFNs) have several advantages over TALENs and CRISPRs. The small size of ZFHDs and ZFs provide an advantage over TALEs to be incorporated into gene delivery vectors

(Holkers et al. 2013) since the typical TALE molecule can be greater than 900 amino acids, where ZFs and HDs are less than 200 amino acids. In addition, both nZFHDs and ZFNs can be functionally encompassed by one type of molecule to simplify delivery of a nuclease to a given system. This can be an advantage over CRISPRs since it requires both the Cas9 protein and sgRNA molecules to generate a DBS while the other nuclease systems only require the nuclease protein to function. Moreover, nZFHD may provide advantages over the CRISPR system due to its high off-target effects (Fu et al. 2013). By expanding HD specificity and applying them for use in artificial nucleases, we have added HDs to the arsenal of DBDs to be used as tools to further biological investigation, biotechnology, and therapeutics.

## **APPENDIX**



**Appendix A-1:**  
**Significance of determinat-triplet correlations**

Helix position	Residue(s)	Triplet preference	Residue(s) present in recovered sequences for target triplets	Residue(s) absent in recovered sequences for target triplets	Residue(s) present in recovered sequences for excluded triplets	Residue(s) absent in recovered sequences for excluded triplet	Odds ratio	95% confidence interval	P-value <sup>a</sup>
47	IVT	Tnn	2430	786	885	8661	30.25	33.66 - 27.18	0
47	KR	Gnn	2182	974	827	8779	23.78	26.41 - 21.41	0
47	N	Ynn	997	5393	293	6079	3.84	4.41 - 3.35	4.61E-99
50	E	nBG	903	1467	259	10133	24.07	28.04 - 20.74	0
50	H	MCn	240	1368	42	11112	46.38	66.31 - 33.13	1.63E-175
50	K	KCC	328	16	499	11919	485.05	851.44 - 294.20	0
50	R	nWC	663	944	672	10483	10.95	12.45 - 9.64	1.44E-283
50	W	nAG	543	261	124	11834	197.94	251.58 - 156.32	0
54	FY	Rnn	2143	4229	77	6313	41.54	52.94 - 32.95	0
54	K	Cyn	1252	314	135	11061	322.92	405.14 - 262.56	0
54	R	CRn	1331	277	1126	10022	42.51	49.23 - 36.77	0

a – Benjamini-Hochberg adjusted P-value is reported to account for multiple hypothesis testing.

## Appendix A-2:

List of validated HD variants that associate with novel and previously defined specificity determinants

KR47 --> Gnn	6_VRVAA	35_VMRWY	269_KVTNF
26_ALKNM	8_RVVSQ	65_GSRWY	
27_LTKDQ	10_KSVMQ		
28_RSKER	11_KSVAQ		K54 --> CYn
29_TLKNQ	12_RGVAA	H50 --> MCn	210_RSNQK
30_LAKDQ	13_ATVKA	267_RVSHT	46_RLDSK
61_KGKEW	17_TRVSA	70_KTSHM	44_RGDSK
62_SHKEY	19_RMVSA	264_KVYHV	67_MKYEK
23_VQKRF	20_QRVSA	66_KTSHM	45_RCYEK
213_KMKES	21_ERVSV	266_KACHS	43_MTNQK
217_KSKEG	56_HRVQA		224_IMNSK
243_ATKSM	60_KTVQV	E50 --> nBG	207_ITYGK
244_KMKSV	201_VRVSQ	28_RSKER	222_SKYGK
249_QLKQS	216_NRVMM	213_KMKES	38_LHYAK
250_AGKTF	235_RAVSV	217_KSKEG	252_LRYSK
261_KSKEA	237_YAVNA	40_STRER	48_EHNAK
23_GTRAY	239_RTVRA	261_KSKEA	47_KMTQK
25_YTRQV	258_RTVQQ	61_KGKEW	220_LTYQK
33_ALRQQ		62_SHKEY	221_RLYQK
35_VMRWY			304_TTNQK
36_ATRRF	N47 --> Ynn		218_KQNQK
65_GSRWY	39_IFNAK	K50 --> KCC	
202_NAREF	43_MTNQK	15_RMIKS	
205_TQRQW	48_EHNAK	32_VRLKY	R54 --> CRn
240_SSRGF	71_MTNNR	13_ATVKA	52_LGMRR
241_GLRAF	210_RSNQK	265_WYSKY	49_LSQSR
242_LQRG	211_TKNQN	246_ISVKY	53_ERVSR
262_QFRAW	218_KQNQK	245_RAVKW	55_LHYVR
35_VGRLY	224_IMNSK		51_MSHWR
310_YRRGA	232_IKNQM		54_LMYQR
311_YRRGF	233_VMNQQ	F54 --> Rnn	253_VANSR
	253_VANSR	59_TRMAF	40_STRER
	303_VMNRK	75_SISRF	168_SRYDR
IVT47 --> Tnn	304_TTNQK	36_ATRRF	251_VGYSR
15_RMIKS		308_RLDRF	257_KLCSR
226_KMISA		203_VQKRF	209_PRDSR
227_YRIAA	R50 --> nWC	31_KITKF	
229_GRISA	52_LGMRR	241_GLRAF	
230_ERISQ	225_SLQRF	73_KLTAF	Y54 --> Gnn
238_QRISV	272_VAQRC	202_NAREF	305_VGRLY
9_KTTQD	36_ATRRF	240_SSRGF	309_RLDRY
14_KGTQM	308_RLDRF	311_YRRGF	35_VMRWY
18_RLTQA	271_KLQRF	37_RFQKF	65_GSRWY
22_RITAA	203_VQKRF	250_AGKTF	34_RTMRY
212_RVTNA		72_KMSNF	32_VRLKY
236_KSTQM	W50 --> nAG	225_SLQRF	265_WYSKY
1_RTVAA	51_MSHWR	271_KLQRF	246_ISVKY
2_RTVSA	76_RAQWF	76_RAQWF	204_RTDYR
4_VRVSA	270_RAQWF	270_RAQWF	
5_TRVAA	301_RSQWH	268_KLQAF	

**Appendix A 3:****Construction and assessment of RF models for predicting HD specificity.**

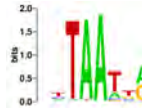
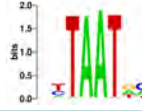
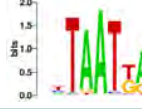
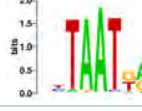
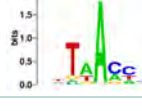
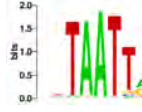
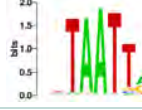
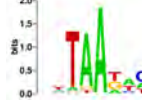
<b>Trial</b>	<b>Model Assessment</b>	<b>Training Set</b>	<b>Prediction Set</b>	<b>MSE per parameter</b>
A	Full	100% extant HDs	100% en mutant HDs	0.053
B	10-fold CV	90% en mutant HDs	10% of en mutant HDs	0.015
C	10-fold CV	90% en mutant HDs + 100% extant HDs	10% of en mutant HDs	0.014
D	Full	100% en mutant HDs	100% extant HDs	0.025
E	Positive control	100% extant HDs	100% extant HDs	0.003
F	Positive control	100% en mutant HDs	100% en mutant HDs	0.004

Extant HDs indicated the set of 246 mouse and fruit fly HDs previously used for modeling (Christensen et al. 2012). En mutant HDs indicates the 151 characterized selected HDs from this study. CV, cross-validation; MSE, mean squared error.

## Appendix A-4: Human HD Predictions

Name ▾	Logo	PFM Matrix
A1L4G3_HUMAN/200..257		A   0.2008 0.1962 0.0212 0.9709 0.9865 0.1785 0.1144 0.5041 0.2028 C   0.2465 0.2269 0.0088 0.0065 0.0084 0.0788 0.1435 0.0106 0.2671 G   0.2883 0.1034 0.0043 0.0053 0.0019 0.0565 0.2367 0.4559 0.3203 T   0.2644 0.4735 0.9656 0.0173 0.0032 0.6862 0.5054 0.0294 0.2098
A4D0Z1_HUMAN/161..221		A   0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535 C   0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239 G   0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676 T   0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550
A4D127_HUMAN/185..245		A   0.2344 0.1869 0.0226 0.9033 0.9814 0.0073 0.0113 0.7340 0.1894 C   0.2384 0.2464 0.0173 0.0095 0.0072 0.0342 0.0514 0.0071 0.3376 G   0.2724 0.1015 0.0038 0.0480 0.0035 0.0164 0.4441 0.2224 0.2769 T   0.2548 0.4652 0.9563 0.0392 0.0080 0.9421 0.4932 0.0365 0.1961
A4D182_HUMAN/32..92		A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826 C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520 G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637 T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017
A6NLG4_HUMAN/10..70		A   0.2216 0.1799 0.0271 0.9614 0.9876 0.0035 0.0280 0.4998 0.2551 C   0.2672 0.2974 0.0107 0.0123 0.0082 0.0037 0.0568 0.0421 0.2260 G   0.2853 0.1350 0.0018 0.0161 0.0018 0.0203 0.0898 0.3923 0.3507 T   0.2259 0.3877 0.9605 0.0102 0.0024 0.9725 0.8254 0.0657 0.1682
A8MWF9_HUMAN/55..115		A   0.2344 0.1869 0.0226 0.9033 0.9814 0.0073 0.0113 0.7340 0.1894 C   0.2384 0.2464 0.0173 0.0095 0.0072 0.0342 0.0514 0.0071 0.3376 G   0.2724 0.1015 0.0038 0.0480 0.0035 0.0164 0.4441 0.2224 0.2769 T   0.2548 0.4652 0.9563 0.0392 0.0080 0.9421 0.4932 0.0365 0.1961
ADNP_HUMAN/763..815		A   0.2437 0.2375 0.0837 0.5709 0.9766 0.2939 0.2199 0.2667 0.1464 C   0.2498 0.1752 0.0459 0.0343 0.0052 0.5463 0.5009 0.3620 0.2379 G   0.2272 0.1584 0.0869 0.1296 0.0054 0.1422 0.0216 0.1255 0.4717 T   0.2793 0.4289 0.7835 0.2651 0.0128 0.0175 0.2575 0.2458 0.1440
ALX1_HUMAN/131..191		A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810 C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916 G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810 T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464
ALX3_HUMAN/152..212		A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810 C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916 G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810 T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464
ALX4_HUMAN/213..273		A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810 C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916 G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810 T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464
ARGFX_HUMAN/77..137		A   0.2261 0.2432 0.0659 0.8778 0.9851 0.1996 0.4513 0.2747 0.1693 C   0.2579 0.2110 0.0199 0.0110 0.0071 0.0000 0.1487 0.4981 0.2496 G   0.2701 0.1098 0.0243 0.0595 0.0027 0.3692 0.0479 0.0000 0.4634 T   0.2460 0.4361 0.8899 0.0517 0.0051 0.4312 0.3521 0.2272 0.1176

# Appendix A-4 contd.

Name ▾	Logo	PFM Matrix
A1L4G3_HUMAN/200..257		A   0.2008 0.1962 0.0212 0.9709 0.9865 0.1785 0.1144 0.5041 0.2028 C   0.2465 0.2269 0.0088 0.0065 0.0084 0.0788 0.1435 0.0106 0.2671 G   0.2883 0.1034 0.0043 0.0053 0.0019 0.0565 0.2367 0.4559 0.3203 T   0.2644 0.4735 0.9656 0.0173 0.0032 0.6862 0.5054 0.0294 0.2098
A4D0Z1_HUMAN/161..221		A   0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535 C   0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239 G   0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676 T   0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550
A4D127_HUMAN/185..245		A   0.2344 0.1869 0.0226 0.9033 0.9814 0.0073 0.0113 0.7340 0.1894 C   0.2384 0.2464 0.0173 0.0095 0.0072 0.0342 0.0514 0.0071 0.3376 G   0.2724 0.1015 0.0038 0.0480 0.0035 0.0164 0.4441 0.2224 0.2769 T   0.2548 0.4652 0.9563 0.0392 0.0080 0.9421 0.4932 0.0365 0.1961
A4D182_HUMAN/32..92		A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826 C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520 G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637 T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017
A6NLG4_HUMAN/10..70		A   0.2216 0.1799 0.0271 0.9614 0.9876 0.0035 0.0280 0.4998 0.2551 C   0.2672 0.2974 0.0107 0.0123 0.0082 0.0037 0.0568 0.0421 0.2260 G   0.2853 0.1350 0.0018 0.0161 0.0018 0.0203 0.0898 0.3923 0.3507 T   0.2259 0.3877 0.9605 0.0102 0.0024 0.9725 0.8254 0.0657 0.1682
A8MWF9_HUMAN/55..115		A   0.2344 0.1869 0.0226 0.9033 0.9814 0.0073 0.0113 0.7340 0.1894 C   0.2384 0.2464 0.0173 0.0095 0.0072 0.0342 0.0514 0.0071 0.3376 G   0.2724 0.1015 0.0038 0.0480 0.0035 0.0164 0.4441 0.2224 0.2769 T   0.2548 0.4652 0.9563 0.0392 0.0080 0.9421 0.4932 0.0365 0.1961
ADNP_HUMAN/763..815		A   0.2437 0.2375 0.0837 0.5709 0.9766 0.2939 0.2199 0.2667 0.1464 C   0.2498 0.1752 0.0459 0.0343 0.0052 0.5463 0.5009 0.3620 0.2379 G   0.2272 0.1584 0.0869 0.1296 0.0054 0.1422 0.0216 0.1255 0.4717 T   0.2793 0.4289 0.7835 0.2651 0.0128 0.0175 0.2575 0.2458 0.1440
ALX1_HUMAN/131..191		A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810 C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916 G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810 T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464
ALX3_HUMAN/152..212		A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810 C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916 G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810 T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464
ALX4_HUMAN/213..273		A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810 C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916 G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810 T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464
ARGFX_HUMAN/77..137		A   0.2261 0.2432 0.0659 0.8778 0.9851 0.1996 0.4513 0.2747 0.1693 C   0.2579 0.2110 0.0199 0.0110 0.0071 0.0000 0.1487 0.4981 0.2496 G   0.2701 0.1098 0.0243 0.0595 0.0027 0.3692 0.0479 0.0000 0.4634 T   0.2460 0.4361 0.8899 0.0517 0.0051 0.4312 0.3521 0.2272 0.1176

# Appendix A-4 contd.

CDX4_HUMAN/172..232		<div>A   0.2142 0.0729 0.0985 0.3510 0.9785 0.0091 0.0977 0.5420 0.1289</div> <div>C   0.1734 0.0712 0.0209 0.0000 0.0069 0.0815 0.0094 0.0154 0.4211</div> <div>G   0.2170 0.0543 0.0111 0.0622 0.0046 0.0305 0.2840 0.4209 0.2441</div> <div>T   0.3954 0.8016 0.8696 0.5868 0.0101 0.8790 0.6089 0.0217 0.2060</div>
CERS2_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
CERS3_HUMAN/69..128	No prediction made	The extracted domain has residue (S) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
CERS4_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
CERS5_HUMAN/81..137	No prediction made	The extracted domain has residue (H) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
CERS6_HUMAN/78..128	No prediction made	The extracted domain has residue (Q) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
CRX_HUMAN/38..98		<div>A   0.2271 0.2000 0.0400 0.9635 0.9883 0.0002 0.0349 0.0327 0.1395</div> <div>C   0.2471 0.2969 0.0041 0.0098 0.0085 0.0051 0.9108 0.7577 0.3630</div> <div>G   0.3047 0.1251 0.0091 0.0116 0.0013 0.0741 0.0043 0.0403 0.2830</div> <div>T   0.2211 0.3780 0.9468 0.0151 0.0019 0.9206 0.0500 0.1693 0.2146</div>
CUX1_HUMAN/1243..1303		<div>A   0.1438 0.1791 0.0787 0.3598 0.9711 0.3373 0.1969 0.3645 0.2414</div> <div>C   0.3657 0.1971 0.0998 0.0277 0.0085 0.1734 0.6559 0.2860 0.2789</div> <div>G   0.2819 0.1291 0.0358 0.5008 0.0031 0.0466 0.0000 0.1814 0.2782</div> <div>T   0.2085 0.4947 0.7857 0.1117 0.0172 0.4428 0.1472 0.1682 0.2015</div>
CUX2_HUMAN/1167..1227		<div>A   0.1438 0.1791 0.0787 0.3598 0.9711 0.3373 0.1969 0.3645 0.2414</div> <div>C   0.3657 0.1971 0.0998 0.0277 0.0085 0.1734 0.6559 0.2860 0.2789</div> <div>G   0.2819 0.1291 0.0358 0.5008 0.0031 0.0466 0.0000 0.1814 0.2782</div> <div>T   0.2085 0.4947 0.7857 0.1117 0.0172 0.4428 0.1472 0.1682 0.2015</div>
D2CFI5_HUMAN/61..121		<div>A   0.2526 0.2040 0.0570 0.6844 0.9728 0.0078 0.0988 0.0385 0.1443</div> <div>C   0.2412 0.2275 0.0675 0.0265 0.0092 0.0189 0.8379 0.7140 0.3504</div> <div>G   0.2737 0.1581 0.0113 0.2183 0.0024 0.0344 0.0177 0.0715 0.2837</div> <div>T   0.2324 0.4103 0.8642 0.0707 0.0156 0.9389 0.0456 0.1761 0.2217</div>
D6R955_HUMAN/88..114	No prediction made	The extracted domain has a gap at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
D6R9U1_HUMAN/88..131	No prediction made	The extracted domain has a gap at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
D6RAR5_HUMAN/45..105		<div>A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006</div> <div>C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872</div> <div>G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068</div> <div>T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053</div>
D6RBB8_HUMAN/180..240		<div>A   0.2524 0.1439 0.0520 0.9014 0.9711 0.0836 0.1018 0.3286 0.2510</div> <div>C   0.2238 0.3478 0.0234 0.0054 0.0061 0.0148 0.0343 0.0514 0.2630</div> <div>G   0.3062 0.1284 0.0219 0.0829 0.0057 0.1413 0.3529 0.5412 0.2866</div> <div>T   0.2176 0.3800 0.9027 0.0104 0.0171 0.7603 0.5110 0.0789 0.1994</div>

# Appendix A-4 contd.

CDX4_HUMAN/172..232		<div>A   0.2142 0.0729 0.0985 0.3510 0.9785 0.0091 0.0977 0.5420 0.1289</div> <div>C   0.1734 0.0712 0.0209 0.0000 0.0069 0.0815 0.0094 0.0154 0.4211</div> <div>G   0.2170 0.0543 0.0111 0.0622 0.0046 0.0305 0.2840 0.4209 0.2441</div> <div>T   0.3954 0.8016 0.8696 0.5868 0.0101 0.8790 0.6089 0.0217 0.2060</div>
CERS2_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
CERS3_HUMAN/69..128	No prediction made	The extracted domain has residue (S) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
CERS4_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
CERS5_HUMAN/81..137	No prediction made	The extracted domain has residue (H) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
CERS6_HUMAN/78..128	No prediction made	The extracted domain has residue (Q) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
CRX_HUMAN/38..98		<div>A   0.2271 0.2000 0.0400 0.9635 0.9883 0.0002 0.0349 0.0327 0.1395</div> <div>C   0.2471 0.2969 0.0041 0.0098 0.0085 0.0051 0.9108 0.7577 0.3630</div> <div>G   0.3047 0.1251 0.0091 0.0116 0.0013 0.0741 0.0043 0.0403 0.2830</div> <div>T   0.2211 0.3780 0.9468 0.0151 0.0019 0.9206 0.0500 0.1693 0.2146</div>
CUX1_HUMAN/1243..1303		<div>A   0.1438 0.1791 0.0787 0.3598 0.9711 0.3373 0.1969 0.3645 0.2414</div> <div>C   0.3657 0.1971 0.0998 0.0277 0.0085 0.1734 0.6559 0.2860 0.2789</div> <div>G   0.2819 0.1291 0.0358 0.5008 0.0031 0.0466 0.0000 0.1814 0.2782</div> <div>T   0.2085 0.4947 0.7857 0.1117 0.0172 0.4428 0.1472 0.1682 0.2015</div>
CUX2_HUMAN/1167..1227		<div>A   0.1438 0.1791 0.0787 0.3598 0.9711 0.3373 0.1969 0.3645 0.2414</div> <div>C   0.3657 0.1971 0.0998 0.0277 0.0085 0.1734 0.6559 0.2860 0.2789</div> <div>G   0.2819 0.1291 0.0358 0.5008 0.0031 0.0466 0.0000 0.1814 0.2782</div> <div>T   0.2085 0.4947 0.7857 0.1117 0.0172 0.4428 0.1472 0.1682 0.2015</div>
D2CFI5_HUMAN/61..121		<div>A   0.2526 0.2040 0.0570 0.6844 0.9728 0.0078 0.0988 0.0385 0.1443</div> <div>C   0.2412 0.2275 0.0675 0.0265 0.0092 0.0189 0.8379 0.7140 0.3504</div> <div>G   0.2737 0.1581 0.0113 0.2183 0.0024 0.0344 0.0177 0.0715 0.2837</div> <div>T   0.2324 0.4103 0.8642 0.0707 0.0156 0.9389 0.0456 0.1761 0.2217</div>
D6R955_HUMAN/88..114	No prediction made	The extracted domain has a gap at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
D6R9U1_HUMAN/88..131	No prediction made	The extracted domain has a gap at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
D6RAR5_HUMAN/45..105		<div>A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006</div> <div>C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872</div> <div>G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068</div> <div>T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053</div>
D6RBB8_HUMAN/180..240		<div>A   0.2524 0.1439 0.0520 0.9014 0.9711 0.0836 0.1018 0.3286 0.2510</div> <div>C   0.2238 0.3478 0.0234 0.0054 0.0061 0.0148 0.0343 0.0514 0.2630</div> <div>G   0.3062 0.1284 0.0219 0.0829 0.0057 0.1413 0.3529 0.5412 0.2866</div> <div>T   0.2176 0.3800 0.9027 0.0104 0.0171 0.7603 0.5110 0.0789 0.1994</div>

# Appendix A-4 contd.

DPRX_HUMAN/15..75		A   0.2271 0.2000 0.0400 0.9635 0.9883 0.0002 0.0349 0.0327 0.1395 C   0.2471 0.2969 0.0041 0.0098 0.0085 0.0051 0.9108 0.7577 0.3630 G   0.3047 0.1251 0.0091 0.0116 0.0013 0.0741 0.0043 0.0403 0.2830 T   0.2211 0.3780 0.9468 0.0151 0.0019 0.9206 0.0500 0.1693 0.2146
DRGX_HUMAN/32..92		A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810 C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916 G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810 T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464
DU4L2_HUMAN/18..78		A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097 C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800 G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175 T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927
DU4L2_HUMAN/93..151		A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499 C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019 G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823 T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659
DU4L3_HUMAN/18..78		A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097 C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800 G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175 T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927
DU4L3_HUMAN/93..151		A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499 C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019 G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823 T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659
DU4L4_HUMAN/18..78		A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097 C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800 G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175 T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927
DU4L4_HUMAN/93..151		A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499 C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019 G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823 T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659
DU4L5_HUMAN/18..78		A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097 C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800 G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175 T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927
DU4L5_HUMAN/93..151		A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499 C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019 G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823 T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659
DU4L6_HUMAN/18..78		A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097 C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800 G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175 T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927



# Appendix A-4 contd.

DU4L6_HUMAN/93..151		<div>A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499</div> <div>C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019</div> <div>G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823</div> <div>T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659</div>
DU4L7_HUMAN/18..78		<div>A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097</div> <div>C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800</div> <div>G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175</div> <div>T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927</div>
DU4L7_HUMAN/93..151		<div>A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499</div> <div>C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019</div> <div>G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823</div> <div>T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659</div>
DUX1_HUMAN/18..78		<div>A   0.2265 0.1816 0.0601 0.6776 0.9747 0.0048 0.0678 0.4547 0.1994</div> <div>C   0.2534 0.1976 0.0334 0.0264 0.0107 0.0404 0.0530 0.0196 0.2876</div> <div>G   0.2514 0.1495 0.0219 0.2362 0.0028 0.1046 0.1838 0.4796 0.3219</div> <div>T   0.2687 0.4713 0.8846 0.0598 0.0118 0.8502 0.6954 0.0461 0.1911</div>
DUX1_HUMAN/93..153		<div>A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499</div> <div>C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019</div> <div>G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823</div> <div>T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659</div>
DUX2_HUMAN/18..78		<div>A   0.2265 0.1816 0.0601 0.6776 0.9747 0.0048 0.0678 0.4547 0.1994</div> <div>C   0.2534 0.1976 0.0334 0.0264 0.0107 0.0404 0.0530 0.0196 0.2876</div> <div>G   0.2514 0.1495 0.0219 0.2362 0.0028 0.1046 0.1838 0.4796 0.3219</div> <div>T   0.2687 0.4713 0.8846 0.0598 0.0118 0.8502 0.6954 0.0461 0.1911</div>
DUX3_HUMAN/120..179		<div>A   0.2376 0.1579 0.0475 0.4023 0.9878 0.0000 0.0415 0.5208 0.2469</div> <div>C   0.2587 0.2934 0.0112 0.0000 0.0074 0.1044 0.0252 0.0000 0.2071</div> <div>G   0.2690 0.1306 0.0089 0.5639 0.0019 0.0542 0.1060 0.4368 0.3747</div> <div>T   0.2348 0.4181 0.9324 0.0338 0.0029 0.8414 0.8273 0.0424 0.1713</div>
DUX3_HUMAN/45..105		<div>A   0.2265 0.1816 0.0601 0.6776 0.9747 0.0048 0.0678 0.4547 0.1994</div> <div>C   0.2534 0.1976 0.0334 0.0264 0.0107 0.0404 0.0530 0.0196 0.2876</div> <div>G   0.2514 0.1495 0.0219 0.2362 0.0028 0.1046 0.1838 0.4796 0.3219</div> <div>T   0.2687 0.4713 0.8846 0.0598 0.0118 0.8502 0.6954 0.0461 0.1911</div>
DUX4C_HUMAN/18..78		<div>A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097</div> <div>C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800</div> <div>G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175</div> <div>T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927</div>
DUX4C_HUMAN/93..151		<div>A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499</div> <div>C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019</div> <div>G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823</div> <div>T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659</div>
DUX4_HUMAN/18..78		<div>A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097</div> <div>C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800</div> <div>G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175</div> <div>T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927</div>
DUX4_HUMAN/93..151		<div>A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499</div> <div>C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019</div> <div>G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823</div> <div>T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659</div>

# Appendix A-4 contd.

DUX5_HUMAN/120..180		A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499 C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019 G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823 T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659
DUX5_HUMAN/45..105		A   0.2265 0.1816 0.0601 0.6776 0.9747 0.0048 0.0678 0.4547 0.1994 C   0.2534 0.1976 0.0334 0.0264 0.0107 0.0404 0.0530 0.0196 0.2876 G   0.2514 0.1495 0.0219 0.2362 0.0028 0.1046 0.1838 0.4796 0.3219 T   0.2687 0.4713 0.8846 0.0598 0.0118 0.8502 0.6954 0.0461 0.1911
DUXA_HUMAN/100..158		A   0.2373 0.1721 0.0590 0.3934 0.9851 0.0159 0.0457 0.4718 0.2404 C   0.2499 0.2761 0.0165 0.0000 0.0078 0.0205 0.0541 0.0177 0.2336 G   0.2633 0.1473 0.0076 0.5528 0.0026 0.0432 0.1387 0.4779 0.3471 T   0.2495 0.4045 0.9169 0.0537 0.0045 0.9204 0.7615 0.0326 0.1788
DUXA_HUMAN/14..72		A   0.2371 0.1566 0.0529 0.3194 0.9859 0.0145 0.0000 0.4186 0.2301 C   0.2461 0.2671 0.0052 0.0000 0.0071 0.0016 0.0644 0.0416 0.2563 G   0.2763 0.1503 0.0110 0.6414 0.0029 0.0433 0.1249 0.5346 0.3358 T   0.2405 0.4260 0.9309 0.0392 0.0042 0.9406 0.8107 0.0052 0.1778
EOYMI7_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
EOYMJ3_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
EOYMJ4_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
EOYMJ5_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
EOYMJ8_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
EOYMJ9_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
E5RGZ2_HUMAN/266..341		A   0.2260 0.2126 0.1212 0.6215 0.9503 0.0503 0.1019 0.5213 0.2296 C   0.2767 0.2221 0.1232 0.0477 0.0000 0.8548 0.3955 0.1597 0.1796 G   0.2235 0.1424 0.0738 0.1213 0.0130 0.0000 0.0753 0.1542 0.4496 T   0.2738 0.4229 0.6818 0.2096 0.0367 0.0949 0.4273 0.1648 0.1413
E7EMR0_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
E7EN04_HUMAN/129..189		A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006 C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872 G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068 T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053
E7EQ07_HUMAN/86..146		A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006 C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872 G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068 T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053

# Appendix A-4 contd.

E7ER53_HUMAN/2112..2172		<div>A   0.2351 0.1770 0.0755 0.8644 0.9854 0.1264 0.6189 0.1422 0.1215</div> <div>C   0.2654 0.2520 0.0169 0.0187 0.0060 0.0136 0.1701 0.6182 0.2285</div> <div>G   0.2370 0.1027 0.0337 0.0968 0.0036 0.7031 0.0000 0.0275 0.5260</div> <div>T   0.2625 0.4683 0.8739 0.0200 0.0050 0.1570 0.2109 0.2121 0.1240</div>
E7ER53_HUMAN/2209..2269		<div>A   0.2519 0.1819 0.0275 0.8227 0.9853 0.0203 0.0536 0.6569 0.2579</div> <div>C   0.2418 0.2492 0.0408 0.0045 0.0075 0.0334 0.0530 0.0141 0.2140</div> <div>G   0.2715 0.1199 0.0276 0.1135 0.0019 0.0237 0.3889 0.2931 0.3775</div> <div>T   0.2348 0.4490 0.9041 0.0594 0.0052 0.9226 0.5044 0.0359 0.1507</div>
E7ER53_HUMAN/2588..2648		<div>A   0.2342 0.1668 0.0273 0.5568 0.9874 0.0101 0.0138 0.4708 0.2177</div> <div>C   0.2530 0.3389 0.0029 0.0000 0.0079 0.0084 0.0216 0.0118 0.2707</div> <div>G   0.2909 0.1231 0.0049 0.4184 0.0017 0.0107 0.0940 0.4760 0.3289</div> <div>T   0.2220 0.3711 0.9649 0.0247 0.0030 0.9708 0.8706 0.0414 0.1826</div>
E7ER53_HUMAN/2912..2972		<div>A   0.2399 0.1483 0.0321 0.9435 0.9809 0.0210 0.0287 0.5619 0.2547</div> <div>C   0.2461 0.3280 0.0106 0.0110 0.0072 0.0028 0.0488 0.0140 0.2428</div> <div>G   0.2837 0.1082 0.0044 0.0325 0.0037 0.0484 0.1220 0.3763 0.3085</div> <div>T   0.2304 0.4155 0.9529 0.0129 0.0083 0.9279 0.8005 0.0478 0.1940</div>
E7ETP3_HUMAN/159..219		<div>A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849</div> <div>C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028</div> <div>G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384</div> <div>T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739</div>
E7EUQ4_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
E7EUW9_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
E7EVZ1_HUMAN/2102..2162		<div>A   0.2351 0.1770 0.0755 0.8644 0.9854 0.1264 0.6189 0.1422 0.1215</div> <div>C   0.2654 0.2520 0.0169 0.0187 0.0060 0.0136 0.1701 0.6182 0.2285</div> <div>G   0.2370 0.1027 0.0337 0.0968 0.0036 0.7031 0.0000 0.0275 0.5260</div> <div>T   0.2625 0.4683 0.8739 0.0200 0.0050 0.1570 0.2109 0.2121 0.1240</div>
E7EVZ1_HUMAN/2199..2259		<div>A   0.2519 0.1819 0.0275 0.8227 0.9853 0.0203 0.0536 0.6569 0.2579</div> <div>C   0.2418 0.2492 0.0408 0.0045 0.0075 0.0334 0.0530 0.0141 0.2140</div> <div>G   0.2715 0.1199 0.0276 0.1135 0.0019 0.0237 0.3889 0.2931 0.3775</div> <div>T   0.2348 0.4490 0.9041 0.0594 0.0052 0.9226 0.5044 0.0359 0.1507</div>
E7EVZ1_HUMAN/2578..2638		<div>A   0.2342 0.1668 0.0273 0.5568 0.9874 0.0101 0.0138 0.4708 0.2177</div> <div>C   0.2530 0.3389 0.0029 0.0000 0.0079 0.0084 0.0216 0.0118 0.2707</div> <div>G   0.2909 0.1231 0.0049 0.4184 0.0017 0.0107 0.0940 0.4760 0.3289</div> <div>T   0.2220 0.3711 0.9649 0.0247 0.0030 0.9708 0.8706 0.0414 0.1826</div>
E7EVZ1_HUMAN/2902..2962		<div>A   0.2399 0.1483 0.0321 0.9435 0.9809 0.0210 0.0287 0.5619 0.2547</div> <div>C   0.2461 0.3280 0.0106 0.0110 0.0072 0.0028 0.0488 0.0140 0.2428</div> <div>G   0.2837 0.1082 0.0044 0.0325 0.0037 0.0484 0.1220 0.3763 0.3085</div> <div>T   0.2304 0.4155 0.9529 0.0129 0.0083 0.9279 0.8005 0.0478 0.1940</div>
E9PB27_HUMAN/159..222		<div>A   0.2331 0.1862 0.0329 0.1150 0.9785 0.0256 0.3941 0.2947 0.2436</div> <div>C   0.2117 0.1101 0.0141 0.0001 0.0070 0.4410 0.0780 0.1906 0.2202</div> <div>G   0.1937 0.1722 0.0228 0.8471 0.0034 0.0585 0.3485 0.2694 0.3604</div> <div>T   0.3615 0.5314 0.9302 0.0378 0.0111 0.4749 0.1794 0.2454 0.1758</div>
		the extracted domain has multiple (A) at position 53 but multiple (M)

## Appendix A-4 contd.

E9PB55_HUMAN/22..80	No prediction made	The extracted domain has residue (Q) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
E9PCM7_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
E9PEK5_HUMAN/154..212		A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499 C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019 G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823 T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659
E9PEK5_HUMAN/79..139		A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097 C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800 G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175 T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927
E9PFV9_HUMAN/216..276		A   0.2275 0.2398 0.0596 0.7268 0.9881 0.0157 0.1254 0.2813 0.2523 C   0.2571 0.2797 0.0096 0.0030 0.0072 0.0173 0.2354 0.2483 0.1896 G   0.3102 0.1459 0.0148 0.2523 0.0021 0.0099 0.1297 0.3206 0.4273 T   0.2052 0.3345 0.9160 0.0179 0.0026 0.9571 0.5094 0.1498 0.1307
E9PG50_HUMAN/273..331	No prediction made	The extracted domain has residue (D) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
E9PG50_HUMAN/390..447	No prediction made	The extracted domain has residue (E) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
E9PGE3_HUMAN/214..274		A   0.2345 0.1541 0.0242 0.5016 0.9882 0.0068 0.0121 0.5322 0.2058 C   0.2632 0.3621 0.0023 0.0019 0.0075 0.0041 0.0125 0.0084 0.2960 G   0.3123 0.1258 0.0047 0.4749 0.0018 0.0058 0.0753 0.4280 0.3341 T   0.1900 0.3580 0.9688 0.0216 0.0025 0.9833 0.9001 0.0314 0.1641
E9PGG2_HUMAN/137..196		A   0.2383 0.1722 0.0787 0.3269 0.9818 0.0000 0.6108 0.4070 0.2617 C   0.2370 0.2267 0.1101 0.0141 0.0065 0.7438 0.1118 0.1451 0.2381 G   0.2454 0.1740 0.0319 0.5116 0.0044 0.0896 0.0554 0.2006 0.3084 T   0.2793 0.4271 0.7792 0.1474 0.0073 0.1666 0.2221 0.2474 0.1918
E9PIN6_HUMAN/234..294		A   0.2473 0.3513 0.1233 0.7905 0.9802 0.0121 0.0171 0.6505 0.3667 C   0.2127 0.2208 0.0182 0.0067 0.0049 0.0036 0.0486 0.2373 0.0974 G   0.3243 0.1216 0.0152 0.0176 0.0043 0.0067 0.4954 0.0479 0.4528 T   0.2158 0.3063 0.8433 0.1853 0.0105 0.9775 0.4389 0.0643 0.0831
E9PIX4_HUMAN/8..68		A   0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535 C   0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239 G   0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676 T   0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550
E9PLE6_HUMAN/55..115		A   0.2184 0.2379 0.0271 0.9604 0.9834 0.0366 0.0387 0.4661 0.1915 C   0.2319 0.2329 0.0092 0.0105 0.0081 0.1007 0.1784 0.0222 0.2702 G   0.3183 0.1328 0.0103 0.0124 0.0025 0.0358 0.3706 0.4677 0.3421 T   0.2314 0.3965 0.9534 0.0167 0.0060 0.8269 0.4122 0.0439 0.1962
E9PNC9_HUMAN/280..340		A   0.2670 0.3899 0.1107 0.7971 0.9786 0.0187 0.0316 0.6638 0.3870 C   0.2142 0.1971 0.0113 0.0073 0.0029 0.0003 0.0657 0.2101 0.0997 G   0.3053 0.1073 0.0149 0.0115 0.0060 0.0103 0.4403 0.0463 0.4244 T   0.2135 0.3057 0.8630 0.1841 0.0126 0.9707 0.4624 0.0798 0.0889

# Appendix A-4 contd.

E9PQ94_HUMAN/37..97		<div>A   0.2184 0.2379 0.0271 0.9604 0.9834 0.0366 0.0387 0.4661 0.1915</div> <div>C   0.2319 0.2329 0.0092 0.0105 0.0081 0.1007 0.1784 0.0222 0.2702</div> <div>G   0.3183 0.1328 0.0103 0.0124 0.0025 0.0358 0.3706 0.4677 0.3421</div> <div>T   0.2314 0.3965 0.9534 0.0167 0.0060 0.8269 0.4122 0.0439 0.1962</div>
E9PQ10_HUMAN/65..125		<div>A   0.2473 0.3513 0.1233 0.7905 0.9802 0.0121 0.0171 0.6505 0.3667</div> <div>C   0.2127 0.2208 0.0182 0.0067 0.0049 0.0036 0.0486 0.2373 0.0974</div> <div>G   0.3243 0.1216 0.0152 0.0176 0.0043 0.0067 0.4954 0.0479 0.4528</div> <div>T   0.2158 0.3063 0.8433 0.1853 0.0105 0.9775 0.4389 0.0643 0.0831</div>
E9PS79_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
EMX1_HUMAN/158..218		<div>A   0.2385 0.1178 0.0427 0.9530 0.9872 0.0316 0.1208 0.6349 0.2395</div> <div>C   0.2516 0.3932 0.0122 0.0128 0.0076 0.0070 0.0419 0.0228 0.2576</div> <div>G   0.2701 0.1232 0.0157 0.0231 0.0021 0.0417 0.2768 0.2853 0.2980</div> <div>T   0.2398 0.3659 0.9294 0.0112 0.0030 0.9196 0.5605 0.0570 0.2049</div>
EMX2_HUMAN/153..213		<div>A   0.2385 0.1178 0.0427 0.9530 0.9872 0.0316 0.1208 0.6349 0.2395</div> <div>C   0.2516 0.3932 0.0122 0.0128 0.0076 0.0070 0.0419 0.0228 0.2576</div> <div>G   0.2701 0.1232 0.0157 0.0231 0.0021 0.0417 0.2768 0.2853 0.2980</div> <div>T   0.2398 0.3659 0.9294 0.0112 0.0030 0.9196 0.5605 0.0570 0.2049</div>
ESX1_HUMAN/138..198		<div>A   0.2358 0.1569 0.0261 0.9545 0.9880 0.0037 0.0363 0.6063 0.2790</div> <div>C   0.2586 0.3220 0.0092 0.0137 0.0070 0.0069 0.0317 0.0163 0.2085</div> <div>G   0.2918 0.1103 0.0025 0.0196 0.0025 0.0095 0.0772 0.3350 0.3569</div> <div>T   0.2139 0.4108 0.9622 0.0123 0.0025 0.9798 0.8549 0.0425 0.1556</div>
EVX1_HUMAN/182..242		<div>A   0.2350 0.1336 0.0185 0.9687 0.9862 0.0056 0.0284 0.7018 0.1918</div> <div>C   0.2644 0.3754 0.0073 0.0092 0.0076 0.0315 0.0744 0.0197 0.3114</div> <div>G   0.2964 0.1184 0.0071 0.0189 0.0024 0.0112 0.4242 0.2305 0.3153</div> <div>T   0.2042 0.3725 0.9672 0.0032 0.0038 0.9517 0.4731 0.0480 0.1814</div>
EVX2_HUMAN/187..247		<div>A   0.2350 0.1336 0.0185 0.9687 0.9862 0.0056 0.0284 0.7018 0.1918</div> <div>C   0.2644 0.3754 0.0073 0.0092 0.0076 0.0315 0.0744 0.0197 0.3114</div> <div>G   0.2964 0.1184 0.0071 0.0189 0.0024 0.0112 0.4242 0.2305 0.3153</div> <div>T   0.2042 0.3725 0.9672 0.0032 0.0038 0.9517 0.4731 0.0480 0.1814</div>
F2Z381_HUMAN/33..93		<div>A   0.3019 0.2505 0.0874 0.8705 0.9843 0.0063 0.0581 0.5512 0.3071</div> <div>C   0.2338 0.2215 0.0037 0.0060 0.0059 0.0072 0.0558 0.2614 0.1645</div> <div>G   0.2729 0.0959 0.0303 0.0190 0.0029 0.0265 0.5642 0.0837 0.4177</div> <div>T   0.1914 0.4321 0.8786 0.1044 0.0068 0.9600 0.3218 0.1038 0.1108</div>
F5GWW6_HUMAN/282..342		<div>A   0.2473 0.3513 0.1233 0.7905 0.9802 0.0121 0.0171 0.6505 0.3667</div> <div>C   0.2127 0.2208 0.0182 0.0067 0.0049 0.0036 0.0486 0.2373 0.0974</div> <div>G   0.3243 0.1216 0.0152 0.0176 0.0043 0.0067 0.4954 0.0479 0.4528</div> <div>T   0.2158 0.3063 0.8433 0.1853 0.0105 0.9775 0.4389 0.0643 0.0831</div>
F5GXB4_HUMAN/44..104		<div>A   0.2526 0.2040 0.0570 0.6844 0.9728 0.0078 0.0988 0.0385 0.1443</div> <div>C   0.2412 0.2275 0.0675 0.0265 0.0092 0.0189 0.8379 0.7140 0.3504</div> <div>G   0.2737 0.1581 0.0113 0.2183 0.0024 0.0344 0.0177 0.0715 0.2837</div> <div>T   0.2324 0.4103 0.8642 0.0707 0.0156 0.9389 0.0456 0.1761 0.2217</div>
F5GZ66_HUMAN/154..212		<div>A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499</div> <div>C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019</div> <div>G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0133 0.0034 0.5000 0.2800</div> <div>T   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499</div>

## Appendix A-4 contd.

		G   0.2020 0.1333 0.0000 0.0700 0.0024 0.0131 0.0004 0.0023 0.3023 T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659
F5GZ66_HUMAN/79..139		A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097 C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800 G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175 T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927
F5GZI2_HUMAN/70..130		A   0.2184 0.2379 0.0271 0.9604 0.9834 0.0366 0.0387 0.4661 0.1915 C   0.2319 0.2329 0.0092 0.0105 0.0081 0.1007 0.1784 0.0222 0.2702 G   0.3183 0.1328 0.0103 0.0124 0.0025 0.0358 0.3706 0.4677 0.3421 T   0.2314 0.3965 0.9534 0.0167 0.0060 0.8269 0.4122 0.0439 0.1962
F5H0K0_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
F5H1R1_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
F5H2R1_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
F5H3E7_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
F5H401_HUMAN/28..88		A   0.2256 0.2329 0.0320 0.9445 0.9838 0.0132 0.0465 0.4633 0.2032 C   0.2579 0.2875 0.0050 0.0096 0.0080 0.0087 0.1394 0.0207 0.3161 G   0.3030 0.1898 0.0023 0.0331 0.0024 0.0090 0.2441 0.4508 0.2789 T   0.2135 0.2899 0.9607 0.0128 0.0058 0.9691 0.5701 0.0652 0.2018
F5H4I8_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
F5H4U9_HUMAN/232..295		A   0.2331 0.1862 0.0329 0.1150 0.9785 0.0256 0.3941 0.2947 0.2436 C   0.2117 0.1101 0.0141 0.0001 0.0070 0.4410 0.0780 0.1906 0.2202 G   0.1937 0.1722 0.0228 0.8471 0.0034 0.0585 0.3485 0.2694 0.3604 T   0.3615 0.5314 0.9302 0.0378 0.0111 0.4749 0.1794 0.2454 0.1758
F5H5U3_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
F5H7Y3_HUMAN/118..197		A   0.2263 0.1863 0.0979 0.7678 0.9589 0.0180 0.0739 0.4643 0.2794 C   0.2852 0.2552 0.0582 0.0348 0.0002 0.9122 0.4205 0.1100 0.1861 G   0.2565 0.0930 0.0324 0.0615 0.0116 0.0068 0.0680 0.2258 0.4069 T   0.2320 0.4655 0.8115 0.1358 0.0293 0.0631 0.4375 0.1999 0.1275
F5H820_HUMAN/495..553	No prediction made	The extracted domain has residue (D) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
F5H820_HUMAN/612..669	No prediction made	The extracted domain has residue (E) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
F5H838_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
F5H8J0_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07

# Appendix A-4 contd.

F6R4Q5_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
F8VSA3_HUMAN/92..152		A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622 C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097 G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140 T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141
F8VSK3_HUMAN/243..306		A   0.2331 0.1862 0.0329 0.1150 0.9785 0.0256 0.3941 0.2947 0.2436 C   0.2117 0.1101 0.0141 0.0001 0.0070 0.4410 0.0780 0.1906 0.2202 G   0.1937 0.1722 0.0228 0.8471 0.0034 0.0585 0.3485 0.2694 0.3604 T   0.3615 0.5314 0.9302 0.0378 0.0111 0.4749 0.1794 0.2454 0.1758
F8VU08_HUMAN/1..24		A   0.2381 0.1853 0.0670 0.7424 0.9831 0.0243 0.1367 0.5249 0.2129 C   0.2325 0.2427 0.0621 0.0000 0.0071 0.1217 0.0921 0.0175 0.2535 G   0.2352 0.1259 0.0533 0.1010 0.0035 0.0476 0.1374 0.4363 0.3466 T   0.2942 0.4461 0.8175 0.1566 0.0063 0.8064 0.6338 0.0213 0.1871
F8VVX3_HUMAN/61..121		A   0.2445 0.2096 0.1271 0.4316 0.9587 0.0205 0.0569 0.5621 0.2182 C   0.2330 0.1654 0.0897 0.0352 0.0052 0.0129 0.0985 0.0624 0.2992 G   0.2606 0.1253 0.0550 0.0718 0.0090 0.0465 0.2375 0.3106 0.3115 T   0.2620 0.4997 0.7282 0.4615 0.0271 0.9201 0.6071 0.0649 0.1712
F8VWZ5_HUMAN/8..68		A   0.2445 0.2096 0.1271 0.4316 0.9587 0.0205 0.0569 0.5621 0.2182 C   0.2330 0.1654 0.0897 0.0352 0.0052 0.0129 0.0985 0.0624 0.2992 G   0.2606 0.1253 0.0550 0.0718 0.0090 0.0465 0.2375 0.3106 0.3115 T   0.2620 0.4997 0.7282 0.4615 0.0271 0.9201 0.6071 0.0649 0.1712
F8VWZ9_HUMAN/1..21		A   0.2583 0.2010 0.0372 0.7846 0.9828 0.0276 0.0965 0.6389 0.2062 C   0.2227 0.1974 0.0304 0.0000 0.0080 0.0015 0.0579 0.0138 0.2991 G   0.2488 0.1189 0.0163 0.0838 0.0026 0.0339 0.2749 0.3091 0.3027 T   0.2703 0.4826 0.9161 0.1316 0.0066 0.9370 0.5707 0.0381 0.1920
F8VXG0_HUMAN/53..113		A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826 C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520 G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637 T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017
F8VXG1_HUMAN/8..68		A   0.1987 0.3362 0.0584 0.9561 0.9824 0.0100 0.0806 0.3828 0.1706 C   0.2379 0.2062 0.0127 0.0160 0.0105 0.0092 0.0670 0.0273 0.3544 G   0.3166 0.1786 0.0059 0.0239 0.0009 0.0302 0.1323 0.5315 0.2907 T   0.2468 0.2791 0.9230 0.0040 0.0062 0.9506 0.7201 0.0585 0.1843
F8VXJ2_HUMAN/127..187		A   0.1987 0.3362 0.0584 0.9561 0.9824 0.0100 0.0806 0.3828 0.1706 C   0.2379 0.2062 0.0127 0.0160 0.0105 0.0092 0.0670 0.0273 0.3544 G   0.3166 0.1786 0.0059 0.0239 0.0009 0.0302 0.1323 0.5315 0.2907 T   0.2468 0.2791 0.9230 0.0040 0.0062 0.9506 0.7201 0.0585 0.1843
F8VXY1_HUMAN/81..137	No prediction made	The extracted domain has residue (H) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
		A   0.2452 0.1830 0.1529 0.7102 0.9874 0.0080 0.0318 0.6574 0.2954

## Appendix A-4 contd.

F8VYP0_HUMAN/218..278		C   0.2423 0.2509 0.0212 0.0057 0.0063 0.0055 0.0256 0.0096 0.2252 G   0.2592 0.1444 0.0178 0.0416 0.0031 0.0211 0.1130 0.2999 0.3132 T   0.2533 0.4217 0.8082 0.2424 0.0032 0.9655 0.8296 0.0330 0.1661
F8WOU5_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
F8W1B5_HUMAN/69..118	No prediction made	The extracted domain has a gap at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
F8W7W6_HUMAN/218..278		A   0.2452 0.1830 0.1529 0.7102 0.9874 0.0080 0.0318 0.6574 0.2954 C   0.2423 0.2509 0.0212 0.0057 0.0063 0.0055 0.0256 0.0096 0.2252 G   0.2592 0.1444 0.0178 0.0416 0.0031 0.0211 0.1130 0.2999 0.3132 T   0.2533 0.4217 0.8082 0.2424 0.0032 0.9655 0.8296 0.0330 0.1661
F8W811_HUMAN/181..226	No prediction made	The extracted domain has a gap at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
F8WBG7_HUMAN/12..72		A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622 C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097 G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140 T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141
G3V138_HUMAN/2128..2188		A   0.2351 0.1770 0.0755 0.8644 0.9854 0.1264 0.6189 0.1422 0.1215 C   0.2654 0.2520 0.0169 0.0187 0.0060 0.0136 0.1701 0.6182 0.2285 G   0.2370 0.1027 0.0337 0.0968 0.0036 0.7031 0.0000 0.0275 0.5260 T   0.2625 0.4683 0.8739 0.0200 0.0050 0.1570 0.2109 0.2121 0.1240
G3V138_HUMAN/2225..2285		A   0.2519 0.1819 0.0275 0.8227 0.9853 0.0203 0.0536 0.6569 0.2579 C   0.2418 0.2492 0.0408 0.0045 0.0075 0.0334 0.0530 0.0141 0.2140 G   0.2715 0.1199 0.0276 0.1135 0.0019 0.0237 0.3889 0.2931 0.3775 T   0.2348 0.4490 0.9041 0.0594 0.0052 0.9226 0.5044 0.0359 0.1507
G3V138_HUMAN/2604..2664		A   0.2342 0.1668 0.0273 0.5568 0.9874 0.0101 0.0138 0.4708 0.2177 C   0.2530 0.3389 0.0029 0.0000 0.0079 0.0084 0.0216 0.0118 0.2707 G   0.2909 0.1231 0.0049 0.4184 0.0017 0.0107 0.0940 0.4760 0.3289 T   0.2220 0.3711 0.9649 0.0247 0.0030 0.9708 0.8706 0.0414 0.1826
G3V138_HUMAN/2928..2988		A   0.2399 0.1483 0.0321 0.9435 0.9809 0.0210 0.0287 0.5619 0.2547 C   0.2282 0.3280 0.0106 0.0110 0.0072 0.0028 0.0488 0.0140 0.2428 G   0.2837 0.1082 0.0044 0.0325 0.0037 0.0484 0.1220 0.3763 0.3085 T   0.2304 0.4155 0.9529 0.0129 0.0083 0.9279 0.8005 0.0478 0.1940
G3V1R3_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
G3V222_HUMAN/36..70		A   0.2421 0.2191 0.0447 0.6895 0.9791 0.0360 0.0390 0.3784 0.2049 C   0.2282 0.1878 0.0643 0.0000 0.0070 0.0639 0.1742 0.0254 0.2888 G   0.2438 0.1422 0.0597 0.0896 0.0042 0.0280 0.3170 0.5490 0.3218 T   0.2859 0.4509 0.8313 0.2209 0.0097 0.8722 0.4698 0.0472 0.1844
G3V243_HUMAN/72..132		A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849 C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028 G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384 T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739



# Appendix A-4 contd.

G3V2N2_HUMAN/217..274		<div>A   0.2537 0.3059 0.0696 0.0818 0.9881 0.0268 0.7430 0.0787 0.1403</div> <div>C   0.2413 0.1870 0.1236 0.0211 0.0087 0.1361 0.2206 0.7112 0.4748</div> <div>G   0.2467 0.2303 0.0879 0.8057 0.0006 0.0668 0.0034 0.0544 0.2127</div> <div>T   0.2583 0.2768 0.7190 0.0915 0.0025 0.7703 0.0331 0.1557 0.1722</div>
G3V2N3_HUMAN/93..153		<div>A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810</div> <div>C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916</div> <div>G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810</div> <div>T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464</div>
G3V2N9_HUMAN/60..120		<div>A   0.2074 0.1845 0.0610 0.8782 0.9839 0.0062 0.0325 0.3769 0.2463</div> <div>C   0.2821 0.2124 0.0406 0.0097 0.0086 0.0161 0.0316 0.1065 0.1939</div> <div>G   0.2880 0.1005 0.0107 0.0617 0.0018 0.0151 0.0904 0.4114 0.3923</div> <div>T   0.2226 0.5026 0.8878 0.0505 0.0056 0.9626 0.8455 0.1051 0.1674</div>
G3V2X8_HUMAN/184..244		<div>A   0.2191 0.1761 0.0263 0.9723 0.9856 0.0560 0.1117 0.3591 0.1986</div> <div>C   0.2675 0.3558 0.0072 0.0080 0.0073 0.0000 0.0736 0.0192 0.2784</div> <div>G   0.2983 0.1247 0.0020 0.0127 0.0030 0.0310 0.1989 0.5711 0.3360</div> <div>T   0.2151 0.3433 0.9644 0.0070 0.0041 0.9130 0.6158 0.0506 0.1870</div>
G3V309_HUMAN/18..78		<div>A   0.2302 0.1742 0.0599 0.3944 0.9753 0.0104 0.0717 0.4626 0.2097</div> <div>C   0.2493 0.2014 0.0340 0.0104 0.0108 0.0410 0.0443 0.0158 0.2800</div> <div>G   0.2460 0.1514 0.0210 0.5221 0.0024 0.1015 0.1874 0.4727 0.3175</div> <div>T   0.2745 0.4730 0.8850 0.0731 0.0115 0.8471 0.6966 0.0490 0.1927</div>
G3V309_HUMAN/93..151		<div>A   0.2372 0.1661 0.0463 0.3135 0.9856 0.0073 0.0133 0.4561 0.2499</div> <div>C   0.2490 0.2693 0.0091 0.0000 0.0076 0.0054 0.0175 0.0124 0.2019</div> <div>G   0.2650 0.1353 0.0085 0.6468 0.0024 0.0131 0.0804 0.5029 0.3823</div> <div>T   0.2487 0.4294 0.9361 0.0397 0.0044 0.9742 0.8888 0.0285 0.1659</div>
G3V397_HUMAN/1..47		<div>A   0.2366 0.2247 0.0435 0.6936 0.9794 0.0514 0.0436 0.3177 0.1926</div> <div>C   0.2288 0.1894 0.0640 0.0000 0.0071 0.0566 0.1835 0.0302 0.2993</div> <div>G   0.2452 0.1460 0.0603 0.0876 0.0041 0.0384 0.3025 0.5936 0.3241</div> <div>T   0.2894 0.4399 0.8321 0.2187 0.0094 0.8536 0.4705 0.0586 0.1840</div>
G3V3J3_HUMAN/37..97		<div>A   0.2271 0.2000 0.0400 0.9635 0.9883 0.0002 0.0349 0.0327 0.1395</div> <div>C   0.2471 0.2969 0.0041 0.0098 0.0085 0.0051 0.9108 0.7577 0.3630</div> <div>G   0.3047 0.1251 0.0091 0.0116 0.0013 0.0741 0.0043 0.0403 0.2830</div> <div>T   0.2211 0.3780 0.9468 0.0151 0.0019 0.9206 0.0500 0.1693 0.2146</div>
G3V3P9_HUMAN/45..105		<div>A   0.2271 0.2000 0.0400 0.9635 0.9883 0.0002 0.0349 0.0327 0.1395</div> <div>C   0.2471 0.2969 0.0041 0.0098 0.0085 0.0051 0.9108 0.7577 0.3630</div> <div>G   0.3047 0.1251 0.0091 0.0116 0.0013 0.0741 0.0043 0.0403 0.2830</div> <div>T   0.2211 0.3780 0.9468 0.0151 0.0019 0.9206 0.0500 0.1693 0.2146</div>
G3V3Q9_HUMAN/164..224		<div>A   0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535</div> <div>C   0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239</div> <div>G   0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676</div> <div>T   0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550</div>
G3V469_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmssearch and a domE cut off of 1e-07
G3V471_HUMAN/38..98		<div>A   0.1987 0.3362 0.0584 0.9561 0.9824 0.0100 0.0806 0.3828 0.1706</div> <div>C   0.2379 0.2062 0.0127 0.0160 0.0105 0.0092 0.0670 0.0273 0.3544</div> <div>G   0.3166 0.1786 0.0059 0.0239 0.0009 0.0302 0.1323 0.5315 0.2907</div> <div>T   0.2468 0.2791 0.9230 0.0040 0.0062 0.9506 0.7201 0.0585 0.1843</div>

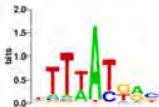

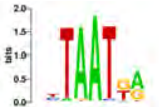
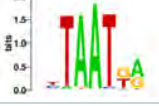
# Appendix A-4 contd.

G3V4M3_HUMAN/38..98		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535  0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239  0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676  0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550 </div>
G3V4N4_HUMAN/73..133		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535  0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239  0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676  0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550 </div>
G3V4Q1_HUMAN/159..219		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535  0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239  0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676  0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550 </div>
G3V4R6_HUMAN/46..93	No prediction made	The extracted domain has a gap at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
G3V567_HUMAN/8..68		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535  0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239  0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676  0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550 </div>
G3V5W7_HUMAN/73..133		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535  0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239  0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676  0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550 </div>
G5E9C1_HUMAN/218..278		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2275 0.2398 0.0596 0.7268 0.9881 0.0157 0.1254 0.2813 0.2523  0.2571 0.2797 0.0096 0.0030 0.0072 0.0173 0.2354 0.2483 0.1896  0.3102 0.1459 0.0148 0.2523 0.0021 0.0099 0.1297 0.3206 0.4273  0.2052 0.3345 0.9160 0.0179 0.0026 0.9571 0.5094 0.1498 0.1307 </div>
GBX1_HUMAN/260..320		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2234 0.1770 0.0728 0.8774 0.9850 0.0102 0.0588 0.5195 0.2173  0.2655 0.3072 0.0302 0.0089 0.0058 0.0176 0.0460 0.0165 0.2762  0.2906 0.1531 0.0144 0.0945 0.0043 0.0158 0.1151 0.4319 0.3286  0.2206 0.3627 0.8826 0.0193 0.0049 0.9564 0.7801 0.0320 0.1779 </div>
GBX2_HUMAN/246..306		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2234 0.1770 0.0728 0.8774 0.9850 0.0102 0.0588 0.5195 0.2173  0.2655 0.3072 0.0302 0.0089 0.0058 0.0176 0.0460 0.0165 0.2762  0.2906 0.1531 0.0144 0.0945 0.0043 0.0158 0.1151 0.4319 0.3286  0.2206 0.3627 0.8826 0.0193 0.0049 0.9564 0.7801 0.0320 0.1779 </div>
GSC2_HUMAN/125..185		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2271 0.2000 0.0400 0.9635 0.9883 0.0002 0.0349 0.0327 0.1395  0.2471 0.2969 0.0041 0.0098 0.0085 0.0051 0.9108 0.7577 0.3630  0.3047 0.1251 0.0091 0.0116 0.0013 0.0741 0.0043 0.0403 0.2830  0.2211 0.3780 0.9468 0.0151 0.0019 0.9206 0.0500 0.1693 0.2146 </div>
GSC_HUMAN/159..219		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2271 0.2000 0.0400 0.9635 0.9883 0.0002 0.0349 0.0327 0.1395  0.2471 0.2969 0.0041 0.0098 0.0085 0.0051 0.9108 0.7577 0.3630  0.3047 0.1251 0.0091 0.0116 0.0013 0.0741 0.0043 0.0403 0.2830  0.2211 0.3780 0.9468 0.0151 0.0019 0.9206 0.0500 0.1693 0.2146 </div>
GSX1_HUMAN/146..206		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2722 0.0944 0.0261 0.9366 0.9863 0.0208 0.0528 0.6940 0.1640  0.2329 0.3918 0.0060 0.0127 0.0063 0.0498 0.0518 0.0237 0.2775  0.3234 0.0880 0.0032 0.0256 0.0036 0.0337 0.2745 0.2522 0.3798  0.1714 0.4259 0.9647 0.0251 0.0037 0.8958 0.6208 0.0301 0.1787 </div>









# Appendix A-4 contd.

GSX2_HUMAN/201..261		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2722 0.0944 0.0261 0.9366 0.9863 0.0208 0.0528 0.6940 0.1640  0.2329 0.3918 0.0060 0.0127 0.0063 0.0498 0.0518 0.0237 0.2775  0.3234 0.0880 0.0032 0.0256 0.0036 0.0337 0.2745 0.2522 0.3798  0.1714 0.4259 0.9647 0.0251 0.0037 0.8958 0.6208 0.0301 0.1787 </div>
HDX_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmssearch and a domE cut off of 1e-07
HESX1_HUMAN/107..167		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2141 0.1981 0.0597 0.9105 0.9847 0.0095 0.0563 0.3875 0.2324  0.2651 0.3035 0.0226 0.0060 0.0060 0.0122 0.0522 0.0473 0.2407  0.3031 0.1458 0.0108 0.0630 0.0039 0.0121 0.0863 0.4951 0.3563  0.2176 0.3526 0.9068 0.0205 0.0054 0.9662 0.8052 0.0701 0.1706 </div>
HHEX_HUMAN/136..196		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2210 0.1638 0.0785 0.6865 0.9798 0.0475 0.1521 0.4341 0.1991  0.2338 0.1882 0.1114 0.0000 0.0068 0.1009 0.0851 0.0203 0.2706  0.2209 0.1197 0.1033 0.1052 0.0040 0.1031 0.1491 0.5147 0.3487  0.3243 0.5283 0.7068 0.2083 0.0094 0.7486 0.6137 0.0309 0.1817 </div>
HLX_HUMAN/275..335		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2445 0.2096 0.1271 0.4316 0.9587 0.0205 0.0569 0.5621 0.2182  0.2330 0.1654 0.0897 0.0352 0.0052 0.0129 0.0985 0.0624 0.2992  0.2606 0.1253 0.0550 0.0718 0.0090 0.0465 0.2375 0.3106 0.3115  0.2620 0.4997 0.7282 0.4615 0.0271 0.9201 0.6071 0.0649 0.1712 </div>
HMBX1_HUMAN/266..341		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2260 0.2126 0.1212 0.6215 0.9503 0.0503 0.1019 0.5213 0.2296  0.2767 0.2221 0.1232 0.0477 0.0000 0.8548 0.3955 0.1597 0.1796  0.2235 0.1424 0.0738 0.1213 0.0130 0.0000 0.0753 0.1542 0.4496  0.2738 0.4229 0.6818 0.2096 0.0367 0.0949 0.4273 0.1648 0.1413 </div>
HME1_HUMAN/302..362		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2234 0.1770 0.0728 0.8774 0.9850 0.0102 0.0588 0.5195 0.2173  0.2655 0.3072 0.0302 0.0089 0.0058 0.0176 0.0460 0.0165 0.2762  0.2906 0.1531 0.0144 0.0945 0.0043 0.0158 0.1151 0.4319 0.3286  0.2206 0.3627 0.8826 0.0193 0.0049 0.9564 0.7801 0.0320 0.1779 </div>
HME2_HUMAN/243..303		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2234 0.1770 0.0728 0.8774 0.9850 0.0102 0.0588 0.5195 0.2173  0.2655 0.3072 0.0302 0.0089 0.0058 0.0176 0.0460 0.0165 0.2762  0.2906 0.1531 0.0144 0.0945 0.0043 0.0158 0.1151 0.4319 0.3286  0.2206 0.3627 0.8826 0.0193 0.0049 0.9564 0.7801 0.0320 0.1779 </div>
HMX1_HUMAN/202..262		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2401 0.1582 0.0836 0.9092 0.9833 0.0388 0.0116 0.1081 0.1841  0.2438 0.1338 0.0473 0.0073 0.0093 0.0188 0.1793 0.0031 0.3333  0.2630 0.1023 0.0325 0.0466 0.0017 0.1763 0.0612 0.8516 0.2754  0.2531 0.6057 0.8366 0.0368 0.0057 0.7660 0.7479 0.0372 0.2072 </div>
HMX2_HUMAN/148..208		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2378 0.1675 0.0771 0.9271 0.9864 0.1035 0.0234 0.1104 0.1835  0.2473 0.1396 0.0472 0.0000 0.0081 0.0794 0.2258 0.0094 0.2838  0.2664 0.0909 0.0236 0.0397 0.0022 0.1250 0.0662 0.8468 0.3293  0.2485 0.6020 0.8520 0.0331 0.0034 0.6921 0.6846 0.0335 0.2034 </div>
HMX3_HUMAN/226..286		<div> <div>A</div> <div>C</div> <div>G</div> <div>T</div> </div> <div> 0.2401 0.1582 0.0836 0.9092 0.9833 0.0388 0.0116 0.1081 0.1841  0.2438 0.1338 0.0473 0.0073 0.0093 0.0188 0.1793 0.0031 0.3333  0.2630 0.1023 0.0325 0.0466 0.0017 0.1763 0.0612 0.8516 0.2754  0.2531 0.6057 0.8366 0.0368 0.0057 0.7660 0.7479 0.0372 0.2072 </div>
HNF1A_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmssearch and a domE cut off of 1e-07
HNF1B_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmssearch and a domE cut off of 1e-07

# Appendix A-4 contd.

HNF6_HUMAN/384..443		<div>A   0.1931 0.1113 0.0222 0.2972 0.9745 0.0000 0.2299 0.4554 0.2274</div> <div>C   0.2696 0.1448 0.0114 0.0000 0.0157 0.3179 0.0929 0.1078 0.2631</div> <div>G   0.2480 0.1308 0.0526 0.6561 0.0000 0.0649 0.0780 0.3108 0.2899</div> <div>T   0.2893 0.6130 0.9138 0.0467 0.0099 0.6172 0.5991 0.1260 0.2195</div>
HOP_HUMAN/3..62	No prediction made	The extracted domain has residue (Q) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
HXA10_HUMAN/335..395		<div>A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006</div> <div>C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872</div> <div>G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068</div> <div>T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053</div>
HXA11_HUMAN/240..300		<div>A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006</div> <div>C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872</div> <div>G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068</div> <div>T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053</div>
HXA13_HUMAN/321..381		<div>A   0.2148 0.0476 0.0387 0.2264 0.9767 0.0135 0.1978 0.5543 0.0909</div> <div>C   0.1365 0.0246 0.0194 0.0161 0.0053 0.2529 0.0163 0.0215 0.2332</div> <div>G   0.1714 0.0213 0.0008 0.0326 0.0062 0.0432 0.4201 0.3712 0.5099</div> <div>T   0.4773 0.9065 0.9411 0.7249 0.0118 0.6904 0.3658 0.0531 0.1661</div>
HXA1_HUMAN/228..288		<div>A   0.2388 0.1524 0.0789 0.8548 0.9809 0.0153 0.0341 0.6234 0.1776</div> <div>C   0.2408 0.2590 0.0302 0.0100 0.0064 0.0150 0.1060 0.0090 0.3514</div> <div>G   0.2657 0.1690 0.0330 0.0811 0.0039 0.0015 0.3173 0.3343 0.3038</div> <div>T   0.2547 0.4196 0.8580 0.0541 0.0088 0.9682 0.5426 0.0333 0.1672</div>
HXA2_HUMAN/142..202		<div>A   0.2350 0.1336 0.0185 0.9687 0.9862 0.0056 0.0284 0.7018 0.1918</div> <div>C   0.2644 0.3754 0.0073 0.0092 0.0076 0.0315 0.0744 0.0197 0.3114</div> <div>G   0.2964 0.1184 0.0071 0.0189 0.0024 0.0112 0.4242 0.2305 0.3153</div> <div>T   0.2042 0.3725 0.9672 0.0032 0.0038 0.9517 0.4731 0.0480 0.1814</div>
HXA3_HUMAN/190..250		<div>A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826</div> <div>C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520</div> <div>G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637</div> <div>T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017</div>
HXA4_HUMAN/214..274		<div>A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826</div> <div>C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520</div> <div>G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637</div> <div>T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017</div>
HXA5_HUMAN/194..254		<div>A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826</div> <div>C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520</div> <div>G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637</div> <div>T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017</div>
HXA6_HUMAN/154..214		<div>A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622</div> <div>C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097</div> <div>G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140</div> <div>T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141</div>
		<div>A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622</div>

# Appendix A-4 contd.

HXA7_HUMAN/129..189		<div>C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097</div> <div>G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140</div> <div>T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141</div>
HXA9_HUMAN/205..265		<div>A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006</div> <div>C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872</div> <div>G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068</div> <div>T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053</div>
HXB13_HUMAN/215..275		<div>A   0.2128 0.0493 0.0343 0.2878 0.9776 0.0112 0.1990 0.5397 0.1105</div> <div>C   0.1383 0.0440 0.0179 0.0066 0.0045 0.2568 0.0179 0.0165 0.2212</div> <div>G   0.1794 0.0343 0.0132 0.0310 0.0066 0.0218 0.3812 0.3905 0.4973</div> <div>T   0.4696 0.8724 0.9345 0.6746 0.0113 0.7103 0.4020 0.0533 0.1710</div>
HXB1_HUMAN/203..262		<div>A   0.2336 0.1371 0.0596 0.8873 0.9778 0.0122 0.0420 0.6226 0.1879</div> <div>C   0.2396 0.2738 0.0441 0.0093 0.0064 0.0234 0.0696 0.0160 0.3520</div> <div>G   0.2672 0.1233 0.0366 0.0476 0.0047 0.0096 0.4171 0.3255 0.2770</div> <div>T   0.2596 0.4658 0.8597 0.0558 0.0111 0.9547 0.4713 0.0360 0.1929</div>
HXB2_HUMAN/142..202		<div>A   0.2350 0.1336 0.0185 0.9687 0.9862 0.0056 0.0284 0.7018 0.1918</div> <div>C   0.2644 0.3754 0.0073 0.0092 0.0076 0.0315 0.0744 0.0197 0.3114</div> <div>G   0.2964 0.1184 0.0071 0.0189 0.0024 0.0112 0.4242 0.2305 0.3153</div> <div>T   0.2042 0.3725 0.9672 0.0032 0.0038 0.9517 0.4731 0.0480 0.1814</div>
HXB3_HUMAN/187..247		<div>A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826</div> <div>C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520</div> <div>G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637</div> <div>T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017</div>
HXB4_HUMAN/161..221		<div>A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826</div> <div>C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520</div> <div>G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637</div> <div>T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017</div>
HXB5_HUMAN/193..253		<div>A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826</div> <div>C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520</div> <div>G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637</div> <div>T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017</div>
HXB6_HUMAN/145..205		<div>A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622</div> <div>C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097</div> <div>G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140</div> <div>T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141</div>
HXB7_HUMAN/136..196		<div>A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622</div> <div>C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097</div> <div>G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140</div> <div>T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141</div>
HXB8_HUMAN/145..205		<div>A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622</div> <div>C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097</div> <div>G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140</div> <div>T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141</div>
		<div>A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006</div> <div>C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872</div>

## Appendix A-4 contd.

HXB9_HUMAN/184..244		G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068 T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053
HXC10_HUMAN/267..327		A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006 C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872 G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068 T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053
HXC11_HUMAN/231..291		A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006 C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872 G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068 T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053
HXC12_HUMAN/213..273		A   0.2252 0.0936 0.0511 0.3292 0.9699 0.0074 0.0978 0.6179 0.1175 C   0.1636 0.0518 0.0120 0.0040 0.0041 0.3050 0.0519 0.0307 0.4384 G   0.1837 0.0610 0.0177 0.0424 0.0082 0.0291 0.4848 0.2911 0.2165 T   0.4275 0.7937 0.9192 0.6244 0.0178 0.6584 0.3655 0.0603 0.2276
HXC13_HUMAN/259..319		A   0.2148 0.0476 0.0387 0.2264 0.9767 0.0135 0.1978 0.5543 0.0909 C   0.1365 0.0246 0.0194 0.0161 0.0053 0.2529 0.0163 0.0215 0.2332 G   0.1714 0.0213 0.0008 0.0326 0.0062 0.0432 0.4201 0.3712 0.5099 T   0.4773 0.9065 0.9411 0.7249 0.0118 0.6904 0.3658 0.0531 0.1661
HXC4_HUMAN/155..215		A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826 C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520 G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637 T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017
HXC5_HUMAN/154..214		A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826 C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520 G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637 T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017
HXC6_HUMAN/140..200		A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622 C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097 G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140 T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141
HXC8_HUMAN/148..208		A   0.2099 0.1346 0.0644 0.6483 0.9649 0.0164 0.0646 0.5106 0.1519 C   0.2343 0.1484 0.0415 0.0154 0.0038 0.0809 0.0638 0.0341 0.4405 G   0.2189 0.1171 0.0531 0.0376 0.0095 0.0172 0.3966 0.3866 0.1973 T   0.3369 0.5998 0.8410 0.2987 0.0218 0.8856 0.4750 0.0687 0.2104
HXC9_HUMAN/191..251		A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006 C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872 G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068 T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053
HXD10_HUMAN/265..325		A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006 C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627

# Appendix A-4 contd.

HXD12_HUMAN/201..261		A   0.2252 0.0936 0.0511 0.3292 0.9699 0.0074 0.0978 0.6179 0.1175 C   0.1636 0.0518 0.0120 0.0040 0.0041 0.3050 0.0519 0.0307 0.4384 G   0.1837 0.0610 0.0177 0.0424 0.0082 0.0291 0.4848 0.2911 0.2165 T   0.4275 0.7937 0.9192 0.6244 0.0178 0.6584 0.3655 0.0603 0.2276
HXD13_HUMAN/275..335		A   0.2148 0.0476 0.0387 0.2264 0.9767 0.0135 0.1978 0.5543 0.0909 C   0.1365 0.0246 0.0194 0.0161 0.0053 0.2529 0.0163 0.0215 0.2332 G   0.1714 0.0213 0.0008 0.0326 0.0062 0.0432 0.4201 0.3712 0.5099 T   0.4773 0.9065 0.9411 0.7249 0.0118 0.6904 0.3658 0.0531 0.1661
HXD1_HUMAN/229..288		A   0.2388 0.1524 0.0789 0.8548 0.9809 0.0153 0.0341 0.6234 0.1776 C   0.2408 0.2590 0.0302 0.0100 0.0064 0.0150 0.1060 0.0090 0.3514 G   0.2657 0.1690 0.0330 0.0811 0.0039 0.0015 0.3173 0.3343 0.3038 T   0.2547 0.4196 0.8580 0.0541 0.0088 0.9682 0.5426 0.0333 0.1672
HXD3_HUMAN/193..253		A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826 C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520 G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637 T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017
HXD4_HUMAN/153..213		A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826 C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520 G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637 T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017
HXD8_HUMAN/196..256		A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622 C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097 G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140 T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141
HXD9_HUMAN/284..344		A   0.1976 0.0924 0.0713 0.3587 0.9481 0.0093 0.0841 0.4919 0.1006 C   0.1817 0.0802 0.0527 0.0226 0.0000 0.2461 0.0627 0.0651 0.4872 G   0.1645 0.0754 0.0438 0.0385 0.0179 0.0276 0.5222 0.3438 0.2068 T   0.4561 0.7520 0.8322 0.5802 0.0340 0.7169 0.3309 0.0992 0.2053
ISL1_HUMAN/180..240		A   0.2524 0.1439 0.0520 0.9014 0.9711 0.0836 0.1018 0.3286 0.2510 C   0.2238 0.3478 0.0234 0.0054 0.0061 0.0148 0.0343 0.0514 0.2630 G   0.3062 0.1284 0.0219 0.0829 0.0057 0.1413 0.3529 0.5412 0.2866 T   0.2176 0.3800 0.9027 0.0104 0.0171 0.7603 0.5110 0.0789 0.1994
ISL2_HUMAN/190..250		A   0.2524 0.1439 0.0520 0.9014 0.9711 0.0836 0.1018 0.3286 0.2510 C   0.2238 0.3478 0.0234 0.0054 0.0061 0.0148 0.0343 0.0514 0.2630 G   0.3062 0.1284 0.0219 0.0829 0.0057 0.1413 0.3529 0.5412 0.2866 T   0.2176 0.3800 0.9027 0.0104 0.0171 0.7603 0.5110 0.0789 0.1994
ISX_HUMAN/81..141		A   0.2185 0.1736 0.0641 0.8991 0.9846 0.0091 0.0574 0.3465 0.2224 C   0.2672 0.3063 0.0229 0.0082 0.0075 0.0099 0.0477 0.0886 0.2461 G   0.2917 0.1363 0.0105 0.0712 0.0032 0.0131 0.0908 0.4519 0.3464 T   0.2227 0.3838 0.9025 0.0214 0.0048 0.9679 0.8041 0.1131 0.1851
LBX1_HUMAN/124..184		A   0.2059 0.1908 0.0454 0.9553 0.9856 0.0293 0.1458 0.6771 0.2052 C   0.2493 0.2771 0.0518 0.0190 0.0078 0.2258 0.1411 0.0375 0.2281 G   0.3054 0.1065 0.0233 0.0179 0.0025 0.0660 0.2322 0.2629 0.3712 T   0.2395 0.4255 0.8796 0.0078 0.0041 0.6790 0.4810 0.0225 0.1955



# Appendix A-4 contd.

LBX2_HUMAN/84..144		A   0.2059 0.1908 0.0454 0.9553 0.9856 0.0293 0.1458 0.6771 0.2052 C   0.2493 0.2771 0.0518 0.0190 0.0078 0.2258 0.1411 0.0375 0.2281 G   0.3054 0.1065 0.0233 0.0179 0.0025 0.0660 0.2322 0.2629 0.3712 T   0.2395 0.4255 0.8796 0.0078 0.0041 0.6790 0.4810 0.0225 0.1955
LEUTX_HUMAN/1..37		A   0.2376 0.2048 0.0760 0.6848 0.9823 0.0094 0.1518 0.0243 0.1691 C   0.2360 0.1920 0.0452 0.0000 0.0078 0.0278 0.7536 0.7188 0.3782 G   0.2459 0.1531 0.0317 0.1196 0.0031 0.0328 0.0000 0.0735 0.2569 T   0.2804 0.4501 0.8471 0.1956 0.0068 0.9301 0.0946 0.1834 0.1959
LHX1_HUMAN/179..239		A   0.2443 0.2089 0.0512 0.9422 0.9841 0.0132 0.0474 0.6244 0.2746 C   0.2498 0.2231 0.0479 0.0078 0.0077 0.0059 0.0327 0.0121 0.2351 G   0.2533 0.1376 0.0275 0.0353 0.0029 0.0190 0.1259 0.3347 0.3081 T   0.2526 0.4304 0.8733 0.0147 0.0054 0.9619 0.7940 0.0289 0.1822
LHX2_HUMAN/265..325		A   0.2538 0.1440 0.0297 0.9307 0.9844 0.0031 0.0704 0.6998 0.2085 C   0.2443 0.2930 0.0144 0.0090 0.0070 0.0453 0.0294 0.0101 0.1934 G   0.2605 0.1008 0.0026 0.0354 0.0030 0.0078 0.1952 0.2659 0.4121 T   0.2415 0.4623 0.9533 0.0250 0.0057 0.9438 0.7050 0.0241 0.1860
LHX3_HUMAN/156..216		A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849 C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028 G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384 T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739
LHX4_HUMAN/156..216		A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849 C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028 G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384 T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739
LHX5_HUMAN/179..239		A   0.2443 0.2089 0.0512 0.9422 0.9841 0.0132 0.0474 0.6244 0.2746 C   0.2498 0.2231 0.0479 0.0078 0.0077 0.0059 0.0327 0.0121 0.2351 G   0.2533 0.1376 0.0275 0.0353 0.0029 0.0190 0.1259 0.3347 0.3081 T   0.2526 0.4304 0.8733 0.0147 0.0054 0.9619 0.7940 0.0289 0.1822
LHX6_HUMAN/218..278		A   0.2345 0.1541 0.0242 0.5016 0.9882 0.0068 0.0121 0.5322 0.2058 C   0.2632 0.3621 0.0023 0.0019 0.0075 0.0041 0.0125 0.0084 0.2960 G   0.3123 0.1258 0.0047 0.4749 0.0018 0.0058 0.0753 0.4280 0.3341 T   0.1900 0.3580 0.9688 0.0216 0.0025 0.9833 0.9001 0.0314 0.1641
LHX8_HUMAN/224..284		A   0.2345 0.1541 0.0242 0.5016 0.9882 0.0068 0.0121 0.5322 0.2058 C   0.2632 0.3621 0.0023 0.0019 0.0075 0.0041 0.0125 0.0084 0.2960 G   0.3123 0.1258 0.0047 0.4749 0.0018 0.0058 0.0753 0.4280 0.3341 T   0.1900 0.3580 0.9688 0.0216 0.0025 0.9833 0.9001 0.0314 0.1641
LHX9_HUMAN/266..326		A   0.2538 0.1440 0.0297 0.9307 0.9844 0.0031 0.0704 0.6998 0.2085 C   0.2443 0.2930 0.0144 0.0090 0.0070 0.0453 0.0294 0.0101 0.1934 G   0.2605 0.1008 0.0026 0.0354 0.0030 0.0078 0.1952 0.2659 0.4121 T   0.2415 0.4623 0.9533 0.0250 0.0057 0.9438 0.7050 0.0241 0.1860
LMX1A_HUMAN/194..254		A   0.2452 0.1830 0.1529 0.7102 0.9874 0.0080 0.0318 0.6574 0.2954 C   0.2423 0.2509 0.0212 0.0057 0.0063 0.0055 0.0256 0.0096 0.2252 G   0.2592 0.1444 0.0178 0.0416 0.0031 0.0211 0.1130 0.2999 0.3132 T   0.2533 0.4217 0.8082 0.2424 0.0032 0.9655 0.8296 0.0330 0.1661
LMX1B_HUMAN/195..255		A   0.2452 0.1830 0.1529 0.7102 0.9874 0.0080 0.0318 0.6574 0.2954 C   0.2423 0.2509 0.0212 0.0057 0.0063 0.0055 0.0256 0.0096 0.2252



# Appendix A-4 contd.

		<div>G   0.2592 0.1444 0.0178 0.0416 0.0031 0.0211 0.1130 0.2999 0.3132</div> <div>T   0.2533 0.4217 0.8082 0.2424 0.0032 0.9655 0.8296 0.0330 0.1661</div>
MEOX1_HUMAN/170..230		<div>A   0.2344 0.1869 0.0226 0.9033 0.9814 0.0073 0.0113 0.7340 0.1894</div> <div>C   0.2384 0.2464 0.0173 0.0095 0.0072 0.0342 0.0514 0.0071 0.3376</div> <div>G   0.2724 0.1015 0.0038 0.0480 0.0035 0.0164 0.4441 0.2224 0.2769</div> <div>T   0.2548 0.4652 0.9563 0.0392 0.0080 0.9421 0.4932 0.0365 0.1961</div>
MEOX2_HUMAN/186..246		<div>A   0.2344 0.1869 0.0226 0.9033 0.9814 0.0073 0.0113 0.7340 0.1894</div> <div>C   0.2384 0.2464 0.0173 0.0095 0.0072 0.0342 0.0514 0.0071 0.3376</div> <div>G   0.2724 0.1015 0.0038 0.0480 0.0035 0.0164 0.4441 0.2224 0.2769</div> <div>T   0.2548 0.4652 0.9563 0.0392 0.0080 0.9421 0.4932 0.0365 0.1961</div>
MIXL1_HUMAN/85..145		<div>A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810</div> <div>C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916</div> <div>G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810</div> <div>T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464</div>
MNX1_HUMAN/240..300		<div>A   0.2256 0.2329 0.0320 0.9445 0.9838 0.0132 0.0465 0.4633 0.2032</div> <div>C   0.2579 0.2875 0.0050 0.0096 0.0080 0.0087 0.1394 0.0207 0.3161</div> <div>G   0.3030 0.1898 0.0023 0.0331 0.0024 0.0090 0.2441 0.4508 0.2789</div> <div>T   0.2135 0.2899 0.9607 0.0128 0.0058 0.9691 0.5701 0.0652 0.2018</div>
MSX1_HUMAN/165..225		<div>A   0.2035 0.2496 0.0360 0.9101 0.9819 0.0093 0.0215 0.3239 0.1946</div> <div>C   0.2511 0.2627 0.0260 0.0253 0.0095 0.0029 0.0284 0.0150 0.2866</div> <div>G   0.2945 0.1800 0.0072 0.0332 0.0017 0.0097 0.0509 0.6125 0.3373</div> <div>T   0.2509 0.3077 0.9308 0.0313 0.0070 0.9781 0.8992 0.0485 0.1815</div>
MSX2_HUMAN/141..201		<div>A   0.2035 0.2496 0.0360 0.9101 0.9819 0.0093 0.0215 0.3239 0.1946</div> <div>C   0.2511 0.2627 0.0260 0.0253 0.0095 0.0029 0.0284 0.0150 0.2866</div> <div>G   0.2945 0.1800 0.0072 0.0332 0.0017 0.0097 0.0509 0.6125 0.3373</div> <div>T   0.2509 0.3077 0.9308 0.0313 0.0070 0.9781 0.8992 0.0485 0.1815</div>
NANG2_HUMAN/37..97		<div>A   0.2184 0.2379 0.0271 0.9604 0.9834 0.0366 0.0387 0.4661 0.1915</div> <div>C   0.2319 0.2329 0.0092 0.0105 0.0081 0.1007 0.1784 0.0222 0.2702</div> <div>G   0.3183 0.1328 0.0103 0.0124 0.0025 0.0358 0.3706 0.4677 0.3421</div> <div>T   0.2314 0.3965 0.9534 0.0167 0.0060 0.8269 0.4122 0.0439 0.1962</div>
NANGN_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmssearch and a domE cut off of 1e-07
NANOG_HUMAN/94..154		<div>A   0.2184 0.2379 0.0271 0.9604 0.9834 0.0366 0.0387 0.4661 0.1915</div> <div>C   0.2319 0.2329 0.0092 0.0105 0.0081 0.1007 0.1784 0.0222 0.2702</div> <div>G   0.3183 0.1328 0.0103 0.0124 0.0025 0.0358 0.3706 0.4677 0.3421</div> <div>T   0.2314 0.3965 0.9534 0.0167 0.0060 0.8269 0.4122 0.0439 0.1962</div>
NANP8_HUMAN/94..154		<div>A   0.2184 0.2379 0.0271 0.9604 0.9834 0.0366 0.0387 0.4661 0.1915</div> <div>C   0.2319 0.2329 0.0092 0.0105 0.0081 0.1007 0.1784 0.0222 0.2702</div> <div>G   0.3183 0.1328 0.0103 0.0124 0.0025 0.0358 0.3706 0.4677 0.3421</div> <div>T   0.2314 0.3965 0.9534 0.0167 0.0060 0.8269 0.4122 0.0439 0.1962</div>
NKX11_HUMAN/258..318		<div>A   0.2191 0.1761 0.0263 0.9723 0.9856 0.0560 0.1117 0.3591 0.1986</div> <div>C   0.2675 0.3558 0.0072 0.0080 0.0073 0.0000 0.0736 0.0192 0.2784</div> <div>G   0.2983 0.1247 0.0020 0.0127 0.0030 0.0310 0.1989 0.5711 0.3360</div> <div>T   0.2151 0.3433 0.9644 0.0070 0.0041 0.9130 0.6158 0.0506 0.1870</div>


# Appendix A-4 contd.

NKX12_HUMAN/162..222		<div>A   0.2191 0.1761 0.0263 0.9723 0.9856 0.0560 0.1117 0.3591 0.1986</div> <div>C   0.2675 0.3558 0.0072 0.0080 0.0073 0.0000 0.0736 0.0192 0.2784</div> <div>G   0.2983 0.1247 0.0020 0.0127 0.0030 0.0310 0.1989 0.5711 0.3360</div> <div>T   0.2151 0.3433 0.9644 0.0070 0.0041 0.9130 0.6158 0.0506 0.1870</div>
NKX21_HUMAN/160..220		<div>A   0.2060 0.1792 0.0541 0.7949 0.9824 0.0017 0.1054 0.2634 0.0920</div> <div>C   0.2620 0.1057 0.3395 0.0024 0.0098 0.0087 0.0094 0.0028 0.3330</div> <div>G   0.2375 0.0509 0.1348 0.1574 0.0015 0.9296 0.0623 0.7215 0.3739</div> <div>T   0.2946 0.6641 0.4715 0.0453 0.0063 0.0600 0.8229 0.0123 0.2012</div>
NKX22_HUMAN/127..187		<div>A   0.2060 0.1792 0.0541 0.7949 0.9824 0.0017 0.1054 0.2634 0.0920</div> <div>C   0.2620 0.1057 0.3395 0.0024 0.0098 0.0087 0.0094 0.0028 0.3330</div> <div>G   0.2375 0.0509 0.1348 0.1574 0.0015 0.9296 0.0623 0.7215 0.3739</div> <div>T   0.2946 0.6641 0.4715 0.0453 0.0063 0.0600 0.8229 0.0123 0.2012</div>
NKX23_HUMAN/147..207		<div>A   0.2060 0.1792 0.0541 0.7949 0.9824 0.0017 0.1054 0.2634 0.0920</div> <div>C   0.2620 0.1057 0.3395 0.0024 0.0098 0.0087 0.0094 0.0028 0.3330</div> <div>G   0.2375 0.0509 0.1348 0.1574 0.0015 0.9296 0.0623 0.7215 0.3739</div> <div>T   0.2946 0.6641 0.4715 0.0453 0.0063 0.0600 0.8229 0.0123 0.2012</div>
NKX24_HUMAN/188..248		<div>A   0.2060 0.1792 0.0541 0.7949 0.9824 0.0017 0.1054 0.2634 0.0920</div> <div>C   0.2620 0.1057 0.3395 0.0024 0.0098 0.0087 0.0094 0.0028 0.3330</div> <div>G   0.2375 0.0509 0.1348 0.1574 0.0015 0.9296 0.0623 0.7215 0.3739</div> <div>T   0.2946 0.6641 0.4715 0.0453 0.0063 0.0600 0.8229 0.0123 0.2012</div>
NKX25_HUMAN/137..197		<div>A   0.2060 0.1792 0.0541 0.7949 0.9824 0.0017 0.1054 0.2634 0.0920</div> <div>C   0.2620 0.1057 0.3395 0.0024 0.0098 0.0087 0.0094 0.0028 0.3330</div> <div>G   0.2375 0.0509 0.1348 0.1574 0.0015 0.9296 0.0623 0.7215 0.3739</div> <div>T   0.2946 0.6641 0.4715 0.0453 0.0063 0.0600 0.8229 0.0123 0.2012</div>
NKX26_HUMAN/131..191		<div>A   0.2060 0.1792 0.0541 0.7949 0.9824 0.0017 0.1054 0.2634 0.0920</div> <div>C   0.2620 0.1057 0.3395 0.0024 0.0098 0.0087 0.0094 0.0028 0.3330</div> <div>G   0.2375 0.0509 0.1348 0.1574 0.0015 0.9296 0.0623 0.7215 0.3739</div> <div>T   0.2946 0.6641 0.4715 0.0453 0.0063 0.0600 0.8229 0.0123 0.2012</div>
NKX28_HUMAN/83..143		<div>A   0.2060 0.1792 0.0541 0.7949 0.9824 0.0017 0.1054 0.2634 0.0920</div> <div>C   0.2620 0.1057 0.3395 0.0024 0.0098 0.0087 0.0094 0.0028 0.3330</div> <div>G   0.2375 0.0509 0.1348 0.1574 0.0015 0.9296 0.0623 0.7215 0.3739</div> <div>T   0.2946 0.6641 0.4715 0.0453 0.0063 0.0600 0.8229 0.0123 0.2012</div>
NKX31_HUMAN/123..183		<div>A   0.2235 0.1651 0.0630 0.9099 0.9811 0.0198 0.0598 0.2550 0.1154</div> <div>C   0.2660 0.2142 0.0465 0.0123 0.0105 0.0000 0.0066 0.0071 0.2901</div> <div>G   0.2598 0.1060 0.0496 0.0451 0.0015 0.8713 0.0516 0.7340 0.4255</div> <div>T   0.2507 0.5147 0.8410 0.0326 0.0070 0.1089 0.8820 0.0040 0.1691</div>
NKX32_HUMAN/205..265		<div>A   0.2131 0.1478 0.0635 0.9117 0.9805 0.0110 0.0618 0.2794 0.1347</div> <div>C   0.2745 0.2056 0.0559 0.0182 0.0119 0.0002 0.0075 0.0010 0.2881</div> <div>G   0.2658 0.1039 0.0534 0.0496 0.0002 0.8587 0.0694 0.7127 0.3851</div> <div>T   0.2467 0.5426 0.8273 0.0205 0.0074 0.1302 0.8613 0.0069 0.1921</div>
NKX61_HUMAN/235..295		<div>A   0.3161 0.2891 0.0259 0.8317 0.9838 0.0151 0.0982 0.6754 0.1698</div> <div>C   0.1581 0.0691 0.0375 0.0011 0.0083 0.0000 0.0511 0.0194 0.3465</div> <div>G   0.2046 0.0789 0.0185 0.0779 0.0022 0.0362 0.3095 0.2825 0.3006</div> <div>T   0.3213 0.5630 0.9181 0.0892 0.0057 0.9487 0.5412 0.0227 0.1831</div>
NKX62_HUMAN/147..207		<div>A   0.3161 0.2891 0.0259 0.8317 0.9838 0.0151 0.0982 0.6754 0.1698</div> <div>C   0.1581 0.0691 0.0375 0.0011 0.0083 0.0000 0.0511 0.0194 0.3465</div> <div>G   0.2046 0.0789 0.0185 0.0779 0.0022 0.0362 0.3095 0.2825 0.3006</div> <div>T   0.3213 0.5630 0.9181 0.0892 0.0057 0.9487 0.5412 0.0227 0.1831</div>

# Appendix A-4 contd.

		A   0.5223 0.5000 0.5204 0.5000 0.5000 0.5000 0.5000 0.5000 0.5000 0.5000 C   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 G   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 T   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
NKX63_HUMAN/138..198		A   0.3161 0.2891 0.0259 0.8317 0.9838 0.0151 0.0982 0.6754 0.1698 C   0.1581 0.0691 0.0375 0.0011 0.0083 0.0000 0.0511 0.0194 0.3465 G   0.2046 0.0789 0.0185 0.0779 0.0022 0.0362 0.3095 0.2825 0.3006 T   0.3213 0.5630 0.9181 0.0892 0.0057 0.9487 0.5412 0.0227 0.1831
NOBOX_HUMAN/271..320	No prediction made	The extracted domain has a gap at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
NOTO_HUMAN/155..215		A   0.2288 0.1814 0.0238 0.9578 0.9874 0.0057 0.0766 0.5059 0.2380 C   0.2586 0.3085 0.0078 0.0099 0.0061 0.1279 0.1358 0.0746 0.1990 G   0.2943 0.1224 0.0041 0.0153 0.0031 0.0411 0.2385 0.3457 0.4227 T   0.2183 0.3876 0.9643 0.0170 0.0034 0.8253 0.5491 0.0738 0.1404
ONEC2_HUMAN/425..484		A   0.1931 0.1113 0.0222 0.2972 0.9745 0.0000 0.2299 0.4554 0.2274 C   0.2696 0.1448 0.0114 0.0000 0.0157 0.3179 0.0929 0.1078 0.2631 G   0.2480 0.1308 0.0526 0.6561 0.0000 0.0649 0.0780 0.3108 0.2899 T   0.2893 0.6130 0.9138 0.0467 0.0099 0.6172 0.5991 0.1260 0.2195
ONEC3_HUMAN/413..472		A   0.1931 0.1113 0.0222 0.2972 0.9745 0.0000 0.2299 0.4554 0.2274 C   0.2696 0.1448 0.0114 0.0000 0.0157 0.3179 0.0929 0.1078 0.2631 G   0.2480 0.1308 0.0526 0.6561 0.0000 0.0649 0.0780 0.3108 0.2899 T   0.2893 0.6130 0.9138 0.0467 0.0099 0.6172 0.5991 0.1260 0.2195
OTP_HUMAN/103..163		A   0.2452 0.1830 0.1529 0.7102 0.9874 0.0080 0.0318 0.6574 0.2954 C   0.2423 0.2509 0.0212 0.0057 0.0063 0.0055 0.0256 0.0096 0.2252 G   0.2592 0.1444 0.0178 0.0416 0.0031 0.0211 0.1130 0.2999 0.3132 T   0.2533 0.4217 0.8082 0.2424 0.0032 0.9655 0.8296 0.0330 0.1661
OTX1_HUMAN/37..97		A   0.2271 0.2000 0.0400 0.9635 0.9883 0.0002 0.0349 0.0327 0.1395 C   0.2471 0.2969 0.0041 0.0098 0.0085 0.0051 0.9108 0.7577 0.3630 G   0.3047 0.1251 0.0091 0.0116 0.0013 0.0741 0.0043 0.0403 0.2830 T   0.2211 0.3780 0.9468 0.0151 0.0019 0.9206 0.0500 0.1693 0.2146
OTX2_HUMAN/37..97		A   0.2271 0.2000 0.0400 0.9635 0.9883 0.0002 0.0349 0.0327 0.1395 C   0.2471 0.2969 0.0041 0.0098 0.0085 0.0051 0.9108 0.7577 0.3630 G   0.3047 0.1251 0.0091 0.0116 0.0013 0.0741 0.0043 0.0403 0.2830 T   0.2211 0.3780 0.9468 0.0151 0.0019 0.9206 0.0500 0.1693 0.2146
P5F1B_HUMAN/229..288		A   0.3019 0.2505 0.0874 0.8705 0.9843 0.0063 0.0581 0.5512 0.3071 C   0.2338 0.2215 0.0037 0.0060 0.0059 0.0072 0.0558 0.2614 0.1645 G   0.2729 0.0959 0.0303 0.0190 0.0029 0.0265 0.5642 0.0837 0.4177 T   0.1914 0.4321 0.8786 0.1044 0.0068 0.9600 0.3218 0.1038 0.1108
PAX3_HUMAN/218..278		A   0.2275 0.2398 0.0596 0.7268 0.9881 0.0157 0.1254 0.2813 0.2523 C   0.2571 0.2797 0.0096 0.0030 0.0072 0.0173 0.2354 0.2483 0.1896 G   0.3102 0.1459 0.0148 0.2523 0.0021 0.0099 0.1297 0.3206 0.4273 T   0.2052 0.3345 0.9160 0.0179 0.0026 0.9571 0.5094 0.1498 0.1307
PAX4_HUMAN/169..229		A   0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535 C   0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239 G   0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676 T   0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550
		A   0.2558 0.1352 0.0223 0.9726 0.9900 0.0053 0.3103 0.3134 0.1535

# Appendix A-4 contd.

PAX6_HUMAN/209..269		C   0.2783 0.3110 0.0044 0.0101 0.0060 0.0141 0.0213 0.4853 0.0239 G   0.2932 0.0638 0.0021 0.0148 0.0033 0.0037 0.3739 0.0424 0.7676 T   0.1728 0.4900 0.9712 0.0026 0.0008 0.9769 0.2945 0.1589 0.0550
PAX7_HUMAN/216..276		A   0.2275 0.2398 0.0596 0.7268 0.9881 0.0157 0.1254 0.2813 0.2523 C   0.2571 0.2797 0.0096 0.0030 0.0072 0.0173 0.2354 0.2483 0.1896 G   0.3102 0.1459 0.0148 0.2523 0.0021 0.0099 0.1297 0.3206 0.4273 T   0.2052 0.3345 0.9160 0.0179 0.0026 0.9571 0.5094 0.1498 0.1307
PBX1_HUMAN/232..295		A   0.2331 0.1862 0.0329 0.1150 0.9785 0.0256 0.3941 0.2947 0.2436 C   0.2117 0.1101 0.0141 0.0001 0.0070 0.4410 0.0780 0.1906 0.2202 G   0.1937 0.1722 0.0228 0.8471 0.0034 0.0585 0.3485 0.2694 0.3604 T   0.3615 0.5314 0.9302 0.0378 0.0111 0.4749 0.1794 0.2454 0.1758
PBX2_HUMAN/243..306		A   0.2331 0.1862 0.0329 0.1150 0.9785 0.0256 0.3941 0.2947 0.2436 C   0.2117 0.1101 0.0141 0.0001 0.0070 0.4410 0.0780 0.1906 0.2202 G   0.1937 0.1722 0.0228 0.8471 0.0034 0.0585 0.3485 0.2694 0.3604 T   0.3615 0.5314 0.9302 0.0378 0.0111 0.4749 0.1794 0.2454 0.1758
PBX3_HUMAN/234..297		A   0.2331 0.1862 0.0329 0.1150 0.9785 0.0256 0.3941 0.2947 0.2436 C   0.2117 0.1101 0.0141 0.0001 0.0070 0.4410 0.0780 0.1906 0.2202 G   0.1937 0.1722 0.0228 0.8471 0.0034 0.0585 0.3485 0.2694 0.3604 T   0.3615 0.5314 0.9302 0.0378 0.0111 0.4749 0.1794 0.2454 0.1758
PBX4_HUMAN/209..272		A   0.2331 0.1862 0.0329 0.1150 0.9785 0.0256 0.3941 0.2947 0.2436 C   0.2117 0.1101 0.0141 0.0001 0.0070 0.4410 0.0780 0.1906 0.2202 G   0.1937 0.1722 0.0228 0.8471 0.0034 0.0585 0.3485 0.2694 0.3604 T   0.3615 0.5314 0.9302 0.0378 0.0111 0.4749 0.1794 0.2454 0.1758
PDX1_HUMAN/145..205		A   0.2263 0.1402 0.0485 0.9456 0.9829 0.0076 0.0573 0.5889 0.1826 C   0.2661 0.3168 0.0059 0.0128 0.0073 0.0304 0.0505 0.0231 0.3520 G   0.2886 0.1183 0.0086 0.0231 0.0033 0.0147 0.4736 0.3364 0.2637 T   0.2191 0.4247 0.9370 0.0185 0.0065 0.9473 0.4186 0.0516 0.2017
PHX2A_HUMAN/89..149		A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810 C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916 G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810 T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464
PHX2B_HUMAN/97..157		A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810 C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916 G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810 T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464
PIT1_HUMAN/213..273		A   0.2631 0.3989 0.1168 0.6712 0.9791 0.0144 0.0535 0.5957 0.4059 C   0.2246 0.1544 0.0168 0.0445 0.0044 0.0063 0.0536 0.2864 0.1167 G   0.2723 0.1050 0.0283 0.1157 0.0047 0.0021 0.4663 0.0446 0.3879 T   0.2401 0.3418 0.8381 0.1687 0.0118 0.9772 0.4266 0.0733 0.0895
PITX1_HUMAN/88..148		A   0.2324 0.1630 0.0267 0.9782 0.9888 0.0026 0.0191 0.0222 0.1627 C   0.2392 0.2854 0.0077 0.0036 0.0066 0.0029 0.9491 0.8396 0.3991 G   0.3303 0.0989 0.0011 0.0110 0.0028 0.0434 0.0012 0.0260 0.2300 T   0.1981 0.4527 0.9645 0.0072 0.0018 0.9511 0.0306 0.1121 0.2083
PITX2_HUMAN/84..144		A   0.2324 0.1630 0.0267 0.9782 0.9888 0.0026 0.0191 0.0222 0.1627 C   0.2392 0.2854 0.0077 0.0036 0.0066 0.0029 0.9491 0.8396 0.3991 G   0.3303 0.0989 0.0011 0.0110 0.0028 0.0434 0.0012 0.0260 0.2300 T   0.1981 0.4527 0.9645 0.0072 0.0018 0.9511 0.0306 0.1121 0.2083

# Appendix A-4 contd.

		T   0.1981 0.4527 0.9645 0.0072 0.0018 0.9511 0.0306 0.1121 0.2083
PITX3_HUMAN/61..121		A   0.2324 0.1630 0.0267 0.9782 0.9888 0.0026 0.0191 0.0222 0.1627 C   0.2392 0.2854 0.0077 0.0036 0.0066 0.0029 0.9491 0.8396 0.3991 G   0.3303 0.0989 0.0011 0.0110 0.0028 0.0434 0.0012 0.0260 0.2300 T   0.1981 0.4527 0.9645 0.0072 0.0018 0.9511 0.0306 0.1121 0.2083
PO2F1_HUMAN/378..438		A   0.2670 0.3899 0.1107 0.7971 0.9786 0.0187 0.0316 0.6638 0.3870 C   0.2142 0.1971 0.0113 0.0073 0.0029 0.0003 0.0657 0.2101 0.0997 G   0.3053 0.1073 0.0149 0.0115 0.0060 0.0103 0.4403 0.0463 0.4244 T   0.2135 0.3057 0.8630 0.1841 0.0126 0.9707 0.4624 0.0798 0.0889
PO2F2_HUMAN/296..356		A   0.2670 0.3899 0.1107 0.7971 0.9786 0.0187 0.0316 0.6638 0.3870 C   0.2142 0.1971 0.0113 0.0073 0.0029 0.0003 0.0657 0.2101 0.0997 G   0.3053 0.1073 0.0149 0.0115 0.0060 0.0103 0.4403 0.0463 0.4244 T   0.2135 0.3057 0.8630 0.1841 0.0126 0.9707 0.4624 0.0798 0.0889
PO2F3_HUMAN/280..340		A   0.2473 0.3513 0.1233 0.7905 0.9802 0.0121 0.0171 0.6505 0.3667 C   0.2127 0.2208 0.0182 0.0067 0.0049 0.0036 0.0486 0.2373 0.0974 G   0.3243 0.1216 0.0152 0.0176 0.0043 0.0067 0.4954 0.0479 0.4528 T   0.2158 0.3063 0.8433 0.1853 0.0105 0.9775 0.4389 0.0643 0.0831
PO3F1_HUMAN/338..398		A   0.2612 0.3687 0.1334 0.7055 0.9766 0.0135 0.0219 0.5028 0.4292 C   0.2211 0.1796 0.0305 0.0041 0.0031 0.0082 0.0480 0.3732 0.1101 G   0.2883 0.1232 0.0400 0.0230 0.0061 0.0037 0.4951 0.0383 0.3881 T   0.2294 0.3284 0.7962 0.2673 0.0142 0.9745 0.4351 0.0857 0.0726
PO3F2_HUMAN/353..413		A   0.2612 0.3687 0.1334 0.7055 0.9766 0.0135 0.0219 0.5028 0.4292 C   0.2211 0.1796 0.0305 0.0041 0.0031 0.0082 0.0480 0.3732 0.1101 G   0.2883 0.1232 0.0400 0.0230 0.0061 0.0037 0.4951 0.0383 0.3881 T   0.2294 0.3284 0.7962 0.2673 0.0142 0.9745 0.4351 0.0857 0.0726
PO3F3_HUMAN/405..465		A   0.2612 0.3687 0.1334 0.7055 0.9766 0.0135 0.0219 0.5028 0.4292 C   0.2211 0.1796 0.0305 0.0041 0.0031 0.0082 0.0480 0.3732 0.1101 G   0.2883 0.1232 0.0400 0.0230 0.0061 0.0037 0.4951 0.0383 0.3881 T   0.2294 0.3284 0.7962 0.2673 0.0142 0.9745 0.4351 0.0857 0.0726
PO3F4_HUMAN/277..337		A   0.2612 0.3687 0.1334 0.7055 0.9766 0.0135 0.0219 0.5028 0.4292 C   0.2211 0.1796 0.0305 0.0041 0.0031 0.0082 0.0480 0.3732 0.1101 G   0.2883 0.1232 0.0400 0.0230 0.0061 0.0037 0.4951 0.0383 0.3881 T   0.2294 0.3284 0.7962 0.2673 0.0142 0.9745 0.4351 0.0857 0.0726
PO4F1_HUMAN/354..414		A   0.3438 0.2235 0.0819 0.8454 0.9818 0.0101 0.0620 0.6215 0.2867 C   0.2166 0.1917 0.0074 0.0091 0.0053 0.0150 0.0321 0.2077 0.1403 G   0.2470 0.0853 0.0348 0.0344 0.0036 0.0097 0.6031 0.0630 0.4496 T   0.1926 0.4995 0.8759 0.1111 0.0093 0.9652 0.3028 0.1077 0.1233
PO4F2_HUMAN/344..404		A   0.3438 0.2235 0.0819 0.8454 0.9818 0.0101 0.0620 0.6215 0.2867 C   0.2166 0.1917 0.0074 0.0091 0.0053 0.0150 0.0321 0.2077 0.1403 G   0.2470 0.0853 0.0348 0.0344 0.0036 0.0097 0.6031 0.0630 0.4496 T   0.1926 0.4995 0.8759 0.1111 0.0093 0.9652 0.3028 0.1077 0.1233
PO4F3_HUMAN/273..333		A   0.3438 0.2235 0.0819 0.8454 0.9818 0.0101 0.0620 0.6215 0.2867 C   0.2166 0.1917 0.0074 0.0091 0.0053 0.0150 0.0321 0.2077 0.1403 G   0.2470 0.0853 0.0348 0.0344 0.0036 0.0097 0.6031 0.0630 0.4496 T   0.1926 0.4995 0.8759 0.1111 0.0093 0.9652 0.3028 0.1077 0.1233

# Appendix A-4 contd.



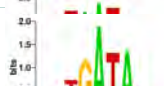






PO5F1_HUMAN/229..289		<div>A   0.3019 0.2505 0.0874 0.8705 0.9843 0.0063 0.0581 0.5512 0.3071</div> <div>C   0.2338 0.2215 0.0037 0.0060 0.0059 0.0072 0.0558 0.2614 0.1645</div> <div>G   0.2729 0.0959 0.0303 0.0190 0.0029 0.0265 0.5642 0.0837 0.4177</div> <div>T   0.1914 0.4321 0.8786 0.1044 0.0068 0.9600 0.3218 0.1038 0.1108</div>
PO5F2_HUMAN/209..266		<div>A   0.2529 0.2213 0.1093 0.5107 0.9663 0.1339 0.0631 0.5415 0.2454</div> <div>C   0.2452 0.1961 0.0639 0.0569 0.0059 0.4328 0.3338 0.1790 0.2473</div> <div>G   0.2668 0.1442 0.0457 0.1208 0.0072 0.0000 0.1223 0.1397 0.3659</div> <div>T   0.2351 0.4384 0.7811 0.3115 0.0206 0.4332 0.4808 0.1398 0.1415</div>
PO6F1_HUMAN/233..293		<div>A   0.2774 0.4172 0.0795 0.7917 0.9789 0.0112 0.0421 0.8048 0.2034</div> <div>C   0.2227 0.1370 0.0124 0.0289 0.0057 0.0117 0.0326 0.1053 0.0576</div> <div>G   0.2710 0.1098 0.0138 0.0729 0.0034 0.0019 0.5778 0.0400 0.6060</div> <div>T   0.2289 0.3360 0.8943 0.1065 0.0119 0.9752 0.3475 0.0498 0.1330</div>
PO6F2_HUMAN/606..666		<div>A   0.2607 0.3976 0.1200 0.7161 0.9777 0.0136 0.0254 0.5760 0.4088</div> <div>C   0.2257 0.1576 0.0225 0.0307 0.0040 0.0081 0.0466 0.3199 0.1088</div> <div>G   0.2789 0.1134 0.0320 0.0503 0.0052 0.0021 0.4992 0.0377 0.4068</div> <div>T   0.2347 0.3314 0.8256 0.2029 0.0131 0.9763 0.4288 0.0664 0.0756</div>
PROP1_HUMAN/68..128		<div>A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849</div> <div>C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028</div> <div>G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384</div> <div>T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739</div>
PRRX1_HUMAN/93..153		<div>A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810</div> <div>C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916</div> <div>G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810</div> <div>T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464</div>
PRRX2_HUMAN/103..163		<div>A   0.2179 0.2059 0.0287 0.9560 0.9867 0.0062 0.0454 0.4502 0.2810</div> <div>C   0.2574 0.2964 0.0141 0.0108 0.0059 0.0139 0.0455 0.0537 0.1916</div> <div>G   0.3067 0.1247 0.0046 0.0216 0.0034 0.0067 0.0717 0.4088 0.3810</div> <div>T   0.2180 0.3731 0.9526 0.0116 0.0039 0.9732 0.8375 0.0873 0.1464</div>
Q2KJ05_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
Q32M63_HUMAN/78..128	No prediction made	The extracted domain has residue (Q) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
Q3ZB87_HUMAN/495..556		<div>A   0.2093 0.1653 0.0865 0.6282 0.9833 0.0972 0.0798 0.2772 0.1611</div> <div>C   0.2440 0.1567 0.1058 0.0006 0.0076 0.0000 0.0622 0.0717 0.2572</div> <div>G   0.2283 0.1041 0.0790 0.2564 0.0028 0.7868 0.1105 0.5713 0.3955</div> <div>T   0.3184 0.5738 0.7287 0.1148 0.0063 0.1160 0.7475 0.0798 0.1862</div>
Q494Z3_HUMAN/218..278		<div>A   0.2275 0.2398 0.0596 0.7268 0.9881 0.0157 0.1254 0.2813 0.2523</div> <div>C   0.2571 0.2797 0.0096 0.0030 0.0072 0.0173 0.2354 0.2483 0.1896</div> <div>G   0.3102 0.1459 0.0148 0.2523 0.0021 0.0099 0.1297 0.3206 0.4273</div> <div>T   0.2052 0.3345 0.9160 0.0179 0.0026 0.9571 0.5094 0.1498 0.1307</div>
Q494Z4_HUMAN/217..277		<div>A   0.2275 0.2398 0.0596 0.7268 0.9881 0.0157 0.1254 0.2813 0.2523</div> <div>C   0.2571 0.2797 0.0096 0.0030 0.0072 0.0173 0.2354 0.2483 0.1896</div>

# Appendix A-4 contd.

Q494Z4_HUMAN/21...211		G   0.3102 0.1459 0.0148 0.2523 0.0021 0.0099 0.1297 0.3206 0.4273 T   0.2052 0.3345 0.9160 0.0179 0.0026 0.9571 0.5094 0.1498 0.1307
Q49AQ3_HUMAN/81..137	No prediction made	The extracted domain has residue (H) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
Q53Y73_HUMAN/136..196		A   0.1987 0.3362 0.0584 0.9561 0.9824 0.0100 0.0806 0.3828 0.1706 C   0.2379 0.2062 0.0127 0.0160 0.0105 0.0092 0.0670 0.0273 0.3544 G   0.3166 0.1786 0.0059 0.0239 0.0009 0.0302 0.1323 0.5315 0.2907 T   0.2468 0.2791 0.9230 0.0040 0.0062 0.9506 0.7201 0.0585 0.1843
Q5SZE1_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
Q5SZE2_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
Q5SZE3_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
Q5SZE4_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
Q5VZ84_HUMAN	No prediction made	No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
Q8IXZ1_HUMAN/195..255		A   0.2198 0.1397 0.0578 0.7688 0.9711 0.0086 0.0609 0.5167 0.1622 C   0.2431 0.1534 0.0278 0.0237 0.0056 0.0621 0.0507 0.0232 0.4097 G   0.2330 0.1142 0.0312 0.0137 0.0071 0.0251 0.4372 0.3948 0.2140 T   0.3041 0.5926 0.8831 0.1937 0.0162 0.9043 0.4512 0.0653 0.2141
RAX2_HUMAN/26..86		A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849 C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028 G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384 T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739
RHF2B_HUMAN/135..193		A   0.2316 0.1860 0.0833 0.6866 0.9789 0.0000 0.0120 0.2694 0.1960 C   0.2442 0.1896 0.0416 0.0000 0.0082 0.0511 0.2420 0.1181 0.3010 G   0.2408 0.1600 0.0300 0.1421 0.0032 0.1144 0.2042 0.5649 0.3107 T   0.2834 0.4644 0.8451 0.1713 0.0097 0.8345 0.5417 0.0476 0.1923
RHXF1_HUMAN/102..162		A   0.2420 0.2234 0.0631 0.5942 0.9852 0.0117 0.1243 0.0418 0.1972 C   0.2423 0.2577 0.0221 0.0000 0.0058 0.0111 0.7415 0.7322 0.3296 G   0.2817 0.1467 0.0226 0.3469 0.0035 0.0387 0.0000 0.0673 0.2854 T   0.2340 0.3722 0.8921 0.0589 0.0056 0.9385 0.1342 0.1587 0.1878
RHXF2_HUMAN/135..193		A   0.2316 0.1860 0.0833 0.6866 0.9789 0.0000 0.0120 0.2694 0.1960 C   0.2442 0.1896 0.0416 0.0000 0.0082 0.0511 0.2420 0.1181 0.3010 G   0.2408 0.1600 0.0300 0.1421 0.0032 0.1144 0.2042 0.5649 0.3107 T   0.2834 0.4644 0.8451 0.1713 0.0097 0.8345 0.5417 0.0476 0.1923
RX_HUMAN/135..195		A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849 C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028 G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384 T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739
		



# Appendix A-4 contd.


SATB1_HUMAN/643..704		<div>A   0.1989 0.1422 0.1056 0.7334 0.9902 0.0585 0.0615 0.3652 0.1648</div> <div>C   0.2936 0.3110 0.0467 0.0000 0.0068 0.0000 0.0436 0.0263 0.2376</div> <div>G   0.2584 0.1334 0.0676 0.2474 0.0021 0.8517 0.1338 0.6082 0.4324</div> <div>T   0.2491 0.4133 0.7801 0.0192 0.0009 0.0898 0.7610 0.0003 0.1651</div>
SATB2_HUMAN/613..674		<div>A   0.2093 0.1653 0.0865 0.6282 0.9833 0.0972 0.0798 0.2772 0.1611</div> <div>C   0.2440 0.1567 0.1058 0.0006 0.0076 0.0000 0.0622 0.0717 0.2572</div> <div>G   0.2283 0.1041 0.0790 0.2564 0.0028 0.7868 0.1105 0.5713 0.3955</div> <div>T   0.3184 0.5738 0.7287 0.1148 0.0063 0.1160 0.7475 0.0798 0.1862</div>
SEBOX_HUMAN/44..104	No prediction made	The extracted domain has residue (K) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
SHOX2_HUMAN/139..199		<div>A   0.2216 0.1799 0.0271 0.9614 0.9876 0.0035 0.0280 0.4998 0.2551</div> <div>C   0.2672 0.2974 0.0107 0.0123 0.0082 0.0037 0.0568 0.0421 0.2260</div> <div>G   0.2853 0.1350 0.0018 0.0161 0.0018 0.0203 0.0898 0.3923 0.3507</div> <div>T   0.2259 0.3877 0.9605 0.0102 0.0024 0.9725 0.8254 0.0657 0.1682</div>
SHOX_HUMAN/116..176		<div>A   0.2216 0.1799 0.0271 0.9614 0.9876 0.0035 0.0280 0.4998 0.2551</div> <div>C   0.2672 0.2974 0.0107 0.0123 0.0082 0.0037 0.0568 0.0421 0.2260</div> <div>G   0.2853 0.1350 0.0018 0.0161 0.0018 0.0203 0.0898 0.3923 0.3507</div> <div>T   0.2259 0.3877 0.9605 0.0102 0.0024 0.9725 0.8254 0.0657 0.1682</div>
SIX1_HUMAN/125..183		<div>A   0.2640 0.3129 0.0587 0.0593 0.9884 0.0137 0.8679 0.0974 0.1561</div> <div>C   0.2307 0.1840 0.1423 0.0176 0.0077 0.0693 0.1216 0.6202 0.4404</div> <div>G   0.2413 0.2526 0.0758 0.8463 0.0014 0.0340 0.0082 0.0985 0.2271</div> <div>T   0.2640 0.2505 0.7232 0.0767 0.0025 0.8831 0.0023 0.1840 0.1765</div>
SIX2_HUMAN/125..183		<div>A   0.2640 0.3129 0.0587 0.0593 0.9884 0.0137 0.8679 0.0974 0.1561</div> <div>C   0.2307 0.1840 0.1423 0.0176 0.0077 0.0693 0.1216 0.6202 0.4404</div> <div>G   0.2413 0.2526 0.0758 0.8463 0.0014 0.0340 0.0082 0.0985 0.2271</div> <div>T   0.2640 0.2505 0.7232 0.0767 0.0025 0.8831 0.0023 0.1840 0.1765</div>
SIX3_HUMAN/205..265		<div>A   0.2759 0.2541 0.0613 0.0757 0.9790 0.0151 0.7994 0.1441 0.1897</div> <div>C   0.2275 0.2023 0.1535 0.0283 0.0089 0.0645 0.1722 0.4797 0.3407</div> <div>G   0.2361 0.2890 0.0602 0.8106 0.0017 0.0170 0.0247 0.1612 0.2517</div> <div>T   0.2605 0.2546 0.7250 0.0855 0.0105 0.9035 0.0037 0.2150 0.2178</div>
SIX4_HUMAN/225..282		<div>A   0.2537 0.3059 0.0696 0.0818 0.9881 0.0268 0.7430 0.0787 0.1403</div> <div>C   0.2413 0.1870 0.1236 0.0211 0.0087 0.1361 0.2206 0.7112 0.4748</div> <div>G   0.2467 0.2303 0.0879 0.8057 0.0006 0.0668 0.0034 0.0544 0.2127</div> <div>T   0.2583 0.2768 0.7190 0.0915 0.0025 0.7703 0.0331 0.1557 0.1722</div>
SIX5_HUMAN/204..260		<div>A   0.2548 0.2795 0.0554 0.1369 0.9843 0.0119 0.7549 0.1121 0.1603</div> <div>C   0.2422 0.1899 0.1010 0.0000 0.0091 0.1279 0.1981 0.6441 0.4112</div> <div>G   0.2470 0.2323 0.0460 0.7659 0.0010 0.1392 0.0210 0.0714 0.2441</div> <div>T   0.2560 0.2982 0.7976 0.0971 0.0055 0.7209 0.0260 0.1724 0.1844</div>
SIX6_HUMAN/127..187		<div>A   0.2759 0.2541 0.0613 0.0757 0.9790 0.0151 0.7994 0.1441 0.1897</div> <div>C   0.2275 0.2023 0.1535 0.0283 0.0089 0.0645 0.1722 0.4797 0.3407</div> <div>G   0.2361 0.2890 0.0602 0.8106 0.0017 0.0170 0.0247 0.1612 0.2517</div> <div>T   0.2605 0.2546 0.7250 0.0855 0.0105 0.9035 0.0037 0.2150 0.2178</div>
TLX1_HUMAN/200..260		<div>A   0.2008 0.1962 0.0212 0.9709 0.9865 0.1785 0.1144 0.5041 0.2028</div> <div>C   0.2465 0.2269 0.0088 0.0065 0.0084 0.0788 0.1435 0.0106 0.2671</div> <div>G   0.2883 0.1034 0.0043 0.0053 0.0019 0.0565 0.2367 0.4559 0.3203</div> <div>T   0.2644 0.4735 0.9656 0.0173 0.0032 0.6862 0.5054 0.0294 0.2098</div>



# Appendix A-4 contd.

TLX2_HUMAN/156..216		A   0.2008 0.1962 0.0212 0.9709 0.9865 0.1785 0.1144 0.5041 0.2028 C   0.2465 0.2269 0.0088 0.0065 0.0084 0.0788 0.1435 0.0106 0.2671 G   0.2883 0.1034 0.0043 0.0053 0.0019 0.0565 0.2367 0.4559 0.3203 T   0.2644 0.4735 0.9656 0.0173 0.0032 0.6862 0.5054 0.0294 0.2098
TLX3_HUMAN/165..225		A   0.2008 0.1962 0.0212 0.9709 0.9865 0.1785 0.1144 0.5041 0.2028 C   0.2465 0.2269 0.0088 0.0065 0.0084 0.0788 0.1435 0.0106 0.2671 G   0.2883 0.1034 0.0043 0.0053 0.0019 0.0565 0.2367 0.4559 0.3203 T   0.2644 0.4735 0.9656 0.0173 0.0032 0.6862 0.5054 0.0294 0.2098
UNC4_HUMAN/104..164		A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849 C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028 G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384 T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739
VAX1_HUMAN/99..159		A   0.2242 0.1402 0.0322 0.9710 0.9886 0.0282 0.1057 0.6090 0.2204 C   0.2657 0.3534 0.0062 0.0095 0.0066 0.0122 0.0543 0.0226 0.2705 G   0.3002 0.1290 0.0022 0.0144 0.0029 0.0205 0.3149 0.3164 0.3518 T   0.2099 0.3774 0.9593 0.0051 0.0019 0.9391 0.5251 0.0520 0.1574
VAX2_HUMAN/101..161		A   0.2242 0.1402 0.0322 0.9710 0.9886 0.0282 0.1057 0.6090 0.2204 C   0.2657 0.3534 0.0062 0.0095 0.0066 0.0122 0.0543 0.0226 0.2705 G   0.3002 0.1290 0.0022 0.0144 0.0029 0.0205 0.3149 0.3164 0.3518 T   0.2099 0.3774 0.9593 0.0051 0.0019 0.9391 0.5251 0.0520 0.1574
VENTX_HUMAN/90..150		A   0.2213 0.2287 0.0296 0.9730 0.9858 0.0253 0.0249 0.2617 0.2181 C   0.2571 0.2676 0.0045 0.0053 0.0066 0.0501 0.1803 0.0300 0.2467 G   0.3055 0.1429 0.0033 0.0143 0.0032 0.0219 0.3492 0.6214 0.3657 T   0.2161 0.3609 0.9626 0.0074 0.0044 0.9028 0.4457 0.0869 0.1695
VSX1_HUMAN/163..223		A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849 C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028 G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384 T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739
VSX2_HUMAN/147..207		A   0.2377 0.1572 0.0336 0.9308 0.9875 0.0039 0.0375 0.5948 0.2849 C   0.2492 0.3294 0.0224 0.0159 0.0079 0.0042 0.0238 0.0207 0.2028 G   0.2792 0.1170 0.0158 0.0405 0.0020 0.0218 0.0738 0.3329 0.3384 T   0.2340 0.3964 0.9282 0.0127 0.0027 0.9701 0.8649 0.0516 0.1739
ZEB1_HUMAN		No matches were found to the Homeobox.hmm Pfam model using the program hmmsearch and a domE cut off of 1e-07
ZFHX2_HUMAN/1594..1654		A   0.2497 0.2247 0.0566 0.9201 0.9860 0.0070 0.0308 0.6598 0.2676 C   0.2494 0.2419 0.0251 0.0094 0.0071 0.0112 0.0570 0.0146 0.2110 G   0.2706 0.1490 0.0000 0.0235 0.0022 0.0288 0.3847 0.3034 0.3756 T   0.2302 0.3844 0.9183 0.0469 0.0047 0.9530 0.5275 0.0222 0.1458
ZFHX2_HUMAN/1856..1916		A   0.2497 0.2148 0.0239 0.4795 0.9877 0.0573 0.2147 0.5618 0.2241 C   0.2443 0.2838 0.0320 0.0000 0.0073 0.0680 0.0101 0.0332 0.2709 G   0.2773 0.1781 0.0219 0.4837 0.0021 0.0000 0.0907 0.3776 0.3267 T   0.2287 0.3232 0.9222 0.0368 0.0029 0.8747 0.6845 0.0274 0.1783
ZFHX2_HUMAN/2064..2124		A   0.2399 0.1483 0.0321 0.9435 0.9809 0.0210 0.0287 0.5619 0.2547 C   0.2461 0.3280 0.0106 0.0110 0.0072 0.0028 0.0488 0.0140 0.2428 G   0.2837 0.1082 0.0044 0.0325 0.0037 0.0484 0.1220 0.3763 0.3085

# Appendix A-4 contd.

		T   0.2304 0.4155 0.9529 0.0129 0.0083 0.9279 0.8005 0.0478 0.1940
ZFHX3_HUMAN/2144..2204		A   0.2351 0.1770 0.0755 0.8644 0.9854 0.1264 0.6189 0.1422 0.1215 C   0.2654 0.2520 0.0169 0.0187 0.0060 0.0136 0.1701 0.6182 0.2285 G   0.2370 0.1027 0.0337 0.0968 0.0036 0.7031 0.0000 0.0275 0.5260 T   0.2625 0.4683 0.8739 0.0200 0.0050 0.1570 0.2109 0.2121 0.1240
ZFHX3_HUMAN/2241..2301		A   0.2502 0.2256 0.0619 0.9035 0.9831 0.0071 0.0308 0.6604 0.2658 C   0.2486 0.2402 0.0224 0.0135 0.0064 0.0113 0.0569 0.0140 0.2083 G   0.2701 0.1463 0.0053 0.0199 0.0032 0.0288 0.3847 0.3024 0.3783 T   0.2311 0.3879 0.9104 0.0631 0.0073 0.9529 0.5276 0.0232 0.1477
ZFHX3_HUMAN/2640..2700		A   0.2342 0.1668 0.0273 0.5568 0.9874 0.0101 0.0138 0.4708 0.2177 C   0.2530 0.3389 0.0029 0.0000 0.0079 0.0084 0.0216 0.0118 0.2707 G   0.2909 0.1231 0.0049 0.4184 0.0017 0.0107 0.0940 0.4760 0.3289 T   0.2220 0.3711 0.9649 0.0247 0.0030 0.9708 0.8706 0.0414 0.1826
ZFHX3_HUMAN/2945..3005		A   0.2399 0.1483 0.0321 0.9435 0.9809 0.0210 0.0287 0.5619 0.2547 C   0.2461 0.3280 0.0106 0.0110 0.0072 0.0028 0.0488 0.0140 0.2428 G   0.2837 0.1082 0.0044 0.0325 0.0037 0.0484 0.1220 0.3763 0.3085 T   0.2304 0.4155 0.9529 0.0129 0.0083 0.9279 0.8005 0.0478 0.1940
ZFHX4_HUMAN/2083..2143		A   0.2351 0.1770 0.0755 0.8644 0.9854 0.1264 0.6189 0.1422 0.1215 C   0.2654 0.2520 0.0169 0.0187 0.0060 0.0136 0.1701 0.6182 0.2285 G   0.2370 0.1027 0.0337 0.0968 0.0036 0.7031 0.0000 0.0275 0.5260 T   0.2625 0.4683 0.8739 0.0200 0.0050 0.1570 0.2109 0.2121 0.1240
ZFHX4_HUMAN/2180..2240		A   0.2519 0.1819 0.0275 0.8227 0.9853 0.0203 0.0536 0.6569 0.2579 C   0.2418 0.2492 0.0408 0.0045 0.0075 0.0334 0.0530 0.0141 0.2140 G   0.2715 0.1199 0.0276 0.1135 0.0019 0.0237 0.3889 0.2931 0.3775 T   0.2348 0.4490 0.9041 0.0594 0.0052 0.9226 0.5044 0.0359 0.1507
ZFHX4_HUMAN/2559..2619		A   0.2342 0.1668 0.0273 0.5568 0.9874 0.0101 0.0138 0.4708 0.2177 C   0.2530 0.3389 0.0029 0.0000 0.0079 0.0084 0.0216 0.0118 0.2707 G   0.2909 0.1231 0.0049 0.4184 0.0017 0.0107 0.0940 0.4760 0.3289 T   0.2220 0.3711 0.9649 0.0247 0.0030 0.9708 0.8706 0.0414 0.1826
ZFHX4_HUMAN/2883..2943		A   0.2399 0.1483 0.0321 0.9435 0.9809 0.0210 0.0287 0.5619 0.2547 C   0.2461 0.3280 0.0106 0.0110 0.0072 0.0028 0.0488 0.0140 0.2428 G   0.2837 0.1082 0.0044 0.0325 0.0037 0.0484 0.1220 0.3763 0.3085 T   0.2304 0.4155 0.9529 0.0129 0.0083 0.9279 0.8005 0.0478 0.1940
ZHX1_HUMAN/465..523	No prediction made	The extracted domain has residue (D) at position 51 but residue (N) is required at this poosition in order to make a prediction (numbering is relative to the reference sequence, en_fly)
ZHX1_HUMAN/663..719	No prediction made	The extracted domain has residue (D) at position 51 but residue (N) is required at this poosition in order to make a prediction (numbering is relative to the reference sequence, en_fly)
ZHX2_HUMAN/440..498	No prediction made	The extracted domain has residue (D) at position 51 but residue (N) is required at this poosition in order to make a prediction (numbering is relative to the reference sequence, en_fly)
ZHX2_HUMAN/532..586	No prediction made	The extracted domain has residue (E) at position 51 but residue (N) is required at this poosition in order to make a prediction (numbering is relative to the reference sequence, en_fly)
ZHX2_HUMAN/405..452	No prediction made	The extracted domain has residue (D) at position 51 but residue (N) is required at this poosition in order to make a prediction (numbering is relative to the reference sequence, en_fly)

Appendix A-4 contd.

ZHX3_HUMAN/473..553	No prediction made	is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)
ZHX3_HUMAN/612..669	No prediction made	The extracted domain has residue (E) at position 51 but residue (N) is required at this position in order to make a prediction (numbering is relative to the reference sequence, en_fly)

Download all predicted PFMs in a single [file](#).  
Download the [FASTA file](#) used to make these predictions

## Appendix A-5:

TAA NNN site	HDs w/ posit ive charge both pos 43 & 46	HDs w/ posit ive charge only pos 43	HDs w/ posit ive charge only pos 46	HDs w/ posit ive charge neith er pos 43 & 46	total numbe r of HDs	expected frequency of cooccurren ce assuming independen ce	actual frequency cooccurren ce	differen ce	odds.ratio	confidence interval 95percent low end	confidenc e interval 95percent high end	B-H adjusted p- value
TTT	31	416	202	57	706	0.209	0.044	0.165	2.12E-02	1.28E-02	3.44E-02	4.96E-86
TTC	78	130	572	24	804	0.209	0.097	0.112	2.54E-02	1.48E-02	4.22E-02	3.44E-67
ATA	86	816	195	241	1338	0.142	0.064	0.077	1.31E-01	9.61E-02	1.76E-01	1.08E-45
ATT	132	223	213	67	635	0.304	0.208	0.096	1.87E-01	1.29E-01	2.68E-01	6.38E-22
TTA	21	293	119	215	648	0.105	0.032	0.072	1.30E-01	7.50E-02	2.16E-01	2.31E-19
TAT	19	107	98	48	272	0.199	0.070	0.129	8.79E-02	4.54E-02	1.64E-01	1.10E-17
CGC	37	265	243	396	941	0.095	0.039	0.056	2.28E-01	1.51E-01	3.36E-01	2.75E-16
TGC	50	79	118	18	265	0.309	0.189	0.120	9.76E-02	4.96E-02	1.84E-01	1.21E-15
TGA	20	303	34	39	396	0.111	0.051	0.061	7.65E-02	3.77E-02	1.52E-01	1.06E-14
AAT	12	365	45	141	563	0.068	0.021	0.046	1.03E-01	4.84E-02	2.06E-01	4.39E-13
TAA	8	282	37	95	422	0.073	0.019	0.054	7.34E-02	2.85E-02	1.67E-01	5.48E-13
CAC	4	334	142	1151	1631	0.019	0.002	0.016	9.71E-02	2.59E-02	2.57E-01	3.78E-10
TGT	19	266	37	76	398	0.101	0.048	0.053	1.48E-01	7.55E-02	2.81E-01	1.01E-09
TTG	9	345	91	590	1035	0.033	0.009	0.024	1.69E-01	7.41E-02	3.42E-01	6.98E-09
AAG	35	378	83	262	758	0.085	0.046	0.039	2.93E-01	1.85E-01	4.55E-01	2.27E-08
AAA	31	517	67	319	934	0.062	0.033	0.028	2.86E-01	1.76E-01	4.55E-01	6.27E-08
GCC	3	38	58	44	143	0.122	0.021	0.101	6.10E-02	1.13E-02	2.11E-01	6.44E-08
TGG	13	190	26	68	297	0.090	0.044	0.046	1.80E-01	8.01E-02	3.87E-01	5.42E-06
ACA	110	347	217	372	1046	0.137	0.105	0.031	5.44E-01	4.10E-01	7.19E-01	3.04E-05
ACC	30	266	63	196	555	0.089	0.054	0.035	3.52E-01	2.11E-01	5.75E-01	3.80E-05
TCG	19	204	86	327	636	0.058	0.030	0.028	3.55E-01	1.97E-01	6.09E-01	1.46E-04
TAC	82	28	570	83	763	0.123	0.107	0.016	4.27E-01	2.57E-01	7.23E-01	3.22E-03
GGG	3	155	27	246	431	0.026	0.007	0.019	1.77E-01	3.38E-02	5.90E-01	3.62E-03
ACG	64	102	121	99	386	0.206	0.166	0.040	5.14E-01	3.33E-01	7.90E-01	3.87E-03
CTC	20	69	84	123	296	0.106	0.068	0.038	4.26E-01	2.27E-01	7.72E-01	8.58E-03
CTT	8	166	51	371	596	0.029	0.013	0.015	3.51E-01	1.41E-01	7.67E-01	1.01E-02
CGG	8	117	32	155	312	0.051	0.026	0.026	3.32E-01	1.27E-01	7.71E-01	1.30E-02
CCG	33	293	89	446	861	0.054	0.038	0.015	5.65E-01	3.57E-01	8.77E-01	1.97E-02
CTA	15	279	69	610	973	0.026	0.015	0.011	4.76E-01	2.48E-01	8.58E-01	1.97E-02
AGT	234	318	299	298	1149	0.223	0.204	0.019	7.34E-01	5.77E-01	9.32E-01	1.97E-02
AAC	7	578	40	1234	1859	0.008	0.004	0.004	3.74E-01	1.40E-01	8.50E-01	2.27E-02
GGT	14	399	40	526	979	0.023	0.014	0.009	4.62E-01	2.29E-01	8.81E-01	3.07E-02
CGA	31	433	63	534	1061	0.039	0.029	0.010	6.07E-01	3.74E-01	9.67E-01	5.71E-02
GTG	2	108	18	224	352	0.018	0.006	0.012	2.31E-01	2.56E-02	9.94E-01	8.37E-02
ATC	17	20	57	137	231	0.051	0.074	-0.022	2.04E+00	9.28E-01	4.43E+00	1.02E-01

	GGA	8	181	23	230	442	0.030	0.018	0.012	4.43E-01	1.67E-01	1.06E+00	1.05E-01
	AGG	85	341	57	326	809	0.092	0.105	-0.013	1.43E+00	9.72E-01	2.10E+00	1.11E-01
	TAG	11	386	12	183	592	0.026	0.019	0.007	4.35E-01	1.70E-01	1.10E+00	1.14E-01
	AGC	56	260	92	303	711	0.093	0.079	0.014	7.10E-01	4.80E-01	1.04E+00	1.27E-01
	GGC	2	54	17	127	200	0.027	0.010	0.017	2.78E-01	3.01E-02	1.24E+00	1.69E-01
	CCA	10	257	36	532	835	0.018	0.012	0.006	5.75E-01	2.51E-01	1.21E+00	2.26E-01
	GCG	2	93	13	200	308	0.015	0.006	0.009	3.32E-01	3.56E-02	1.51E+00	2.45E-01
	GTT	30	172	120	497	819	0.045	0.037	0.009	7.23E-01	4.50E-01	1.13E+00	2.57E-01
	CAA	1	422	13	1271	1707	0.002	0.001	0.001	2.32E-01	5.44E-03	1.55E+00	3.06E-01
	ACT	56	135	74	138	403	0.153	0.139	0.014	7.74E-01	4.96E-01	1.20E+00	3.44E-01
	TCC	4	65	31	262	362	0.018	0.011	0.007	5.21E-01	1.29E-01	1.55E+00	3.70E-01
	TCA	10	166	34	362	572	0.024	0.017	0.006	6.42E-01	2.76E-01	1.37E+00	4.12E-01
	CAT	23	770	79	2036	2908	0.010	0.008	0.002	7.70E-01	4.58E-01	1.25E+00	4.12E-01
	GCA	0	66	4	158	228	0.005	0.000	0.005	0.00E+00	0.00E+00	3.72E+00	4.25E-01
	GCT	8	101	25	201	335	0.032	0.024	0.008	6.38E-01	2.40E-01	1.52E+00	4.25E-01
	CCC	34	424	88	885	1431	0.027	0.024	0.004	8.07E-01	5.17E-01	1.23E+00	4.53E-01
	GAC	0	20	15	165	200	0.008	0.000	0.008	0.00E+00	0.00E+00	2.52E+00	4.57E-01
	ATG	42	282	50	276	650	0.071	0.065	0.006	8.22E-01	5.14E-01	1.31E+00	5.19E-01
	CTG	16	70	18	55	159	0.116	0.101	0.015	7.00E-01	3.03E-01	1.60E+00	5.19E-01
	TCT	1	88	14	428	531	0.005	0.002	0.003	3.48E-01	8.13E-03	2.34E+00	5.57E-01
	AGA	99	270	135	413	917	0.103	0.108	-0.005	1.12E+00	8.20E-01	1.53E+00	5.57E-01
	CAG	18	205	58	528	809	0.026	0.022	0.004	8.00E-01	4.32E-01	1.42E+00	5.62E-01
	CGT	13	277	13	365	668	0.017	0.019	-0.003	1.32E+00	5.53E-01	3.14E+00	6.04E-01
	GTC	8	51	41	187	287	0.035	0.028	0.007	7.16E-01	2.73E-01	1.68E+00	6.08E-01
	CCT	5	188	14	374	581	0.011	0.009	0.002	7.11E-01	1.97E-01	2.13E+00	6.67E-01
	GAG	7	141	17	265	430	0.019	0.016	0.003	7.74E-01	2.65E-01	2.02E+00	6.96E-01
	GAT	2	56	14	260	332	0.008	0.006	0.002	6.64E-01	7.13E-02	3.02E+00	7.72E-01
	GTA	17	169	70	614	870	0.021	0.020	0.002	8.82E-01	4.74E-01	1.57E+00	7.95E-01
	GAA	8	104	49	552	713	0.013	0.011	0.001	8.67E-01	3.44E-01	1.92E+00	8.50E-01
ent ire dat ase t		HDs w/ pos iti ve cha rge bot h pos 43 & 46	HDs w/ posit ive charg e only pos 43	HDs w/ posit ive charg e only pos 46	HDs w/ posit ive charg e neith er pos 43 & 46	total numbe r of HDs	expected frequency of cooccurre nce assuming independen ce	actual frequency cooccurren ce	differe nce	odds.ratio	conf.inte rval.95perc ent.low	conf.inte rval.95perc ent.hig h	p-value
		188 1	15043	5532	22025	44481	0.063	0.042	0.021	4.97E-01	4.70E-01	5.27E-01	3.74E-140

## Appendix A-6:

Final sequence cloned into pBIH2 $\omega$ 2-12En between NotI and XbaI to create pBIH2 $\omega$ 2-12En(SB):

```
GCGGCCGCGGACTACAAGGATGACGACGACAAGTTCCGGACCGGCTCCAAGACCCCGCCGCACGGCACGGCGCGCCC  
ATATGCTTGCCCTGTCGAGTCCTGCGATCGCCGCTTTTCTCGCTCGGATGAGCTTACCCGCCATATCCGCATCCACA  
CAGGCCAGAAGCCCTTCCAGTGTCTGAATCTGCATGCGTAACTTCAGTCGTAGTGACCACCTTACCACCCACATCCGC  
ACCCACACCGGTACCGGCCGTGAGAAGCGTCCACGCACCGCGTTCTCCAGCGAGCAGTTGGCCCGCCTTAAGCGGGA  
GTTCAACGAGAATCGCTATCTGACCGAGCGGAGACGCCAGCAGCTGAGCAGCGAGCTCGGCCTGAACGAGGCGCAGA  
TCAAGATCTGGTTCCAGAACAAGCGGGCCAAGATCAAGAAGTCGGGATCCTAATCTAGA
```

en K55 library:

```
CGGCCTGAACGAGNNSCAGATCNNSNNSTGGTTCNNSAACAAGCGGNNSAAGATCAAGAAGTCGG
```

en Library 5p comp:

```
CTCGTTCAGGCCGAGCT
```

en Library 3p comp:

```
GATCCCGACTTCTTGATCTT
```

# Appendix A-7:

Barcode	Barcode Strand 1 Sequence (no phosphorylation)	Barcode Strand 2 Sequence (no phosphorylation)
AA	aaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTttT
CA	caAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtgT
GA	gaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtcT
TA	taAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtaT
AC	acAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgtT
CC	ccAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTggT
GC	gcAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgcT
TC	tcAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgaT
AG	agAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTctT
CG	cgAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcgT
GG	ggAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTccT
TG	tgAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcaT
AT	atAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTatT
CT	ctAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTagT
GT	gtAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTacT
TT	ttAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTaaT
TTT	aaaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtttT
TGT	acaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtgtT
TCT	agaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtctT
TAT	ataAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtatT
GTT	gtaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcaaT
GGT	ggtAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTccaT
GCT	gctAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcgaT
GAT	ctaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgatT
CTT	gaaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcttT
CGT	gcaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcgtT
CCT	ggaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcctT
CAT	gtaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcatT
ATT	taaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTattT
AGT	tcaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTagtT
ACT	tgaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTactT
AAT	ttaAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTaatT
TTCT	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTttctT
TGGA	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtggT
TCAC	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtcacT
TATG	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtatgT
GTGC	gcacAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgtgcT

	T	
	cgccAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
GGCG	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTggcgT
	aagcAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
GCTT	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgcttT
	tttcAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
GAAA	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgaaaT
	ctagAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
CTAG	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTctagT
	gacgAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
CGTC	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcgtcT
	tgggAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
CCCA	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcccaT
	actgAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
CAGT	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcagtT
	taatAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
ATTA	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTattaT
	atctAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
AGAT	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTagatT
	ggttAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	
AACC	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTaaccT



**Appendix A-8:**

The triplet binding site of the HD selection	Barcode	Restriction Site
TTT	TTCT	BamHI
TTG	TGGA	BamHI
TTC	TCAC	BamHI
TTA	TATG	BamHI
TGT	GTGC	BamHI
TGG	GGCG	BamHI
TGC	GCTT	BamHI
TGA	GAAA	BamHI
TCT	CTAG	BamHI
TCG	CGTC	BamHI
TCC	CAGT	BamHI
TCA	CCCA	BamHI
TAT	ATTA	BamHI
TAG	AGAT	BamHI
TAC	ACGG	BamHI
TAA	AACC	BamHI
GTT	TTT	BamHI
GTG	TGT	BamHI
GTC	TCT	BamHI
GTA	TAT	BamHI
GGT	GTT	BamHI
GGG	GGT	BamHI
GGC	GCT	BamHI
GGA	GAT	BamHI
GCT	CTT	BamHI
GCG	CGT	BamHI
GCC	CCT	BamHI
GCA	CAT	BamHI
GAT	ATT	BamHI
GAG	AGT	BamHI
GAC	ACT	BamHI
GAA	AAT	BamHI
CTT	TTCT	XbaI
CTG	TGGA	XbaI
CTC	TCAC	XbaI
CTA	TATG	XbaI
CGT	GTGC	XbaI
CGG	GGCG	XbaI
CGC	GCTT	XbaI

CGA	GAAA	XbaI
CCT	CTAG	XbaI
CCG	CGTC	XbaI
CCC	CCCA	XbaI
CCA	CAGT	XbaI
CAT	ATTA	XbaI
CAG	AGAT	XbaI
CAC	ACGG	XbaI
CAA	AACC	XbaI
ATT	TTT	XbaI
ATG	TGT	XbaI
ATC	TCT	XbaI
ATA	TAT	XbaI
AGT	GTT	XbaI
AGG	GGT	XbaI
AGC	GCT	XbaI
AGA	GAT	XbaI
ACT	CTT	XbaI
ACG	CGT	XbaI
ACC	CCT	XbaI
ACA	CAT	XbaI
AAT	ATT	XbaI
AAG	AGT	XbaI
AAC	ACT	XbaI
AAA	AAT	XbaI





## Appendix A-9:

ATA\_LRWNS\_Top  
CGAGctcCAGATCcgctggTGGTTCaatAACAAGCGGagc  
ATA\_LRWNS\_Bottom  
TCTTGCTCCGCTTGTTATTGAACCACCAGCGGATCTGGAG

GCA\_RHDRA\_Top  
CGAGcggCAGATCcacgacTGGTTCcggAACAAGCGGgcc  
GCA\_RHDRA\_Bottom  
TCTTGCCCCGCTTGTTCCGGAACCAGTCGTGGATCTGCCG

GCA\_RYDRA\_Top  
CGAGcggCAGATCtacgacTGGTTCcggAACAAGCGGgcc  
GCA\_RYDRA\_Bottom  
TCTTGCCCCGCTTGTTCCGGAACCAGTCGTAGATCTGCCG

CTC\_VMNRK\_Top  
CGAGgttCAGATCatgaacTGGTTCcggAACAAGCGGaaa  
CTC\_VMNRK\_Bottom  
TCTTTTTCCGCTTGTTCCGGAACCAGTTCATGATCTGAAC

GTC\_YRRGA\_Top  
CGAGtacCAGATCcgtcggTGGTTCggtAACAAGCGGgcc  
GTC\_YRRGA\_Bottom  
TCTTGCCCCGCTTGTTACCGAACCACCGACGGATCTGGTA

GTC\_YRRGF\_Top  
CGAGtacCAGATCcgtcggTGGTTCggtAACAAGCGGttc  
GTC\_YRRGF\_Bottom  
TCTTGAACCGCTTGTTACCGAACCACCGACGGATCTGGTA

AAG\_RSQWH\_Top  
CGAGcgtCAGATCagccagTGGTTCtggAACAAGCGGcac  
AAG\_RSQWH\_Bottom  
TCTTGTGCCGCTTGTTCCAGAACCACCTGGCTGATCTGACG

# Appendix A-10:

# of HD_The triplet the HD variant was selected from_key residues of HD	Sequencing Barcode	Restriction Site for Illumina adaptor attachment
1_TTT_RTVA	TTT	BamHI
2_TTT_RTVA	TGT	BamHI
4_TTC_VRVSA	TCT	BamHI
5_TTC_TRVAA	TAT	BamHI
6_TTA_VRVAA	GTT	BamHI
7_TTA_RVLRA	GGT	BamHI
8_TGT_RVVSQ	GCT	BamHI
9_TGG_KTTQD	GAT	BamHI
10_TGG_KSVMQ	AA	BamHI
11_TGA_KSVAQ	CA	BamHI
12_TGA_RGVAA	GA	BamHI
13_TCT_ATVKA	TA	BamHI
14_TCG_KGTQM	AC	BamHI
15_TCC_RMIKS	CC	BamHI
17_TAT_TRVSA	GC	BamHI
18_TAG_RLTQA	TC	BamHI
19_TAG_RMVSA	AG	BamHI
20_TAC_QRVSA	CG	BamHI
21_TAC_ERVSV	GG	BamHI
22_TAA_RITAA	TG	BamHI
23_GTT_GTRAY	AT	BamHI
24_GTG_HLIQY	CT	BamHI
25_GTA_YTRQV	GT	BamHI
26_GGT_ALKNM	TT	BamHI
27_GGT_LTKDQ	TTCT	BamHI
28_GGG_RSKER	TGGA	BamHI
29_GGC_TLKNQ	TCAC	BamHI
30_GGA_LAKDQ	TATG	BamHI
31_GCT_KITKF	GTGC	BamHI
32_GCC_VRLKY	GGCG	BamHI
33_GCA_ALRQQ	GCTT	BamHI
34_GAT_RTMR	GAAA	BamHI
35_GAG_VMRWY	CTAG	BamHI
36_GAC_ATRRF	CGTC	BamHI
37_GAA_RFQKF	CCCA	BamHI
38_CTA_LHYAK	CAGT	BamHI
39_CTA_IFNAK	ATTA	BamHI

40_CGG_STRER	CTT	NcoI
41_CGC_RVMSR	CGT	NcoI
42_CGA_TFYAA	CCT	NcoI
43_CCT_MTNGK	CAT	NcoI
44_CCT_RGDSK	ATT	NcoI
45_CCG_RCYEK	AGT	NcoI
46_CCC_RLDSK	ACT	NcoI
47_CCA_KMTQK	AAT	NcoI
48_CCA_EHNAK	AA	NcoI
49_CAT_LSQSR	CA	NcoI
50_CAT_MMCSR	GA	NcoI
51_CAG_MSHWR	TA	NcoI
52_CAC_LGMRR	AC	NcoI
53_CAC_ERVSR	CC	NcoI
54_CAA_LMYQR	GC	NcoI
55_CAA_LHYVR	TC	NcoI
56_ATT_HRVQA	AG	NcoI
57_ATG_LTYQW	CG	NcoI
58_ATG_RVYQW	GG	NcoI
59_ATC_TRMAF	TG	NcoI
60_ATA_KTVQV	AT	NcoI
61_AGT_KGKEW	CT	NcoI
62_AGG_SHKEY	GT	NcoI
63_AGC_QSRNV	TT	NcoI
64_AGA_AFRAH	TTCT	NcoI
65_AGA_GSRWY	TGGA	NcoI
66_ACT_KTSHM	TCAC	NcoI
67_ACT_MKYEK	TATG	NcoI
68_ACG_SRYDR	GTGC	NcoI
69_ACG_VKYER	GGCG	NcoI
70_ACC_KTSHM	GCTT	NcoI
71_ACA_MTNNR	GAAA	NcoI
72_AAT_KMSNF	CTAG	NcoI
73_AAT_KLTAF	CGTC	NcoI
74_AAG_STSAH	CCCA	NcoI
75_AAC_SISRF	CAGT	NcoI
76_AAA_RAQWF	ATTA	NcoI
77_AAA_KEYVH	AGAT	NcoI
168_ACG_SRYDR	TGGA	BamHI
201_TGC_VRVSQ	TCAC	BamHI
202_GTG_NAREF	TATG	BamHI
203_GTC_VQKRF	GTGC	BamHI
204_GCG_RTDRY	GGCG	BamHI

205_GAA_TQRQW	GCTT	BamHI
207_CTT_ITYGK	GAAA	BamHI
208_CTC_HFNRK	CTAG	BamHI
209_CGC_PRDSR	CGTC	BamHI
210_CCG_RSNQK	CCCA	BamHI
211_ATT_TKNQN	CAGT	BamHI
212_ATA_RVTNA	ATTA	BamHI
213_AGG_KMKES	AGAT	BamHI
215_TTC_KRLAA	AACC	BamHI
216_TGC_NRVMM	TTT	BamHI
217_GGG_KSKEG	TGT	BamHI
218_CTG_KQNQK	TCT	BamHI
219_CTG_KVYER	TAT	BamHI
220_CTG_LTYQK	GTT	BamHI
221_CTG_RLYQK	GGT	BamHI
222_CTC_SKYGK	GCT	BamHI
223_CTC_RTFGK	GAT	BamHI
224_CCC_IMNSK	CTT	BamHI
225_AAC_SLQRF	CGT	BamHI
226_TTT_KMISA	ATT	BamHI
227_TTT_YRIAA	AGT	BamHI
228_TTG_KMLQA	ACT	BamHI
229_TTC_GRISA	AAT	BamHI
230_TGC_ERISQ	AA	BamHI
232_TCG_IKNQM	CA	BamHI
233_TCG_VMNQQ	GA	BamHI
234_TCA_AMVQR	TA	BamHI
235_TAT_RAVSV	AC	BamHI
236_TAG_KSTQM	CC	BamHI
237_TAG_YAVNA	GC	BamHI
238_TAC_QRISV	TC	BamHI
239_TAA_RTVRA	TTCT	NcoI
240_GTT_SSRGF	TGGA	NcoI
241_GTT_GLRAF	TCAC	NcoI
242_GTC_LQRG	TATG	NcoI
243_GGT_ATKSM	GTGC	NcoI
244_GGT_KMKSV	GGCG	NcoI
245_GCT_RAVKW	GCTT	NcoI
246_GCT_ISVKY	GAAA	NcoI
247_GCG_RTDRS	CTAG	NcoI
249_GCA_QLKQS	CGTC	NcoI
250_GAT_AGKTF	CCCA	NcoI
251_CTT_VGYSR	CAGT	NcoI



252_CTC_LRYSK	ATTA	NcoI
253_CGT_VANSR	AGAT	NcoI
255_CCT_RADGK	AACC	NcoI
256_CCA_RLYQK	TTT	NcoI
257_CAT_KLCSR	TGT	NcoI
258_ATT_RTVQQ	TCT	NcoI
259_ATA_KMYAW	TAT	NcoI
260_ATA_KAYNA	GTT	NcoI
261_AGG_KSKEA	GGT	NcoI
262_AGA_QFRAW	GCT	NcoI
263_AGA_VRFAA	GAT	NcoI
264_ACT_KVYHV	CTT	NcoI
265_ACG_WYSKY	CGT	NcoI
266_ACC_KACHS	CCT	NcoI
267_ACA_RVSHT	CAT	NcoI
268_AAT_KLQAF	ATT	NcoI
269_AAT_KVTNF	AGT	NcoI
270_AAG_RAQWF	ACT	NcoI
271_AAC_KLQRF	AAT	NcoI
272_AAC_VAQRC	AG	NcoI
1_AAG_RSQWH	TTCT	NcoI
2_ATA_LRWNS	TGGA	NcoI
3_CTC_VMNRK	TCAC	NcoI
4_CTG_TTNQK	TATG	NcoI
5_GAT_VGRLY	GTGC	NcoI
6_GCA_RHDRA	GGCG	NcoI
7_GCA_RYDRA	GCTT	NcoI
8_GCG_RLDRF_	GAAA	NcoI
9_GCG_RLDRY	CTAG	NcoI
10_GTC_YRRGA	CGTC	NcoI
11_GTC_YRRGF	CCCA	NcoI
12_no_HD	CAGT	NcoI
13_12-En(SB)	ATTA	NcoI

## REFERENCES:

- Ades SE, Sauer RT. 1994. Differential DNA-binding specificity of the engrailed homeodomain: the role of residue 50. *Biochemistry* **33**(31): 9187-9194.
- . 1995. Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex. *Biochemistry* **34**(44): 14601-14608.
- Alleyne TM, Pena-Castillo L, Badis G, Talukder S, Berger MF, Gehrke AR, Philippakis AA, Bulyk ML, Morris QD, Hughes TR. 2009. Predicting the binding preference of transcription factors to individual DNA k-mers. *Bioinformatics* **25**(8): 1012-1018.
- Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. 2004. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* **14**(3): 283-291.
- Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**(5935): 1720-1723.
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**(Web Server issue): W202-208.
- Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36.
- Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**(5967): 836-840.
- Bedell VM, Wang Y, Campbell JM, Poshusta TL, Starker CG, Krug RG, 2nd, Tan W, Penheiter SG, Ma AC, Leung AY et al. 2012. In vivo genome editing using a high-efficiency TALEN system. *Nature* **491**(7422): 114-118.
- Beerli RR, Segal DJ, Dreier B, Barbas CF, 3rd. 1998. Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proc Natl Acad Sci U S A* **95**(25): 14628-14633.
- Benos PV, Lapedes AS, Stormo GD. 2002. Probabilistic code for DNA recognition by proteins of the EGR family. *J Mol Biol* **323**(4): 701-727.
- Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET et al. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**(7): 1266-1276.
- Bogdanove AJ, Voytas DF. 2011. TAL effectors: customizable proteins for DNA targeting. *Science* **333**(6051): 1843-1846.
- Brenner S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* **77**(1): 71-94.
- Brivanlou AH, Darnell JE, Jr. 2002. Signal transduction and the control of gene expression. *Science* **295**(5556): 813-818.
- Burglin TR. 2011. Homeodomain subtypes and functional diversity. *Subcell Biochem* **52**: 95-122.

- Carlson DF, Tan W, Lillico SG, Stverakova D, Proudfoot C, Christian M, Voytas DF, Long CR, Whitelaw CB, Fahrenkrug SC. 2012. Efficient TALEN-mediated gene knockout in livestock. *Proc Natl Acad Sci U S A* **109**(43): 17382-17387.
- Carroll D. 2011. Zinc-finger nucleases: a panoramic view. *Curr Gene Ther* **11**(1): 2-10.
- Carroll D, Beumer KJ, Trautman JK. 2010. High-efficiency gene targeting in *Drosophila* with zinc finger nucleases. *Methods Mol Biol* **649**: 271-280.
- Choo Y, Sanchez-Garcia I, Klug A. 1994. In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature* **372**(6507): 642-645.
- Christensen RG, Enuameh MS, Noyes MB, Brodsky MH, Wolfe SA, Stormo GD. 2012. Recognition Models to Predict DNA-binding Specificities of Homeodomain Proteins. *Bioinformatics*.
- Chu SW, Noyes MB, Christensen RG, Pierce BG, Zhu LJ, Weng Z, Stormo GD, Wolfe SA. 2012. Exploring the DNA-recognition potential of homeodomains. *Genome Res* **22**(10): 1889-1898.
- Connolly JP, Augustine JG, Francklyn C. 1999. Mutational analysis of the engrailed homeodomain recognition helix by phage display. *Nucleic Acids Res* **27**(4): 1182-1189.
- Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**(6): 1188-1190.
- Damante G, Pellizzari L, Esposito G, Fogolari F, Viglino P, Fabbro D, Tell G, Formisano S, Di Lauro R. 1996. A molecular code dictates sequence-specific DNA recognition by homeodomains. *EMBO J* **15**(18): 4992-5000.
- De Masi F, Grove CA, Vedenko A, Alibes A, Gisselbrecht SS, Serrano L, Bulys ML, Walhout AJ. 2011. Using a structural and logics systems approach to infer bHLH-DNA binding specificity determinants. *Nucleic Acids Res* **39**(11): 4553-4563.
- Deppmann CD, Alvania RS, Taparowsky EJ. 2006. Cross-species annotation of basic leucine zipper factor interactions: Insight into the evolution of closed interaction networks. *Molecular biology and evolution* **23**(8): 1480-1492.
- Doyon Y, McCammon JM, Miller JC, Faraji F, Ngo C, Katibah GE, Amora R, Hocking TD, Zhang L, Rebar EJ et al. 2008. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nat Biotechnol* **26**(6): 702-708.
- Doyon Y, Vo TD, Mendel MC, Greenberg SG, Wang J, Xia DF, Miller JC, Urnov FD, Gregory PD, Holmes MC. 2011. Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nat Methods* **8**(1): 74-79.
- Ekker SC, Jackson DG, von Kessler DP, Sun BI, Young KE, Beachy PA. 1994. The degree of variation in DNA sequence recognition among four *Drosophila* homeotic proteins. *EMBO J* **13**(15): 3551-3560.
- Elrod-Erickson M, Rould MA, Neklodova L, Pabo CO. 1996. Zif268 protein-DNA complex refined at 1.6 Å: a model system for understanding zinc finger-DNA interactions. *Structure* **4**(10): 1171-1180.
- Emerson RO, Thomas JH. 2009. Adaptive evolution in zinc finger transcription factors. *PLoS Genet* **5**(1): e1000325.

- Enuameh MS, Asriyan Y, Richards A, Christensen RG, Hall VL, Kazemian M, Zhu C, Pham H, Cheng Q, Blatti C et al. 2013a. Global analysis of *Drosophila* Cys2-His2 zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome Res.*
- 2013b. Global analysis of *Drosophila* Cys(2)-His(2) zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome Res* **23**(6): 928-940.
- Fraenkel E, Rould MA, Chambers KA, Pabo CO. 1998. Engrailed homeodomain-DNA complex at 2.2 Å resolution: a detailed view of the interface and comparison with other engrailed structures. *J Mol Biol* **284**(2): 351-361.
- Friedland AE, Tzur YB, Esvelt KM, Colaiacovo MP, Church GM, Calarco JA. 2013. Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. *Nat Methods* **10**(8): 741-743.
- Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, Sander JD. 2013. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol.*
- Gehring WJ, Affolter M, Burglin T. 1994a. Homeodomain proteins. *Annu Rev Biochem* **63**: 487-526.
- Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, Affolter M, Otting G, Wuthrich K. 1994b. Homeodomain-DNA recognition. *Cell* **78**(2): 211-223.
- Grant RA, Rould MA, Klemm JD, Pabo CO. 2000. Exploring the role of glutamine 50 in the homeodomain-DNA interface: crystal structure of engrailed (Gln50 → ala) complex at 2.0 Å. *Biochemistry* **39**(28): 8187-8192.
- Gratz SJ, Cummings AM, Nguyen JN, Hamm DC, Donohue LK, Harrison MM, Wildonger J, O'Connor-Giles KM. 2013. Genome Engineering of *Drosophila* with the CRISPR RNA-Guided Cas9 Nuclease. *Genetics*.
- Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, Walhout AJ. 2009. A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138**(2): 314-327.
- Guo J, Gaj T, Barbas CF, 3rd. 2010. Directed evolution of an enhanced and highly efficient FokI cleavage domain for zinc finger nucleases. *J Mol Biol* **400**(1): 96-107.
- Gupta A, Christensen RG, Rayla AL, Lakshmanan A, Stormo GD, Wolfe SA. 2012. An optimized two-finger archive for ZFN-mediated gene targeting. *Nat Methods* **9**(6): 588-590.
- Gupta A, Meng X, Zhu LJ, Lawson ND, Wolfe SA. 2010. Zinc finger protein-dependent and -independent contributions to the in vivo off-target activity of zinc finger nucleases. *Nucleic Acids Res* **39**(1): 381-392.
- 2011. Zinc finger protein-dependent and -independent contributions to the in vivo off-target activity of zinc finger nucleases. *Nucleic Acids Res* **39**(1): 381-392.
- Hanes SD, Brent R. 1991. A genetic model for interaction of the homeodomain recognition helix with DNA. *Science* **251**(4992): 426-430.
- Herskowitz I. 1989. A regulatory hierarchy for cell specialization in yeast. *Nature* **342**(6251): 749-757.

- Holkers M, Maggio I, Liu J, Janssen JM, Miselli F, Mussolino C, Recchia A, Cathomen T, Goncalves MA. 2013. Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic Acids Res* **41**(5): e63.
- Holt N, Wang J, Kim K, Friedman G, Wang X, Taupin V, Crooks GM, Kohn DB, Gregory PD, Holmes MC et al. 2010. Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nat Biotechnol* **28**(8): 839-847.
- Hovde S, Abate-Shen C, Geiger JH. 2001. Crystal structure of the Msx-1 homeodomain/DNA complex. *Biochemistry* **40**(40): 12013-12021.
- Hwang WY, Fu Y, Reyon D, Maeder ML, Kaini P, Sander JD, Joung JK, Peterson RT, Yeh JR. 2013. Heritable and Precise Zebrafish Genome Editing Using a CRISPR-Cas System. *PLoS One* **8**(7): e68708.
- Ihaka R, Gentleman R. 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5**(3): 299-314.
- Jauch R, Ng CK, Saikatendu KS, Stevens RC, Kolatkar PR. 2008. Crystal structure and DNA binding of the homeodomain of the stem cell transcription factor Nanog. *J Mol Biol* **376**(3): 758-770.
- Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**(6096): 816-821.
- Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. 2013. RNA-programmed genome editing in human cells. *Elife* **2**: e00471.
- Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**(1-2): 327-339.
- Joshi R, Passner JM, Rohs R, Jain R, Sosinsky A, Crickmore MA, Jacob V, Aggarwal AK, Honig B, Mann RS. 2007. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131**(3): 530-543.
- Joung JK, Sander JD. 2013. TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol* **14**(1): 49-55.
- Kapiloff MS, Farkash Y, Wegner M, Rosenfeld MG. 1991. Variable effects of phosphorylation of Pit-1 dictated by the DNA response elements. *Science* **253**(5021): 786-789.
- Kim E, Kim S, Kim DH, Choi BS, Choi IY, Kim JS. 2012. Precision genome engineering with programmable DNA-nicking enzymes. *Genome Res* **22**(7): 1327-1333.
- Kim YG, Cha J, Chandrasegaran S. 1996. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc Natl Acad Sci U S A* **93**(3): 1156-1160.
- Kim YG, Chandrasegaran S. 1994. Chimeric restriction endonuclease. *Proc Natl Acad Sci U S A* **91**(3): 883-887.
- Kissinger CR, Liu BS, Martin-Blanco E, Kornberg TB, Pabo CO. 1990. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell* **63**(3): 579-590.
- Klug A. 2010. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu Rev Biochem* **79**: 213-231.

- Kornberg TB. 1993. Understanding the homeodomain. *J Biol Chem* **268**(36): 26813-26816.
- Koshland DE, Jr. 2002. Special essay. The seven pillars of life. *Science* **295**(5563): 2215-2216.
- Kuziora MA, McGinnis W. 1989. A homeodomain substitution changes the regulatory specificity of the deformed protein in *Drosophila* embryos. *Cell* **59**(3): 563-571.
- Lin SY, Riggs AD. 1972. Lac repressor binding to non-operator DNA: detailed studies and a comparison of equilibrium and rate competition methods. *J Mol Biol* **72**(3): 671-690.
- Liu J, Stormo GD. 2008. Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics* **24**(17): 1850-1857.
- Liu LA, Bradley P. 2012. Atomistic modeling of protein-DNA interaction specificity: progress and applications. *Curr Opin Struct Biol* **22**(4): 397-405.
- Liu PQ, Rebar EJ, Zhang L, Liu Q, Jamieson AC, Liang Y, Qi H, Li PX, Chen B, Mendel MC et al. 2001. 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. *J Biol Chem* **276**(14): 11323-11334.
- Maeder ML, Linder SJ, Reyon D, Angstman JF, Fu Y, Sander JD, Joung JK. 2013. Robust, synergistic regulation of human gene expression using TALE activators. *Nat Methods* **10**(3): 243-245.
- Magari SR, Rivera VM, Iulucci JD, Gilman M, Cerasoli F, Jr. 1997. Pharmacologic control of a humanized gene therapy system implanted into nude mice. *J Clin Invest* **100**(11): 2865-2872.
- Mahony S, Auron PE, Benos PV. 2007. Inferring protein-DNA dependencies using motif alignments and mutual information. *Bioinformatics* **23**(13): i297-304.
- Mann RS, Hogness DS. 1990. Functional dissection of Ultrabithorax proteins in *D. melanogaster*. *Cell* **60**(4): 597-610.
- Mann RS, Lelli KM, Joshi R. 2009. Hox specificity unique roles for cofactors and collaborators. *Curr Top Dev Biol* **88**: 63-101.
- Mathias JR, Zhong H, Jin Y, Vershon AK. 2001. Altering the DNA-binding specificity of the yeast Matalpha 2 homeodomain protein. *J Biol Chem* **276**(35): 32696-32703.
- McGinnis W, Levine MS, Hafen E, Kuroiwa A, Gehring WJ. 1984. A conserved DNA sequence in homoeotic genes of the *Drosophila* Antennapedia and bithorax complexes. *Nature* **308**(5958): 428-433.
- Meng X, Noyes MB, Zhu LJ, Lawson ND, Wolfe SA. 2008. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nat Biotechnol* **26**(6): 695-701.
- Miller JC, Holmes MC, Wang J, Guschin DY, Lee YL, Rupniewski I, Beausejour CM, Waite AJ, Wang NS, Kim KA et al. 2007. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat Biotechnol* **25**(7): 778-785.
- Moore M, Choo Y, Klug A. 2001. Design of polyzinc finger peptides with structured linkers. *Proc Natl Acad Sci U S A* **98**(4): 1432-1436.

- Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**(5967): 876-879.
- Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML. 2013. DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proc Natl Acad Sci U S A* **110**(30): 12349-12354.
- Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008a. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**(7): 1277-1289.
- Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA. 2008b. A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Res* **36**(8): 2547-2560.
- Nusslein-Volhard C, Wieschaus E. 1980. Mutations affecting segment number and polarity in *Drosophila*. *Nature* **287**(5785): 795-801.
- Osborn MJ, Starker CG, McElroy AN, Webber BR, Riddle MJ, Xia L, DeFeo AP, Gabriel R, Schmidt M, von Kalle C et al. 2013. TALEN-based gene correction for epidermolysis bullosa. *Mol Ther* **21**(6): 1151-1159.
- Otting G, Qian YQ, Billeter M, Muller M, Affolter M, Gehring WJ, Wuthrich K. 1990. Protein--DNA contacts in the structure of a homeodomain--DNA complex determined by nuclear magnetic resonance spectroscopy in solution. *EMBO J* **9**(10): 3085-3092.
- Pabo CO, Sauer RT. 1992. Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem* **61**: 1053-1095.
- Passner JM, Ryoo HD, Shen L, Mann RS, Aggarwal AK. 1999. Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature* **397**(6721): 714-719.
- Pattanayak V, Ramirez CL, Joung JK, Liu DR. 2011. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat Methods* **8**(9): 765-770.
- Percival-Smith A, Muller M, Affolter M, Gehring WJ. 1990. The interaction with DNA of wild-type and mutant fushi tarazu homeodomains. *EMBO J* **9**(12): 3967-3974.
- Perez-Pinera P, Ousterout DG, Gersbach CA. 2012. Advances in targeted genome editing. *Curr Opin Chem Biol* **16**(3-4): 268-277.
- Persikov AV, Osada R, Singh M. 2009. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics* **25**(1): 22-29.
- Persikov AV, Singh M. 2011. An expanded binding model for Cys2His2 zinc finger protein-DNA interfaces. *Phys Biol* **8**(3): 035010.
- Piper DE, Batchelor AH, Chang CP, Cleary ML, Wolberger C. 1999. Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. *Cell* **96**(4): 587-597.
- Pomerantz JL, Sharp PA. 1994. Homeodomain determinants of major groove recognition. *Biochemistry* **33**(36): 10851-10858.
- Pomerantz JL, Sharp PA, Pabo CO. 1995. Structure-based design of transcription factors. *Science* **267**(5194): 93-96.

- Porteus MH, Baltimore D. 2003. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**(5620): 763.
- Ptashne M. 1992. *A Genetic Switch*.
- Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**(5): 1173-1183.
- Ramirez CL, Certo MT, Mussolino C, Goodwin MJ, Cradick TJ, McCaffrey AP, Cathomen T, Scharenberg AM, Joung JK. 2012. Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects. *Nucleic Acids Res* **40**(12): 5560-5568.
- Ramirez CL, Foley JE, Wright DA, Muller-Lerch F, Rahman SH, Cornu TI, Winfrey RJ, Sander JD, Fu F, Townsend JA et al. 2008. Unexpected failure rates for modular assembly of engineered zinc fingers. *Nat Methods* **5**(5): 374-375.
- Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**(7457): 172-177.
- Rebar EJ, Huang Y, Hickey R, Nath AK, Meoli D, Nath S, Chen B, Xu L, Liang Y, Jamieson AC et al. 2002. Induction of angiogenesis in a mouse model using engineered transcription factors. *Nat Med* **8**(12): 1427-1432.
- Reece-Hoyes JS, Pons C, Diallo A, Mori A, Shrestha S, Kadreppa S, Nelson J, Diprima S, Dricot A, Lajoie BR et al. 2013. Extensive Rewiring and Complex Evolutionary Dynamics in a *C. elegans* Multiparameter Transcription Factor Network. *Mol Cell* **51**(1): 116-127.
- Reyon D, Khayter C, Regan MR, Joung JK, Sander JD. 2012a. Engineering designer transcription activator-like effector nucleases (TALENs) by REAL or REAL-Fast assembly. *Curr Protoc Mol Biol* **Chapter 12**: Unit 12 15.
- Reyon D, Tsai SQ, Khayter C, Foden JA, Sander JD, Joung JK. 2012b. FLASH assembly of TALENs for high-throughput genome editing. *Nat Biotechnol* **30**(5): 460-465.
- Rivera VM, Clackson T, Natesan S, Pollock R, Amara JF, Keenan T, Magari SR, Phillips T, Courage NL, Cerasoli F, Jr. et al. 1996. A humanized system for pharmacologic control of gene expression. *Nat Med* **2**(9): 1028-1032.
- Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. 2010. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* **79**: 233-269.
- Ryan MP, Jones R, Morse RH. 1998. SWI-SNF complex participation in transcriptional activation at a step subsequent to activator binding. *Mol Cell Biol* **18**(4): 1774-1782.
- Ryder SP, Recht MI, Williamson JR. 2008. Quantitative analysis of protein-RNA interactions by gel mobility shift. *Methods Mol Biol* **488**: 99-115.
- Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**(5981): 1036-1040.
- Silva G, Poirot L, Galetto R, Smith J, Montoya G, Duchateau P, Paques F. 2011. Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. *Curr Gene Ther* **11**(1): 11-27.



- Simon MD, Shokat KM. 2004. Adaptability at a protein-DNA interface: re-engineering the engrailed homeodomain to recognize an unnatural nucleotide. *J Am Chem Soc* **126**(26): 8078-8079.
- Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ et al. 2011. Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. *Cell* **147**(6): 1270-1282.
- Soldner F, Laganieri J, Cheng AW, Hockemeyer D, Gao Q, Alagappan R, Khurana V, Golbe LI, Myers RH, Lindquist S et al. 2011. Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell* **146**(2): 318-331.
- Song J, Singh M. 2013. From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interactome at different scales of organization. *PLoS Comput Biol* **9**(2): e1002910.
- Steadman DJ, Giuffrida D, Gelmann EP. 2000. DNA-binding sequence of the human prostate-specific homeodomain protein NKX3.1. *Nucleic Acids Res* **28**(12): 2389-2395.
- Szczepek M, Brondani V, Buchel J, Serrano L, Segal DJ, Cathomen T. 2007. Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nat Biotechnol* **25**(7): 786-793.
- Thomas MC, Chiang CM. 2006. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* **41**(3): 105-178.
- Treisman J, Gonczy P, Vashishtha M, Harris E, Desplan C. 1989. A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell* **59**(3): 553-562.
- Tucker-Kellogg L, Rould MA, Chambers KA, Ades SE, Sauer RT, Pabo CO. 1997. Engrailed (Gln50-->Lys) homeodomain-DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity. *Structure* **5**(8): 1047-1054.
- Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC. 2005. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**(7042): 646-651.
- Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. 2010. Genome editing with engineered zinc finger nucleases. *Nat Rev Genet* **11**(9): 636-646.
- Valton J, Dupuy A, Daboussi F, Thomas S, Marechal A, Macmaster R, Melliand K, Juillerat A, Duchateau P. 2012. Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *J Biol Chem* **287**(46): 38427-38432.
- van Nimwegen E. 2003. Scaling laws in the functional content of genomes. *Trends Genet* **19**(9): 479-484.
- Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**(4): 252-263.
- Wah DA, Bitinaite J, Schildkraut I, Aggarwal AK. 1998. Structure of FokI has implications for DNA cleavage. *Proc Natl Acad Sci U S A* **95**(18): 10564-10569.

- Wang J, Friedman G, Doyon Y, Wang NS, Li CJ, Miller JC, Hua KL, Yan JJ, Babiarz JE, Gregory PD et al. 2012. Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Res* **22**(7): 1316-1326.
- Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR et al. 2010. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J* **29**(13): 2147-2160.
- Wirt SE, Porteus MH. 2012. Development of nuclease-mediated site-specific genome modification. *Curr Opin Immunol* **24**(5): 609-616.
- Wolberger C, Vershon AK, Liu B, Johnson AD, Pabo CO. 1991. Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* **67**(3): 517-528.
- Wolfe SA, Grant RA, Pabo CO. 2003. Structure of a designed dimeric zinc finger protein bound to DNA. *Biochemistry* **42**(46): 13401-13409.
- Wood AJ, Lo TW, Zeitler B, Pickle CS, Ralston EJ, Lee AH, Amora R, Miller JC, Leung E, Meng X et al. 2011. Targeted genome editing across species using ZFNs and TALENs. *Science* **333**(6040): 307.
- Wyman C, Kanaar R. 2006. DNA double-strand break repair: all's well that ends well. *Annu Rev Genet* **40**: 363-383.
- Yanover C, Bradley P. 2011. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Res* **39**(11): 4564-4576.
- Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. 2011. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat Biotechnol* **29**(2): 149-153.
- Zhao H, Yang Y, Zhou Y. 2010. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. *Bioinformatics* **26**(15): 1857-1863.
- Zhu C, Gupta A, Hall VL, Rayla AL, Christensen RG, Dake B, Lakshmanan A, Kuperwasser C, Stormo GD, Wolfe SA. 2013. Using defined finger-finger interfaces as units of assembly for constructing zinc-finger nucleases. *Nucleic Acids Res* **41**(4): 2455-2465.
- Zhu C, Smith T, McNulty J, Rayla AL, Lakshmanan A, Siekmann AF, Buffardi M, Meng X, Shin J, Padmanabhan A et al. 2011. Evaluation and application of modularly assembled zinc-finger nucleases in zebrafish. *Development* **138**(20): 4555-4564.
- Zhuang JJ, Hunter CP. 2012. RNA interference in *Caenorhabditis elegans*: uptake, mechanism, and regulation. *Parasitology* **139**(5): 560-573.
- Zu Y, Tong X, Wang Z, Liu D, Pan R, Li Z, Hu Y, Luo Z, Huang P, Wu Q et al. 2013. TALEN-mediated precise genome modification by homologous recombination in zebrafish. *Nat Methods* **10**(4): 329-331.