2009-03-20

An Omega-Based Bacterial One-Hybrid System for the Determination of Transcription Factor Specificity

Marcus Blaine Noyes
University of Massachusetts Medical School

AN OMEGA-BASED BACTERIAL ONE-HYBRID SYSTEM FOR THE DETERMINATION OF TRANSCRIPTION FACTOR SPECIFICITY

Marcus B Noyes

University of Massachusetts Medical School

AN OMEGA-BASED BACTERIAL ONE-HYBRID SYSTEM FOR THE DETERMINATION OF TRANSCRIPTION FACTOR SPECIFICITY

A Dissertation Presented

By

Marcus Blaine Noyes

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

November 7th, 2008

Department of Biochemistry and Molecular Pharmacology

Program in Gene Function and Expression

AN OMEGA-BASED BACTERIAL ONE-HYBRID SYSTEM FOR THE DETERMINATION OF TRANSCRIPTION FACTOR SPECIFICITY

A Dissertation Presented
By
Marcus Blaine Noyes

The signatures of the Dissertation Defense Committee signifies
Completion and approval as to style and content of the Dissertation

Scot Wolfe, Ph.D., Thesis Advisor

Keith Joung, MD, Ph.D., Member of Committee

Ken Knight, Ph.D., Member of Committee

Stephen Miller, 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

Martin Marinus, 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

Department of Biochemistry and Molecular Pharmacology
Program in Gene Function and Expression
November 7th, 2008

This and all of my future works are dedicated to my wife Tania Marie Castañeda. You are the passion, the motivation, and the happiness which makes me so certain that everyday will be a good day.

ACKNOWLEDGEMENTS

I would like to start my acknowledgements by thanking three very important members of my family. First, my wife. It is not easy to marry a scientist, especially a crazy one. She has been incredibly understanding of all the time I am in the lab late at night, early mornings, on weekends, weeknights, through holidays…I am so thankful that she is in my life and been with me through this process and I am extremely appreciative of her understanding. We have shared together the struggles and elations that come with this work and it is so much more meaningful with her. She has supported me not only with belief and encouragement but with her own sacrifices. I am honored to be with her every day.

I would also like to thank my parents. What wonderful encouragement they have provided throughout my life. There has never been a moment when I felt there was something I couldn't do. I was taught to believe that I had the ability to whatever I wanted and that whatever that was, I had their support. When I was a child and I said I wanted to be president, of course I could! The next day when I wanted to be a rodeo clown, of course I could! They have always been behind me 100%, for good or bad. They have always believed in me and therefore I believe in me.

There are a number of people at UMass Medical I would also like to thank. Of course there are several faculty. I would like to thank my committee members for their input over the last 4 years Ken Knight, Martin Marinus, Stephen Miller, and Marian Walhout. I would especially like to thank Marian Walhout, Nathan Lawson, and Mike Brodsky for their input on my research, help in preparation of presentations that I've given, and their support in my transition to the next phase of my career. I would also like to thank Joe McNulty, a member of the Wolfe lab, for all of the responsibilities he has taken on over the past few year that have enabled me to focus on my research relatively uninterrupted. Finally I would like to thank Mike Cole. If it weren't for the impressive and professional recruitment program that he runs, I may not have chosen UMass Medical in the first place.

The greatest contribution to my professional development and perhaps my entire career has come from my mentor, Scot Wolfe. As Scot's first graduating student, I had the honor of working with Scot side by side when I first joined his almost brand new lab. I had what seems to have become a rare experience; I was trained directly by my mentor. I personally think this was a huge advantage. Scot not only taught me how to do an experiment but why to do it, why each control was critical and how it was fitting into a larger picture. I think this comprehensive understanding from his point of view taught me to think of problems in a similar manner. Now I feel because I was trained by Scot that I will graduate not only with a developed skill set but the ability to deconstruct almost any problem and develop a reasonable plan to solve it.

In addition to the skills I have developed under Scot's guidance, I think the greatest thing I will take away from this experience is his example of leadership. In or out of the lab, Scot is an incredibly hard worker and his example has set the standard for the entire lab. In fact, the number of times that I have been able to get into the lab before Scot, I could count on one hand. But hard work doesn't make a leader. Scot's real leadership is demonstrated by how he treats people. Scot has always been very supportive. I have never felt belittled by Scot, overlooked or talked down to. On the contrary, more often than not Scot gives too much praise and does not take enough credit for his own contribution to a project. He also seems tireless in his ability to help just about anyone work through a problem and takes every inquiry seriously. Perhaps the best demonstration of Scot's leadership is that I have never heard him criticize anyone. Jokingly or otherwise, Scot always finds something good to say about people. He conveys an admirable level of respect for everyone, both collaborators and competitors. I hope that someday when I have my own lab that I am able to lead with the same level of grace and dignity.

ABSTRACT

From the yeast genome completed in 1996 to the 12 *Drosophila* genomes published earlier this year; little more than a decade has provided an incredible amount of genomic data. Yet even with this mountain of genetic information the regulatory networks that control gene expression remain relatively undefined. In part, this is due to the enormous amount of non-coding DNA, over 98% of the human genome, which needs to be made sense of. It is also due to the large number of transcription factors, potentially 2,000 such factors in the human genome, which may contribute to any given network directly or indirectly. Certainly, one of the central limitations has been the paucity of transcription factor (TF) specificity data that would aid in the prediction of regulatory targets throughout a
genome.

The general lack of specificity data has hindered the prediction of regulatory targets for individual TFs as well as groups of factors that function within a common regulatory pathway. A large collection of factor specificities would allow for the combinatorial prediction of regulatory targets that considers all factors actively expressed in a given cell, under a given condition. Herein we describe substantial improvements to a previous bacterial one-hybrid system with increased sensitivity and dynamic range that make it amenable for the high-throughput analysis of sequence-specific TFs. Currently we have characterized 108 (14.3%) of the predicted TFs in *Drosophila* that fall into a broad range

of DNA-binding domain families, demonstrating the feasibility of characterizing a large number of TFs using this technology.

To fully exploit our large database of binding specificities, we have created a GBrowse-based search tool that allows an end-user to examine the overrepresentation of binding sites for any number of individual factors as well as combinations of these factors in up to six *Drosophila* genomes (veda.cs.uiuc.edu/cgi-bin/gbrowse/gbrowse/Dmel4). We have used this tool to demonstrate that a collection of factor specificities within a common pathway will successfully predict previously validated *cis*-regulatory modules within a genome. Furthermore, within our database we provide a complete catalog of DNA-binding specificities for all 84 homeodomains in *Drosophila.* This catalog enabled us to propose and test a detailed set of recognition rules for homeodomains and use this information to predict the specificities of the majority of homeodomains in the human genome.

Table of Contents

DEDICATION iv

ACKNOWLEDGMENTS v

ABSTRACT viii

TABLE OF CONTENTS x

LIST OF TABLES xiv

LIST OF FIGURES xv

PREFACE xix

CHAPTER I INTRODUCTION 1

Summary 2

Transcription 3

Regulation 7

DNA-binding domains 12

Regulatory Networks 25

Yeast Regulatory Network 26

A Metazoan Developmental Network 28

Identifying regulatory elements in metazoans 44

Characterizing Transcription Factor Specificity 48

CHAPTER II CHARACTERIZATION OF A BACTERIAL ONE-HYBRID SYSTEM FOR THE DETERMINATION OF DNA-BINDING SPECIFICITY 57

Portions of Chapter II have been published previously as:

Noyes MB, Meng, X. Wakabayashi A, Sinha S, Brodsky M, and Wolfe S. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Research, ***36****, 2547-60.*

Introduction 58

Results 63

Discussion 85

Experimental Procedures 87

Acknowledgements 111

CHAPTER III THE SYSTEMATIC CHARACTERIZATION OF DNA-BINDING SPECIFICITIES FOR FACTORS INVOLVED IN THE ANTERIOR – POSTERIOR SEGMENTATION OF THE *DROSOPHILA* EMBRYO 112

Portions of Chapter III have been published previously as:

Noyes MB, Meng, X. Wakabayashi A, Sinha S, Brodsky M, and Wolfe S. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Research, ***36****, 2547-60.*

Introduction 113

Results 118

Discussion 147

Experimental Procedures 151

Acknowledgements 155

CHAPTER IV A COMPREHENSIVE CATALOG OF HOMEODOMAIN DNA-BINDING SPECIFICITIES FROM *D. MELANOGASTER* 156

Chapter IV has been published previously as:

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, and Wolfe SA (2008) Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites. Cell, 133, 1277-89.

Introduction 157

Results 163

Discussion 240

Experimental Procedures 250

Acknowledgements 264

CHAPTER V GENERAL DISCUSSION 265

The Bacterial One-hybrid System 266

The Genome Surveyor 268

A Catalog of Homeodomain Specificities 270

Specificity Determinants 272

CRM Prediction with Homeodomain Specificities 275

Artificial Domains 277

REFERENCES 280

APPENDIX 300

List of Tables

Table 3.1 Evaluation of CRM scoring function using 48 defined CRMs and 4800 random regions. 134

Table 3.2 Top twenty matches in a genome-wide search with Bcd, Hb, Hkb, Kr & Tll. 140

Table 4.1 Selected sequences for Bicoid, Deformed and En. 181

Table 4.2 Crossvalidation analysis of fly homeodomains. 232

Table 4.3 Prediction of TRANSFAC homeodomains 236

Table 4.4 Examples of functional (*in vivo*) homeodomain binding sites consistent with their monomeric specificity 244

Table A.1 Amino acid sequence and selected binding sites for AP *Drosophila* factor 300

Table A.2 Amino acid sequence and selected binding sites *Drosophila* homeodomains. 323

Table A.3 Master alignment of selected binding sites. 362

Table A.4 Mutual Information product comparison. 378

Table A.5 Fly factors used to predict the human homeodomains and confidence scores. 388

List of Figures

Figure 1.1 Transcription factors involved in A-P patterning. TFs involved in A-P patterning function in a hierarchical network to subdivide the embryo into 14 segments. 31

Figure 1.2 Schematic of the CRM and the transcription factors that controls the second stripe of eve's early expression. 34

Figure 1.3 Schematic of eve's early expression regulated by the 3+7 and 4+6 CRMs. 38

Figure 1.4 Schematic of bacterial one and two-hybrid systems. 50

Figure 2.1 Overview of the omega-based B1H system. 61

Figure 2.2 PCR products of rpoZ locus. 65

Figure 2.3 Maps of the bait plasmids used in the omega-B1H system. 67

Figure 2.4 Comparison of the activity of three different factors at three different promoter strengths. 70

Figure 2.5 Western blots of 3 ω-TF's at 3 different promoter strengths. 72

Figure 2.6 Comparison of ω-Zif268 activity with 4 different binding sites. 74

Figure 2.7 The omega subunit and the positions utilized by omega-TF hybrids. 78

Figure 2.8 Comparison of Zif268, Mig1 and Rap1 logos. 80

Figure 2.9 The ZFHD1 model. 83

Figure 3.1 Transcription factors involved in A-P patterning. 115

Figure 3.2 Specificities of the 35 TF in the A-P regulatory pathway characterized by the omega-B1H system. 121

Figure 3.3 Comparison of B1H-generated recognition motifs to previously published data. 123

Figure 3.4 Motifcount plots over 48 CRMs from *D. melanogaster* that drive early patterned expression. 127

Figure 3.5 Motifcount plots for the DNase and B1H PMWs over syntenic regions for the 48 CRMs within the *D. pseudoobscura* and *D. mojavensis* genomes. 130

Figure 3.6 Genome Surveyor display interface. 135

Figure 3.7 Hits two through five from the anterior TF search. 142

Figure 4.1 DNA recognition by the homeodomain family. 160

Figure 4.2 The ZFHD1 framework for determining DNA-binding specificity. 165

Figure 4.3 Sequence logos for 84 homeodomains determined in this study. 167

Figure 4.4 ClustalW alignment of the homeodomain motifs from the 84 homeodomains examined in this study. 170

Figure 4.5 Binding site logos for the eight *D. melanogaster* HOX factors displayed in order of anterior to posterior expression. 173

Figure 4.6 Competitive gel shift assays for nine different fly homeodomains representing nine different specificity groups. 175

Figure 4.7 Determining the orientation of the homeodomain on a selected site. 178

Figure 4.8 Specificity-based clustering of the 84 Drosophila homeodomains. 184

Figure 4.9 Secondary effects influence recognition for Thr47 containing homeodomains. 188

Figure 4.10 Residues at positions 6 and 7 of the N-terminal arm can influence the specificity at position 1 of the binding site. 192

Figure 4.11 Correlations are observed within the atypical homeodomains between the presence of Arg at positions 54 and 55 and the preferred specificity at binding site positions 2 and 4. 195

Figure 4.12 The role of position 8 in organizing the N-terminal arm. 199

Figure 4.13 Glu59 may influence the specificity imparted by Arg55 in Caup. 202

Figure 4.14 Heat plot of Mutual Information (MI) analysis between the amino acid sequences of the homeodomain and the master alignment of their selected binding sites. 205

Figure 4.15 Heat plot of the joint-ranked Mutual Information (jrMI) analysis. 207

Figure 4.16 Catalog of common specificity determinants for Asn51-containing homeodomains. 210

Figure 4.17 Exploring DNA-binding specificity through mutagenesis. 215

Figure 4.18 Binding site selection data for Bicoid collected from a selection performed at 5mM 3-AT instead of the typical stringency of 10 mM. 217

Figure 4.19 Clustergram of engineered En homeodomain that has TGIF-like specificity (**En_2_TGIF**; R3K, I47N, Q50A, A54R, K55R) with the entire set of fly homeodomains. 223

Figure 4.20 ClustalW alignment of the *D. melanogaster* and *H. sapiens* Iroquois family members. 225

Figure 4.21 Sequence logos of the 153 predicted human homeodomains. 229

Figure 4.22 Comparison of the predicted and actual recognition motifs for 6 human homeodomains. 234

Figure 4.23 Splash page of interactive web-based homeodomain specificity prediction tool at http://ural.wustl.edu/flyhd/. 238

Figure 4.24 Conversion of Engrailed into a factor with an NK-2 type specificity. 248

PREFACE

The work reported in this dissertation has been published in the following articles. Significant contributions of coauthors are cited below.

Noyes MB, Meng, X. Wakabayashi A, Sinha S, Brodsky M, and Wolfe S. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Research*, **36**, 2547-60.

A significant contribution was made to this publication by Saurabh Sinha. He created the Gbrowse tool, the "genome surveyor" that allows for the combinatorial prediction of CRMs using combinations of multiple factor specificities within our dataset.

Xiangdong Meng created the previous, alpha-based version of the bacterial one-hybrid system which this new system is based upon. Several of the constructs used in this work were modified versions of those developed in the alpha system. In addition, he built the 28 base pair library used in this article.

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, and Wolfe SA (2008) Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites. *Cell*, **133**, 1277-89.

Ryan Christensen in Gary Stromo's lab made a great contribution to this article by providing the cluster analysis of the homeodomain specificities, mutual information analysis and the creation of the homeodomain web page for the prediction of DNA-binding specificity.

CHAPTER I: INTRODUCTION

Summary

From the yeast genome completed in 1997 to the 12 *Drosophila* genomes published earlier this year; little more than a decade has provided an incredible amount of genomic data (Cherry et al., 1997; Drosophila 12 Genomes Consortium, 2008). Yet even with this mountain of genetic information the regulatory networks that control gene expression remain relatively undefined. In part, this is due to the enormous amount of non-coding DNA, over 98% of the human genome, which needs to be made sense of (Lander et al, 2001). It is also due to the large number of transcription factors, potentially over 2,000 such human factors, which may contribute to any given network directly or indirectly (Wilson et al., 2008; Tupler et al., 2001). Certainly, one of the central limitations has been the paucity of transcription factor specificity data that would aid in the prediction of regulatory targets throughout a genome. The general lack of specificity data has not only hindered the prediction of regulatory targets for an individual transcription factor but also how it may function within a set of complimentary factors. A large collection of factor specificities would allow for the combinatorial prediction of gene regulation that considers all factors actively expressed in a given cell, under a given condition. Furthermore, the amount of effort required for the characterization of a transcription factor's specificity, at least until recently, has resulted in the sporadic characterization of factors with mostly coincidental common themes or similarities. A comprehensive catalog of specificities for a given DNA-binding domain family from a metazoan would provide a detailed understanding of that domain's interaction at the protein-DNA interface. Such a catalog might allow the prediction of DNA-binding specificity for like factors in other genomes which may alleviate or limit the need to characterize every factor in every genome.

Transcription

To understand the regulation of transcription it must first be understood how transcription works and at what points might its mechanism be encouraged or retarded. A very basic understanding of transcription begins with RNA polymerase being able to find and bind the core promoter of a specific gene. Binding the DNA is followed by separation of the strands of a short sequence of DNA where one of these strands will be used as the template for the polymerase to copy. The actual duplication of this "template strand" is accomplished by base pairing with complimentary ribonucleotides, which is catalyzed by the polymerase, adding to the growing RNA molecule one nucleotide at a time. Though the polymerase may struggle initially, once the RNA molecule reaches beyond roughly 10 bases the reaction becomes favorable. At this point the polymerization becomes more efficient and the polymerase clears the promoter transitioning to the elongation stage (Mooney and Landick, 1999; Cramer 2004). There are many influences that can promote this transition from initiation to elongation and those that stabilize and enhance elongation itself. For example, in the development of the *Drosophila* embryo it appears that RNA polymerase may be loaded but stalled at many promoters and it is the repression of elongation that is regulating transcription (Wang et al., 2007). However, for the purposes of this discussion we will focus primarily on the first step where much of regulation occurs; how the polymerase gets to a specific promoter to initiate transcription.

In prokaryotes, RNA polymerase is able to engage the promoter and initiate transcription with only a small number of essential subunits (Borukhov and Nudler, 2003). The core polymerase, which is able to catalyze the polymerization reaction, is made up of two α subunits, the β and β' subunits (Vassylyev et al, 2002; Sweester, et al, 1987). However, the core polymerase is capable of initiating transcription from essentially any sequence. It is the addition of a fifth subunit, the σ subunit, that provides the promoter specificity that allows for initiation to begin at an appropriate sequence (Gross et al, 1998). Though there are several σ subunits, most genes are transcribed with σ_{70} during exponential growth and σ_{38} while in stationary phase (Lonetta et al, 1992). These five subunits comprise the minimal holoenzyme that is capable of initiating transcription from a given promoter though there are several other nonessential subunits that play varying roles. Essentially, the two factors influencing prokaryotic transcription are the local concentration of the polymerase near a promoter and strength of the interaction between the holoenzyme and the specific promoter sequence.

Eukaryotic transcription is quite complex in comparison to the prokaryotic system. There are three eukaryotic polymerases that transcribe DNA into RNA, each with a separate purpose. RNA polymerase II is responsible for the transcription of protein coding genes and will therefore be the focus of this discussion. Like its prokaryotic counterpart, the core 12 subunit RNA polymerase II is able to initiate transcription on special transcripts *in vitro* but is not able to recognize a promoter sequence (Armache et al., 2003). Several proteins, the general transcription factors (GTFs), must assemble at the promoter before transcription is initiated (Green 2005; Roeder 1980). These GTFs include RNA polymerase II and the auxiliary protein complexes TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. The first to assemble is TFIID which contains the TATA-binding protein (TBP) and the TBP-associated factors (TAFs) (Maldonado et al, 1990; Smale and Kadonaga, 2003). TBP has been shown to bind *in vitro* to the core A-T rich sequence of the promoter, the TATA box, creating a major distortion in the DNA, around which the remaining GTFs assemble (Patikoglou et al. 1999). TFIID binding is assisted by the binding of TFIIA and TFIIB (Maldonado et al, 1990). Once this complex is formed, RNA polymerase II along with TFIIF assembles followed by TFIIE and TFIIH. TFIIH contains a helicase that melts the promoter which not only marks the beginning of initiation but creates a much more stable open complex of polymerase and the DNA.

In addition to the TATA-box, there are other conserved sequence elements in eukaryotic promoters that subunits of RNA polymerase II interact with. In fact, the TATA-box is found in only 20% of eukaryotic promoters (Bucher, 1990). It appears that at many of these TATA-less promoters TBP does not make base specifying contacts (Coleman and Pugh, 1995). Rather, TBP simply acts as a scaffold for other proteins to assemble upon, the first of which may be tissue specific TAFs, thereby providing the first example of specific regulation of transcription. Other elements that encourage transcription initiation are the TFIIB regulatory element (BRE), the initiator sequence (Inr) and the downstream promoter element (DPE) (Smale and Kadonaga, 2003). The BRE is the only element recognized by a GTF other than TFIID. TFIIB will bind cooperatively with TBP at the BRE when a TATA-box is present and appears to provide directionality to the TBP (Tsai and Sigler, 2000). Inr is the most conserved of these elements with roughly 70% of the genes in *Drosophila* containing such a sequence (Smale and Kadonaga, 2003). Typically 30 base pairs downstream of where the TATA box would be, this Inr marks the spot where initiation will begin and is recognized by TAFs. Finally, found in roughly 40% of eukaryotic promoters the TAF recognized DPE functions cooperatively with the Inr and is usually found in promoters that lack a TATA-box. Any given promoter may have a unique combination of these sequences as well as their relative match to their consensus. Therefore, the unique combination of promoter elements and available TAFs in a given cell, is the first level of sequence specific regulation of transcription.

The core RNA polymerase II and the GTFs are sufficient to initiate transcription *in vitro*. However, *in vivo* an approximately 25-30 subunit complex, the Mediator, also assembles with the GTFs (Conaway et al., 2005). The subunits of the Mediator, thought to be the target of many transcription activators, acts by integrating the signals of these trans-activating factors and stimulating the assembly of the GTFs (Maston et al, 2006). This understates the importance of the Mediator. For example, while TFIID is only required for activation from 16% of yeast promoters, the Mediator protein Srb4 is required in 93% (Holstege et al, 1999). It is clear that subunits of the Mediator are absolutely necessary for transcription at most promoters *in vivo* therefore it is now commonly considered an additional GTF. Together these subunits, the core RNA polymerase II, the additional GTFs, and the Mediator form the preinitiation complex (PIC) at the promoter where TFIIH will separate the stands of DNA to initiate transcription.

Regulation

The first step in the regulation of transcription is to regulate the ability of the PIC to assemble at a given promoter. Therefore, the most straightforward way to regulate transcription might be to simply control the availability of a given promoter. In fact, the most common mode of gene regulation in prokaryotes is for a repressor protein to bind the promoter region of a target gene and by doing so make it inaccessible to RNA polymerase (Thiel et al, 2004). Of course, this also implies that the natural state of a given promoter in a prokaryote is active, which is not all too surprising considering the relatively minimal polymerase components that are required for transcription. However this is not necessarily the case. Activation in prokaryotes often requires recruitment of the holoenzyme to the target promoter by a transactivating factor.

Activation of transcription by recruitment of the holoenzyme is accomplished through the binding of the transcriptional activator to a specified sequence that neighbors its target promoter. An additional domain will contact a subunit of the holoenzyme and either recruit the enzyme to the promoter or stimulate the transition to an open complex and promoter clearance. Therefore, stimulation of transcription is mediated by a protein-protein interaction between the transcription factor and a subunit of the polymerase. The contact is often made with C-terminal domain of the α subunit (Hochschild and Dove, 1998). This may be due to the presence of two α subunits, the accessibility of the C-terminal domain or simply the DNA-binding properties of the α subunit. Stabilizing the interaction between the α subunit and the DNA might be an efficient way for a transcription factor to stabilize the polymerase-DNA interaction and therefore stimulate transcription. Moreover, the σ subunit which also binds DNA, is another common target of activators. However, it is certainly possible that different activators utilize interactions with any of the subunits of the holoenzyme or even a nonessential domain. In fact, a transcription factor expressed as a direct fusion to the nonessential ω subunit was able to stimulate transcription by 70 fold (Dove and Hochschild, 1998). Also, alternative protein-protein interactions can stimulate transcription if one protein is fused to a DNA-binding domain and the other is fused to either the α or ω subunits (Dove et al, 1997; Dove and Hochschild,1998). In this case, when the target sequence of the DNA-binding domain is placed upstream of the promoter, the alternative protein-protein interaction mediates the recruitment of RNA polymerase through the fusion to one of its subunits. Furthermore, the Hochschild lab also demonstrated that multiple contacts lead to a synergistic activation of transcription (Joung et al, 1993; Joung et al. 1994). First, they demonstrated that two factors that are able to contact different subunits of the polymerase will result in a much greater level of transcription than either interaction independently.

Secondly, they demonstrated that if a 2nd, artificial interaction is engineered into a single factor, this too results in a synergistic activation of transcription in comparison to the natural and artificial interaction alone. Therefore, transcription can be influenced by multiple factors that each make a different contact with the polymerase resulting in a synergistic activation of transcription.

In contrast to prokaryotic systems, eukaryotic promoters might be thought of as naturally inactive because of the inability of RNA polymerase II to transcribe without the assembly of many times more proteins and the organization of DNA in chromatin, making any given promoter at least to some degree inaccessible. One way in which promoter accessibility is controlled in chromatin is by influencing the strength of the DNA-chromatin interaction. The basic subunit of chromatin is the histone, the core of which is composed 2 sets of the H2A, H2B, H3, and H4 subunits (Widom J, 1998; Woodhead and Johns, 1976). Roughly 150 base pairs wrap around each histone which are linked by short sequences of DNA. The N-terminal tails of the histone domains are the targets of the chromatin modifying enzymes. The strength of the DNA-histone interaction can be influenced by acetylation or methylation of lysines, and sometimes arginines, of these N-terminal tails (Wade et al., 1997; Ng and Bird, 2000; Martinowich et al. 2003). For example, acetylation decreases the net positive charge of the histone and therefore weakens the interaction with the negatively charged DNA and making it more accessible (Wade et al, 1997).

Inducing higher order structure in chromatin, such as the densely packed heterochromatin, will also have a large impact on promoter accessibility. This is primarily induced by methylation of the lysine 9 residue of the H3 subunit (Rea et al., 2000; Jenuwein T, 2001; Grewal and Elgin, 2002). This methylated residue is the target of heterochromatin protein 1 (HP1) which can dimerize. It is the dimerization of the bound HP1 that leads to an expanding compaction of the DNA (Nielsen et al., 2001; Grewal and Elgin, 2002). Therefore it is the extent of methylation at lysine 9 and the availability of HP1 that controls the transition to heterochromatin.

Based on the concept of an inaccessible promoter, it might seem likely that activation of transcription is as simple as remodeling the chromatin. However activation is not that simple. There are several lines of evidence that argue for the required recruitment of factors that will stimulate the assembly of the PIC for transcription to be activated (Kuras and Struhl, 1999; Ptashne and Gann, 1997). Perhaps the most convincing work came out of the Ptashne lab in the late 1980's and 1990's. The Ptashne lab eloquently demonstrated the fundamental concept of recruitment and its contribution to activation of transcription. They were able to demonstrate that a DNA-binding and activating domain were sufficient to stimulate transcription at a specified promoter but only in the presence of that DNA-binding domain's target sequence (Giniger et al, 1985; Brent and Ptashne, 1985; Ma and Ptashne, 1987). Furthermore, if the DNA-binding domain is fused directly to a component of the PIC, transcription at a specific promoter can be activated when its target sequence has been installed upstream of that promoter (Barberis, et al, 1995). By comparison, when the same type of fusion is made with a chromatin modifying protein, transcription is only weakly activated (Georgakopoulous et al, 1995). This demonstrates that recruitment of the PIC to a promoter appears sufficient to activate transcription while localized chromatin remodeling is not.

Through these and other studies, the general mode of transcriptional activation has been relatively well outlined for some time now. The DNA-binding domain of an activator binds its target sequence which is within some proximity of the promoter it regulates. In yeast, this distance is typically within a few hundred base pairs (Levine and Tjian, 2003). In humans, this distance can span tens of thousands of base pairs. No matter the distance, once bound to its target sequence the activator uses an additional domain, the activation domain (AD), to contact some component of the PIC. The contact between the activator and the component of the PIC stimulates the assembly of the PIC at the target promoter. In other words, through its contact with a PIC component, the activator recruits RNA polymerase II to a specific promoter and thus stimulates transcription of a specific gene. Ptashne demonstrated that the Gal4 activation domain was able to interact specifically with a mutated form of the Mediator subunit Gal11 (Gal11P) to stimulate PIC assembly and transcription at a specified promoter (Barberis et al, 1995). It has since been shown that Gal4 naturally interacts with Tra1, a component of the SAGA complex (Bhaumik et al., 2004). SAGA in turn interacts with the Mediator and leads to the TAF independent recruitment of TBP to the promoter. Both contacts demonstrate interactions with the Mediator, direct and indirect, that lead to assembly of the PIC. Of course, there are activators that influence chromatin structure or regulate elongation such as in the *Drosophila* embryo but stimulating PIC assembly appears to be a common mode of site directed transcriptional activation and certainly the most documented.

DNA-binding domains

Overview

Trans acting activators and repressors are referred to as sequence specific transcription factors because their effect on transcription is dependant on the availability of a specific sequence of DNA. These transcription factors are able to recognize their genomic targets by utilizing a specialized subdomain, their DNA-binding domain. These domains are able to interact with a small degenerative set of DNA sequences which at least partially determines which genomic sequences it interacts with and therefore, which gene's transcription the factor may regulate. Several different strategies for such protein-DNA interactions have evolved but the most common theme is the DNA-binding domain is able to position an alpha helix into the major groove where the side chains of its amino acids can make direct contact with the base pairs. However, this is an over simplification of the interaction at the protein-DNA interface. The local sequence of DNA influences the depth, width and flexibility of the double helix (Klug A, 1995). A given domain may not only recognize a pattern of chemical groups in the major or minor groove but also the unique shape and local environment which that sequence provides. Therefore, these secondary influences on specificity may play a powerful role in a domains ability to discriminate one sequence from another.

A great number of domains have evolved for the purpose of specifying DNA sequences but most transcription factors utilize one of a small number of domain families. In fact, 84% of the transcription factors in the human genome use one of only the 7 different domains that follow (Wilson et al., 2008; Tupler et al., 2001). Though all of these domains bring a unique element to specificity determination, they can be broadly clustered in two catagories: those that are able to bind DNA independently and those that require dimerization. Below, a brief description of the 7 most common DBDs in the human genome is provided as well as their relative abundance.

Monomers

These classes of DBDs are considered monomers because not only are they able to bind DNA independently, but they are able to carry out their mode of action independently. Whether the TF's function is to recruit other TFs, stimulate PIC formation, or modify chromatin, these factors are able to do that as a single protein. In contrast, there is some evidence that dimers may initially bind DNA as monomers but they require dimerization to carry out the function of the TF that encodes them. Of course, the DBD families listed in this category might also function as dimers. The most obvious example is the homeodomain that is often thought of as forming a dimer to provide extended specificity and therefore determine which genomic targets will be regulated (Ryoo and Mann, 1999; Joshi et al., 2007). However, this is primarily based on evidence that the HOX factors are able to dimerize with extradenticle and perhaps other homeodomains that fall into the atypical subfamily of this domain. This dimerization does indeed extend and modify the specificity of the HOX monomer. Still, outside of the HOX factors there is very little evidence that homeodomains function as dimers and there are several examples of monomeric regulation even by members of the HOX family (see discussion Chapter 4). An interesting wrinkle in the dimerization discussion is the ability of the winged helix DBD families, the Forkhead and Ets families, to form heterotypic dimers, which are dimers formed between two factors from different DBD families (Amoutzias et al, 2008). Both the Forkhead and Ets families have the ability to dimerize with hormone receptors and some homeodomains. Again, these factors are indeed able to bind DNA and function independently though dimerization, both homotypic and heterotypic, may be an important component of their regulatory repertoire. For the purposes of this discussion I will focus on their monomeric binding.

Cys_2His_2 Zinc fingers

When only considering the longest transcripts for each coding sequence and therefore ignoring any relevant splice variants, there are approximately 750 transcription factors in the human genome that utilize Cys_2His_2 zinc fingers making this by far the most common DBD (Wilson et al, 2008; Tupler et al, 2001). These factors represent roughly 50% of the sequence specific transcription factors in the human genome and regulate a diverse set of cellular functions. This DBD was first noticed as 9 tandem repeats in the coding sequence of TFIIIA of *Xenopus* (Brown et al, 1985; Miller et al, 1985) and as more and more Cys_2His_2 zinc fingers have been characterized, their alignments have shown that the domain can be described by the consensus sequence: (F/Y)-X-C-X_{2-5}-C-X_3-(F/Y)-X_5-□-X_2-H-X_{3-5}-H, where X can be any residue and □ is a hydrophobic residue. The domain utilizes tetrahedral zinc coordination by its two cysteines and two histidines to stably fold the polypeptide into a $\beta\beta\alpha$ structure (Pavletich and Pabo, 1991; Elrod-Erickson et al, 1996). This zinc coordination provides stability to these roughly 30 amino acid domains that are not large enough to form a stable hydrophobic core and therefore are not able to fold in the absence of the zinc ion.

Zinc finger transcription factors typically contain multiple fingers that each bind adjacent 3bp subsites (Wolfe et al, 2000; Elrod-Erickson et al, 1996). Hypothetically, this means that a two fingered protein will bind a 6bp site and each additional finger will add 3 bases of specificity. The base specifying contacts of each finger are then made with the DNA by positioning the helix in the major groove where positions -1, 2, 3, and 6 of the helix will most often provide the specificity for each finger (Elrod-Erickson et al, 1996; Luscombe et al, 2000). These factors will typically function as monomers and can bind with sufficient specificity and affinity to regulated sets of genes independently of other factors. In fact, three finger proteins will often bind with nM to pM affinities (Elrod-Erickson and Pabo, 1999; Greismann and Pabo, 1997) and provide excellent discrimination between target sequences.

In practice, zinc finger binding is not so straight forward. These factors in the human genome have an average of 12 fingers per protein (Wilson et al, 2008; Huntley et al, 2006). If each finger were to bind as proposed, the average transcription factor would be specifying 36 bases. But the proposition that each additional finger will bind an adjacent 3bp sequence assumes that those fingers are canonically linked and function cooperatively. That is to say that the linker between two fingers is 5 amino acids from the last histidine of one finger to the first hydrophobic residue of the following finger, the linker sequence is similar to TGEKP, and the binding of each finger is at least partially dependent on the binding of the other. This may be the case for a significant portion of fingers but certainly not all of them. Out of the over 10,000 estimated individual zinc fingers coded by the human genome, approximately 4,200 of them are canonically linked (Letunic et al, 2006). Many more fingers have short linkers, but noncanonical in sequence, that are likely restricted in their flexibility. Still, many other fingers have large insertions between them. Since a given transcription factor will often have all of these various linker types, its possible that these large insertions are acting as spacers or even barriers that separate the functions of sets of canonically linked fingers within the same protein. Each set may act as a functional unit and therefore provide diversity in the functions that a single protein might perform.

Homeodomains

In humans, as well as many other metazoans, homeodomains comprise the second largest class of sequence-specific transcription factors with an estimated 233 such factors

(Wilson et al, 2008; Tupler et al, 2001). Homeodomains were first identified as Homeotic genes in *D. melanogaster* where the altered activity of a particular gene can lead to the morphological transformation of one segment into that of its neighbor, leading to dramatic phenotypes such as the appearance of a second pair of wings (Lewis, 1978) . Cloning of these genes and analysis of their products led to the observation that they contain a common sequence motif that encodes a DNA binding domain (Gehring et al., 1994a).

The homeodomain consists of approximately 60 amino acids that fold into a stable 3-helix bundle preceded by a flexible N-terminal arm. Interactions with a 4 to 7 basepair DNA binding site are formed by positioning the third helix, the recognition helix, in the major groove and the N-terminal arm in the minor groove. Despite a common DNA binding architecture, there is significant variation in the sequence composition within the homeodomain family; for example the two superclasses of homeodomains, denoted as typical and atypical (Banerjee-Basu and Baxevanis, 2001; Gehring et al., 1994a; Mukherjee and Burglin, 2007), share low sequence identity and generally recognize substantially different DNA sequences. Nonetheless, the docking of typical and atypical homeodomains with the DNA is nearly identical (Kissinger et al., 1990; Wolberger et al., 1991), likely facilitated by common sets of contacts to the phosphodiester backbone.

Though the vast majority of homeodomains appear to function as monomers, there are several examples of dimerization between members of the Hox and TALE families of homeodomains (Ryoo and Mann, 1999; Joshi et al., 2007). This dimerization appears to be mediated by a YPWM motif that is encoded N-terminal to the homeodomain in the Hox factor and the three amino acid loop extension (TALE) that connects helix 1 and helix 2 of the TALE factor. The dimerization of these homeodomains not only increases the size of the core recognition sequence to eight or more bases, but it also modifies the specificity contributed by each of the individual homeodomains. However, it must be noted there is also evidence that these same homeodomains are able to function and regulate transcription as homo-oligomers (Galant et al, 2002; Lohmann et al, 2002).

Forkhead domains

Forkhead domains represent the fifth most common family of DBDs in the human genome with 49 representative members and they are the first of two families that utilize a "winged helix" motif to bind DNA (Tupler et al, 2001). The common architecture of this domain includes 3 β-sheets, 3 α-helices, and two loops or "wings" in the following order: α1 - β1 - α2 - α3 - β2-W1- β3-W2. The 2nd and 3rd helix form a helix-turn-helix motif that positions the 3rd helix in the major groove where the base specifying contacts are made (Luscombe et al, 2000). However, this 3rd helix is highly conserved which may contribute to the fact that all Forkhead domains characterized to date bind a core sequence of (A/C)AA(C/T)A sequence element (Wijchers et al, 2006). This 5bp sequence is required but not sufficient for binding which implies there are other base specifying contacts. These additional contacts are thought to come from the less conserved turn immediately N-terminal to the 3rd helix and from contacts made in the minor groove by the second wing. These additional contacts may lead to the extended specificities that are required for a given subfamily of Forkhead domains. For example, the FoxO family specifies a (T/C)(G/A)AAACAA sequences (core sequence element is underlined).

For the most part Forkhead proteins function as monomers; however, there are exceptions. The FoxP subfamily requires dimerization to function. Also, as previously noted, the Forkhead family can form heterotypic dimers with other DBD families. For example, members of the FoxO family are able to interact with the androgen, glucocorticoid, and retinoic acid receptors. Foxa2 has been shown to interact with several homeodomains including PITX2, HOXA10, and engrailed (Foucher et al, 2003; Marshak et al, 2000). It is not clear how these heterotypic interactions are forming and their influences on function are inconsistent. Some interactions enhance while others interfere with the regulation of the binding partner's target. It is also unclear whether these interactions modify the specificity of either DBD and thus revise the regulatory targets of either domain, or if they alter the regulatory activity of their partner at their common set of targets.

Ets domains

Ets domains were first identified by homology to the *v-ets* oncogene from the E26 avian erythroblastosis virus (**E**-**T**wenty-**S**ix) (Sharrocks et al, 1997). There are 28 members in the human genome, representing the 7th largest family (Wilson et al, 2008). Structurally they are very similar to the Forkhead domain in that both domains utilize a winged helix motif. In fact, the only major difference appears to be that most Ets domains only have a single wing compared to the common two wings of the Forkhead domain (Luscombe et al, 2000). The main difference appears to lie in the specificity. While the Forkhead domain requires a (A/C)AA(C/T)A core sequence for DNA binding, the Ets domain requires a GGA(A/T) motif (Oikawa and Yamada, 2002). These differences in core specificity appear to be determined by different residues on the 3rd helix, the recognition helix. Also, like the Forkhead domain, the Ets domain requires subfamily dependent sequences that flank the core motif binding. These typically result in 7-9bp elements and are thought to be specified by the residues adjacent to the recognition helix.

Also like the Forkhead domain, Ets domains primarily function as monomers though recent structural data indicates it is possible for some members to homodimerize. However, a more striking similarity is that members of both of these domains are able to form heterotypic dimers with both hormone receptors and homeodomains. There are a handful of examples for each partnering domain but a very interesting example is demonstrated between the Ets-1 factor and the vitamin D receptor (VDR) (Tolon et al, 2000). Like most hormone receptors, ligand binding to VDR (in this case vitamin D) induces a conformational change that allows the receptor to function as a transcription factor. Remarkably, the interaction between Ets-1 and VDR appears to produce the same conformational change, relieving the receptor of its vitamin D dependance and stimulating its regulatory activity. The Forkhead and Ets domains seem so similar in structure and function that one might think of them as two superclasses of one master DNA-binding domain family, the winged helix family, but thus far their classification has remained separate.

Dimers

Dimeric DNA-binding domains add an additional level of complexity to gene regulation. Because one monomer can potentially bind with several partners, that same monomer is able to regulate different genes and networks. Functionally, a specific partner might modify the DNA-binding specificity of the dimer and therefore lead to the regulation of a unique set of genomic targets (Ryoo and Mann, 1999; Amoutzias et al, 2008). On the other hand, different partners might simply provide different auxiliary domains, leading to different outcomes at the same group of targets. For example, the basic helix-loop-helix domain Max can dimerize with either Myc or Mad. However, Myc and Mad are not able to dimerize with one another (Luscher, 2001). If Max is bound to Myc it recruits SWI/SNF, a histone acetyl transferase, to its genomic targets. If Max is bound to Mad it recruits histone deacetylases, leading to the exact opposite result as the Myc-Max dimer at a targeted promoter (Grandori et al, 2000). Therefore, the ability of a dimer to regulate a target sequence is dependant on additional influences: the concentration of both monomers, the affinity the monomers have for one another, and the affinity that the dimer then has for the target sequence. These influences provide additional interactions that can be regulated and therefore provided tighter control than might be possible with monomeric factors.

Another advantage of dimeric transcription factors is that they provide potential diversity in their targets with a minimal number of parts. For example, if we assume that all monomers of a given DBD family are able to form functional dimers, the 51 basic leucine zipper (bZip) domains in the human genome (Wilson et al, 2008) would provide 1,326 dimer combinations, close to 150% of all the Cys_2His_2 zinc finger proteins. However, it is estimated that there are actually only 350 unique dimers in the human genome that are functional because of the dimerization specificity between bZip monomers (Grigoryan and Keating, 2006; Newman and Keating, 2003). In fact, dimeric DBD families often have “hub” proteins that are able to dimerize with many partners while the peripheral monomers bind with relatively few. For example, the E2A bHLH domain and the RXR hormone receptor are predicted to have 38 and 24 partners, respectively (Armoutzias et al, 2004; Amoutzias et al, 2007). The dimerization specificity obviously limits the number of potential combinations but still, 350 unique bZip combinations is 50% more than all of the monomeric homeodomains and it implies that each bZip monomer has on average 7 different dimer combinations. Below, the DNA binding properties of the three most common dimeric DBD families are briefly described. When we consider all the homodimeric and heterodimeric combinations of these three domains and then add into the equation that some factors, such as Ets factors, are able to form heterotypic multimers with other DBD families, it becomes clear that the protein-protein interactions between site-specific transcription factors are every bit as important as the protein-DNA interactions that they govern.

Basic Leucine Zippers

The basic Leucine Zipper (bZip) consists of a single α-helix, roughly 60 amino acids long (Luscombe et al, 2000). The helix is made of two parts; the C-terminal dimerization domain and the N-terminal basic region . The 30 amino acid C-terminal domains have a leucine, or similar residue, positioned roughly every 7 amino acids or two turns of the helix. These hydrophobic side chains from each helix utilize hydrophobic contacts to pack together side by side and “zipper” the helices together, forming a coiled-coil. The leucine zipper is a very common dimerization domain utilized by DBDs. Subfamilies of the bHLH family use one and sometimes two leucine zippers to specify potential dimerization partners (Amoutzias et al, 2008). There are also a few examples of zinc fingers and homeodomains that utilize Leucine zippers for dimerization.

In the context of the basic leucine zipper, the basic region provides the DNA-binding specificity. The basic region is simply an extension of the helix that is positioned into the major groove where its side chains make base specifying contacts with ½ of the dimer’s target sequence. However, it must be noted the basic region is not ordered in the absence of DNA. It is presently unclear whether the mode of action is for the dimers to form independently and the basic regions of both monomers will then be ordered once in contact with the entire target sequence or if the monomers bind DNA independently and dimerize when they come in contact with a potential partner (Kohler and Schepartz, 2000). It is known that dimers are able to form in the absence of DNA. However, there is evidence that the presence of DNA speeds up the dimerization process implying that by binding the DNA first, monomers are better able to dimerize.

Basic Helix-Loop-Helix

The basic helix-loop-helix (bHLH) domain is the third most common DBD family in the human genome with 118 members (Wilson et al, 2008; Tupler et al, 2001). The bHLH domain dimerizes by forming a 4 helix bundle on the DNA where each monomer is contributing a helix-loop-helix (Amoutzias et al, 2008). The basic region is an extension of the first helix of the helix-loop-helix motif and, like the bZip domain, is disordered in solution but forms an alpha helix when contacting DNA (Jones, 2004). Contacts are made between this basic alpha helix and bases in the major groove. In contrast to the bZip domain, the loop between helices in the bHLH domain provides more flexibility in positioning the helices in the major groove. The resulting 4 helix bundle is then centered on the E-box sequence CANNTG.

Hormone Receptors

The hormone receptor family of DBD's function by binding their ligand in the cytoplasm and then translocating to the nuclease where they regulate a subset of genes in response to this environmental cue. The vast majority of this domain family members function as a dimer with each monomer having a ligand binding, DNA-binding, and transcriptional regulatory domain (Amoutzias et al, 2007). The canonical domain architecture is loop-helix-loop-helix, the fold of which is stabilized by the coordination of two zinc ions, one by the 4 cysteines in each loop leading to its designation as a C_4 zinc finger (Luscombe et al, 2000). Contact is made with the DNA by both of the helices. The first helix provides the base specifying contacts in the major groove. The second helix positions itself perpendicular to the first helix and makes nonspecific contacts with the DNA backbone. Dimerization is mediated by the second loop which positions the two monomers so that each of their six base pair half sites are separated by 3-6 base pairs. One exception to the hormone receptor family is the knirps, and knirps-like, protein. These factors function as monomers utilizing the same C_4 zinc finger architecture but instead of relying on dimerization to provide additional specificity and/or affinity, knirps contains an additional DBD known as the "knirps box" (Rothe et al, 1989). It is not clear how the knirps box makes contact with the DNA but truncation analysis has demonstrated that it is necessary for DNA binding. Because of the lack of three dimensional structure or quality specificity data for the knirps protein, it is difficult to say whether the knirps box provides specificity, affinity, or both.

Regulatory Networks

A single transcription factor is able to influence transcription of its regulatory targets by binding to a sequence or set of sequences in a genome and influencing the transcription of neighboring genes, positively or negatively as previously detailed. However, in higher eukaryotes genes regulated by a single protein are less common (Spiegelman and Heinrich, 2004; Harbison et al., 2004; Kim and Kim, 2006). Rather, the transcription of a given gene is the result, or the sum, of all the combinatorial inputs of the factors that are available and able to influence initiation of transcription at that gene's promoter. Therefore, transcription may be the result of a network of regulatory influences rather than a one to one relationship between a transcription factor and its target. Since each transcription factor holds a different piece of information (e.g. Is a hormone present? Is the cell dividing or stationary?) the cell, or more appropriately a given promoter, must be able to decode this information in order to determine what is the proper outcome. The units that decipher all of this information are the DNA sequence elements that all the relevant transcription factors are able to bind to and that neighbor a given gene's promoter (Davidson, 2001). These sequence elements are most often thought of as enhancers but might include enhancers, silencers or insulators. Together, these elements are referred to as "cis-regulatory modules" (CRMs). Each CRM is typically greater than 300bp in length and contains roughly 10 binding sites for four or more transcription factors (Levine and Tjian, 2003). However, because the promoters each CRM regulates can be tens of thousands of base pairs away, and not always the nearest promoter, it is rather difficult to pinpoint what sequence elements and promoters are functional pairs.

Yeast Regulatory Networks

Since the sequencing of the yeast genome (Cherry et al., 1997), an enormous effort has been made to map the regulatory networks in this organism, providing a great deal of what is now understood about these systems. This effort has been aided by several advantages provided by this organism. First, it is relatively straight forward to site specifically modify a genomic sequence in yeast by homologous recombination. Second, there are few introns and the intergenic sequence space is compact in comparison to higher eukaryotes such as humans. Together, this allowed the Young lab to create knock-ins for most of the transcriptional regulators in yeast with the addition of a C-terminal epitope tag (Lee et al., 2002). This allowed for tagged versions of each regulator to be expressed under its natural physiological control. A tagged version of each regulator could then be chromatin-immunoprecipitated (ChIP) under rich growth conditions, or an alternative condition, to recover potentially all of the genomic regions that a given factor was able to bind to. DNA microarrays were used to determine which genomic sequences were enriched by a given factor. This approach was referred to as a genome-wide location analysis (Ren et al., 2000).

The first pass at a large scale genome-wide location analysis provided several insights into the fundamental workings of eukaryotic networks (Lee et al., 2002). Analysis of 106 yeast factors demonstrated an average factor bound 38 different promoter regions. Conversely, it was found that roughly 37% of the promoter regions were bound by at least one transcription factor. Furthermore, more than a third of these promoter were bound by two or more factors. Within this data set, a number of network motifs were described that form the building blocks of higher network structure. There are several examples of motifs such as autoregulation, feedforward loops, and single or multiple-input loops. However, this analysis was limited in that it provided which promoter a factor or set of factors might bind but not how many times they might bind or how multiple factors might influence the binding of one another within a given promoter.

A second pass by the Young lab included more factors (203), 84 of which were tested at multiple growth conditions, and a comparison of the phylogenetic conservation between related yeast species at an enriched genomic region to predict the binding site specificity for a given factor (Harbison, et al., 2004). This analysis allowed for the prediction of how many binding sites a given factor might have within a specific promoter. Promoters that contained single binding sites, repeating binding sites, and sites for multiple different factors were all common. Furthermore, multiple growth conditions provided insight into how sets of factors respond differently to external cues. For example, some factors were termed ‘condition-invariant’ because the same sets of promoters were recovered under multiple conditions. Others were only active once the conditions were modified, in other words they were ‘condition enabled’. Still, other sets demonstrated an expansion of the promoters they bound under new conditions. This global analysis of the yeast regulatory network provides a powerful foundation to expand upon in higher eukaryotes.

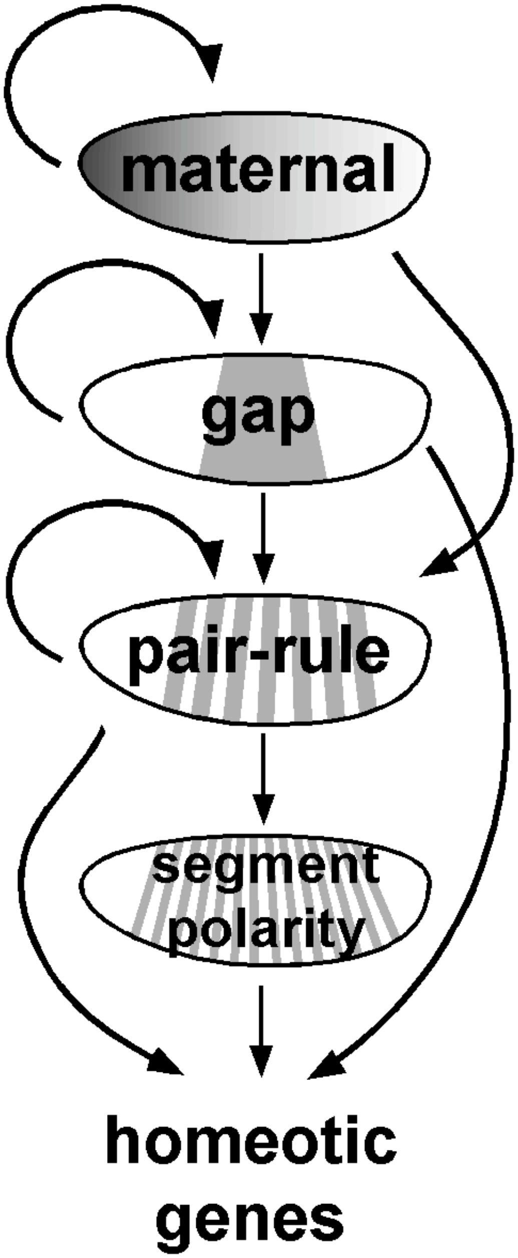
A Metazoan Development Network

Much of what we know about metazoan regulatory networks comes from decades of genetic studies that detail the development of the *Drosophila* body plan and the genes that control it (Carroll, 1990; Ingham, 1988). Remarkably, execution of the adult body plan begins extremely early in development where the fate of a given nucleus is determined by its position in the embryo and the patterned distribution of critical transcription factors (Gaul and Jackle, 1987; Ingham 1988; Small et al., 1991). Even before fertilization the body plan is being organized in the embryo by the distribution of maternal mRNA. These maternal morphogens, such as bicoid, are arranged in a broad gradient across the embryo. Upon fertilization, it is the concentration of these maternal factors and the number of binding sites that neighbor their regulatory targets that will determine which zygotic genes are initially expressed.

The *Drosophila* embryo has provided an excellent platform to investigate how various concentrations of multiple transcription factors and numbers of binding sites are processed to result in discrete outputs. This is at least partially because of the unique, rapid expansion of nuclei that takes place soon after fertilization. The syncytial blastoderm, also referred to as the pre-cellular blastoderm, that is formed contains thousands of nuclei across the embryo, none of which are separated by cell membranes. This provides for an almost seamless flow of information (the expression and availability of transcription factors) between nuclei across the embryo. It is therefore the discrete position of each nucleus within the embryo that determines the concentration of maternal factors that will influence the expression of a unique set of genes. Neighboring nuclei will then have similar expression patterns and therefore set up secondary gradients of transcription factor expression. It is this cascade of transcription that leads to the more and more refined regions with unique portfolios of gene expression and ultimately the morphology that will develop from that region of the embryo (Carroll, 1990; Arnosti, 2003; Arnosti et al., 1996).

The factors that determine the anterior-posterior segmentation of the *Drosophila* embryo form a hierarchy of groups of transcription factors; the Maternal, Gap, Pair-ruled and Segment polarity groups (St Johnston and Nusslein-Volhard, 1992; Schroeder et al., 2004) (Figure 1.1). Their function can be described as a regulatory cascade where the expression of one group regulates the expression of the proceeding group and so on (Peel et al., 2005; Pick, 1998). Activated by the unique, local concentration of Maternal factors, the first set of activated zygotic genes are the gap factors. These are expressed in broad bands and, in concert with the maternal factors they regulate the expression of the pair-ruled genes. Expressed in 7 stripes roughly perpendicular to the anterior-posterior axis, the pair-rule gene expression gives the first glimpse at the developing segmented body plan of the *Drosophila*. The pair-ruled factors then regulate 14 stripes of segment polarity gene expression, providing the foundation for the segmented adult body. Members of all of these groups may participate in the regulation of the downstream homeotic factors that influence the further development of each segment.

Figure 1.1.



maternal genes

bcd, cad, hb

gap genes

head:
btd, D, Gsc, oc, ems, Optix
trunk:
hb, Kr, Ttk, gt, nub, kni, cad
terminal:
tll, Hkb, fkh

pair-rule genes

h, slp1, prd, opa, odd,
Blimp-1, eve, ftz, run*

segment polarity genes

en, inv

homeotic genes

lab, Dfd, Scr, pb, Antp,
Ubx, Abd-A, Abd-B

Figure 1.1. Transcription factors involved in A-P patterning. TFs involved in A-P patterning function in a hierarchical network to subdivide the embryo into 14 segments. The early maternal factors are expressed in broad gradients, with subsequent TF groups expressed in patterns that are increasingly refined. TFs involved in early segmentation or expressed in early patterns that were characterized in this study are grouped according to their initial stage of expression and they are color-coded to indicate the type of DBD (Cys_2His_2 Zinc fingers = Blue, homeodomains = Green, bHLH = Gray, bZip = Red, Winged helix = Pink, Nuclear Hormone Receptor = Orange, POU motif= light Blue, Paired motif = Yellow, and HMG = lavender. Runt is black and was characterized as an alpha fusion (Meng et al., 2005)).

Early expression of even-skipped

Perhaps the best characterized example of the combinatorial control of a gene's expression is demonstrated by the early expression of Even-skipped (*eve*) in the *Drosophila* embryo. Like the other Pair-ruled genes, eve is initially expressed as 7 stripes that transverse the pre-cellular embryo, each of which is 5-6 nuclei wide (Small et al., 1991). These 7 stripes are controlled by 5 different CRM's that sit in a roughly 16kb region that surrounds the *eve* gene (Goto et al, 1989; Harding et al, 1989). Three of these CRM's control the expression of a single stripe (1, 2, and 5) while two others are responsible for two stripes each (3+7 and 4+6). Therefore, there are at least 5 different combinations of transcription factors that can lead to the expression of *eve* and it is the unique availability and discreet positioning these factors along the anterior-posterior axis that leads to the striped pattern.

The stripe 2 CRM is the best characterized of the 5 and is perhaps the absolute classic example of a single CRM's combinatorial control of gene expression in a metazoan. This CRM is roughly 500 bp in length and its regulation can be described by 3 gap and 1 maternal factors: Giant (Gt), Kruppel (Kr), Bicoid (Bcd), and Hunchback (Hb) (Small et al., 1991; Small et al., 1992). The CRM sequence contains 3 Giant binding sites, 3 Kruppel sites, 5 bicoid sites and 1 Hunchback site (Figure 1.2 top). Giant and Kruppel act as repressors while Bicoid and Hunchback act as activators and it is the unique pattern of these factors across the embryo that leads to the unique stripe of *eve* expression

Figure 1.2

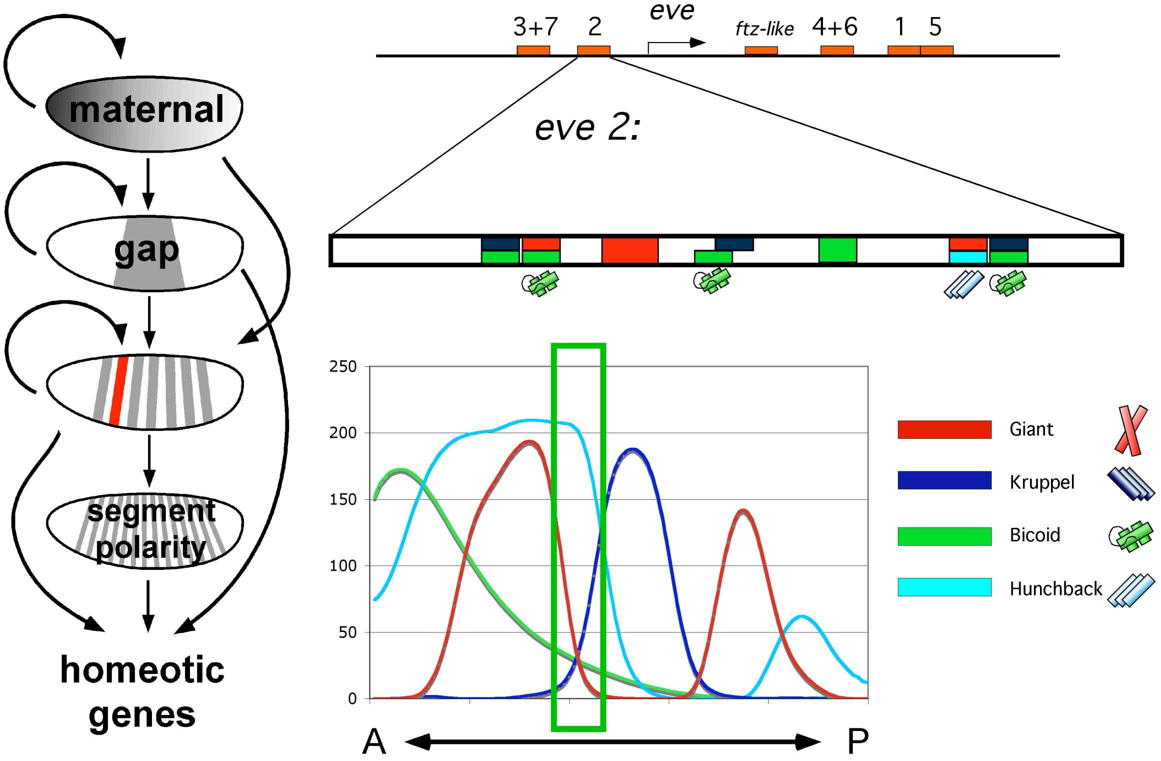

maternal
gap
segment
polarity
homeotic
genes
3+7
2
eve
ftz-like
4+6
1
5
eve 2:
250
200
150
100
50
0
Giant
Kruppel
Bicoid
Hunchback
A
P

Figure 1.2. Schematic of the CRM and the transcription factors that control the second stripe of eve's early expression. Top. The roughly 16kb genomic sequences that include the eve gene and control elements of eve's early expression are depicted. Each stripe CRM has been placed in its approximate position relative to eve's start site. The 480bp stripe 2 element is blown up to show the binding sites for the four critical transcription factors. These factors and binding sites are color coded as follows: Kruppel is dark blue, Hunchback is light blue, Bicoid is green, and Giant red. Cartoon depictions of each factor are shown in the key and what their binding might look like under stripe 2 activating conditions are shown. Bottom. The expression patterns of all four transcription factors are displayed from anterior to posterior, color coded as in A. The region that should lead to expression of the eve gene from stripe two is boxed. Notice the relatively low level of Bicoid. It is apparently Bicoid's ability to act synergistically with Hunchback that leads to stripe two activation.

controlled by this CRM. Both Bicoid and Hunchback are present as a gradient along the anterior-posterior axis with their highest concentration at the anterior pole (Myasnikova et al., 2001), though Hb has another band of expression at the posterior pole (Figure 1.2 bottom). Certainly, both activators are present in the relatively anterior position of eve's second stripe though Bicoid is tailing off. The repressors, Giant and Kruppel, are expressed in wide bands along the axis with Giant expressed more anterior and Kruppel expressed in the central embryo. Therefore, any nuclei too anterior or too central in the embryo will not be able to express eve from the stripe 2 CRM since these repressors will be present. Moreover, it is the location of the repressors, Giant on the anterior border and Kruppel on the posterior border, which sets the boundaries of where eve is able to be expressed by this CRM. With this in mind, it might seem possible that *eve* would be expressed in the posterior portion of the embryo from this CRM since both repressors are absent. However, the lack of just a single factor, Bicoid results in no expression from the CRM in the posterior embryo (Small et al., 1992). In fact, disruption of any of these genes will change the morphology of the 2nd stripe of eve expression. For example, in a Giant mutant embryo, stripes 1 and 2 are fused while in a Hunchback mutant, stripe 2 is gone or reduced.

In contrast to the stripe 2 CRM, the 3+7 and 4+6 CRMs appear to be predominantly controlled by repression. In fact, though activation from the stripe 3+7 CRM appears to be at lest partly dependant on the ubiquitously expressed dSTAT92E, additional activators for this and the 4+6 CRM are unknown (Small et al, 1996; Clyde et al., 2003).

Regardless of the activator, the borders of these stripes appear to be completely controlled by the availability of the repressors Knirps and Hunchback. Hunchback in this case is likely to act as a repressor through the recruitment of coregulators discussed below. Knirps is found in a gradiant band of the embryo roughly in the range of stripes 4-6 with its highest concentration centered on stripe 5 (Figure 1.3 Top). Hunchback, as mentioned previously, is found in gradients that are at their highest concentration near the poles and decrease towards the center of the embryo with a broader concentration near the anterior side. The internal borders of the stripes controlled by these two CRMs (posterior side of stripes 3 and 4; anterior side of stripes 6 and 7) are then set by knirps repression. The external borders of these stripes (anterior side of stripes 3 and 4; posterior side of stripes 6 and 7) are set by Hunchback repression.

An interesting comparison between the functional activity of a transcription factor and DNA-binding, both the number and strength of these events, is demonstrated by these two CRMs. Both CRMs are able to control the sharp borders of eve expression necessary to define their associated stripes yet both do this with drastically different concentrations of the repressors. This is demonstrated by the different positions of the stripes in relation to the gradients of both hunchback and knirps expression (Clyde et al, 2003). The internal borders of stripes 3 and 7 are set at much lower concentrations of knirps than stripes 4 and 6 (Figure 1.3 Bottom). This appears to be due to the greater number (12 vs 4) and quality of knirps binding sites found in the 3+7 CRM (quality is determined by

Figure 1.3

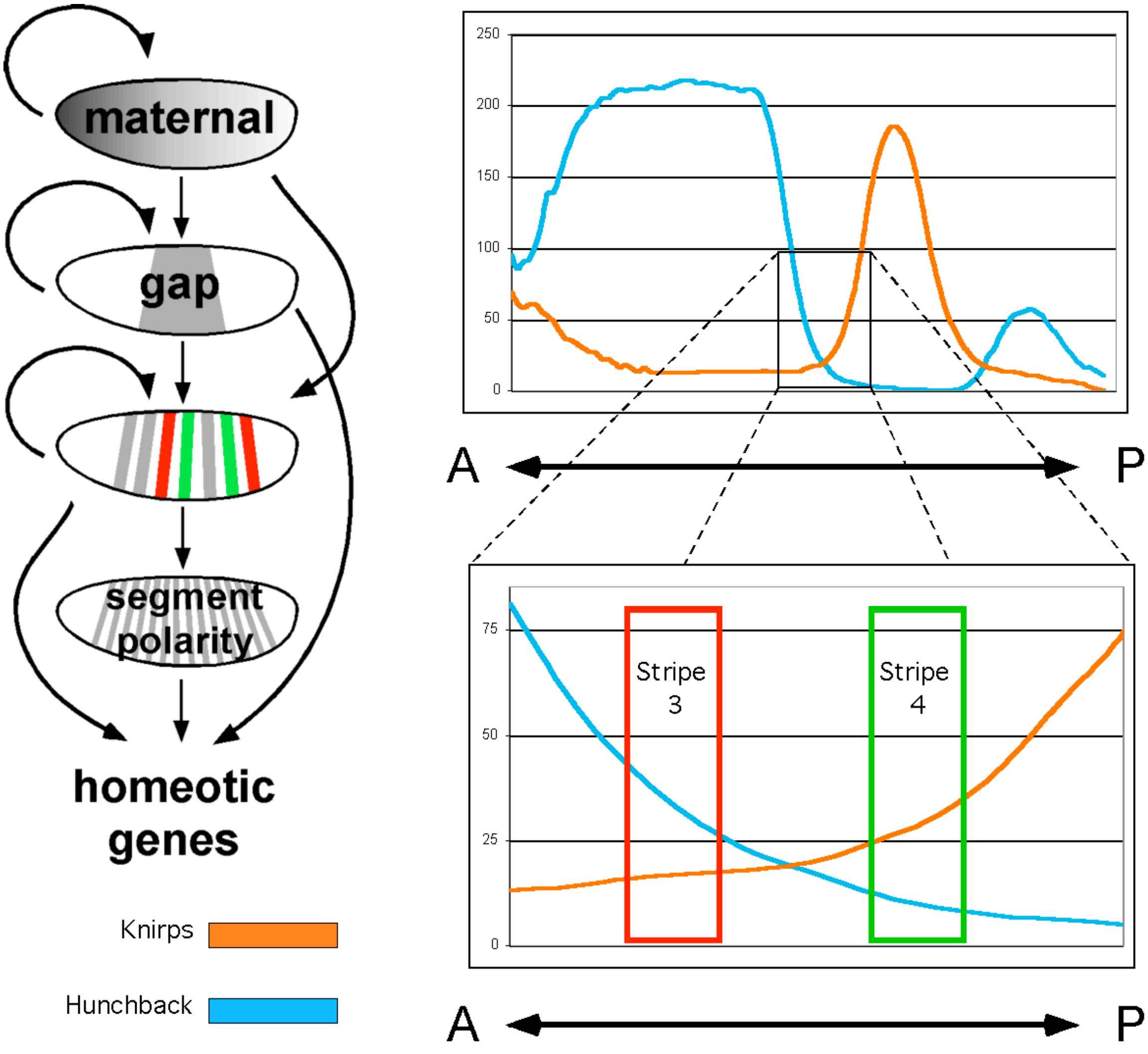

maternal
gap
segment polarity
homeotic genes
Knirps
Hunchback
250
200
150
100
50
0
A
P
75
50
25
0
Stripe 3
Stripe 4
A
P

Figure 1.3. Schematic of eve's early expression regulated by the 3+7 and 4+6 CRMs. Top. The expression patterns of Hunchback and Knirps are depicted from anterior to posterior, Hunchback = light blue, Knirps = orange. Bottom. The expression of these two factors in the more central region of the embryo is magnified. Relative positions of the stripes 3 and 4 are indicated by green boxes. Notice the posterior border of stripe 3 is at a relatively low concentration of Knirps. The 3+7 CRM contains a greater number of high quality Knirps sites in comparison to the 4+6 CRM (12 vs 4) presumably to accommodate this low concentration. Likewise, the anterior border of stripe 4 is at a relatively low concentration of Hunchback. This CRM (4+6) contains a greater number of high quality Hunchback sites.

how well the site matches a position weight matrix (PWM) of sites created for knirps by comparison to the known *in vivo* knirps binding sites. The composition of bases at each position within a set of binding sites can be connected to the free energy of binding, discussed below). Likewise, the external borders of stripes 4 and 6 are set at lower concentrations of Hunchback in comparison to stripes 3 and 7. This also appears to be due to a greater number of high affinity hunchback sites (10 vs 6) in the 4+6 CRM. In both cases lower repressor concentration has been compensated for with greater number and/or affinity binding sites in the corresponding CRM. This may also explain in part how bicoid is able to contribute to the activation of stripe 2 at relatively low concentrations. A close inspection of the stripe 2 CRM (Small et al., 1992) shows that 3 of the 5 bicoid sites are perfect matches to bicoid's consensus sequence (TAATCC). The other two sites differ by only one base (TAATC(G/T)) which should also be high affinity sites (see Chapter 4). Its possible that the low concentration of bicoid has been compensated for with 5 high affinity binding sites.

The relative quality of a binding site for a given transcription factor (e.g. bicoid, hunchback, knirps) can be approximated because the fractional composition of bases observed at each position in a set of sequences infer the relative energetic contribution of each base at each position in the binding site when compared to a reference sequence (Stormo and Fields, 1998; Stormo 1998). The reference sequence is typically the consensus sequence which we can assume is the high affinity site if each position of the binding site contributes independently to the free energy of binding. Under this assumption, the contribution to the equilibrium constant (K) for a specific base (b) at a specific position (l) should be proportional to the fraction (f) with which that base occurs in the set of sites divided by the probability (p) of that base occurring in the specific genome (not all genomes have an equal distribution of bases and therefore the background probability can be different for each Watson-Crick pair). In other words, the "weight" of a specific base can be calculated as $W_{(b,l)} = \ln f_{(b,l)}/p_{(b)}$. A partial equilibrium constant can then be calculated as $Ka \approx e^{W}/\Sigma_{b,l}\, e^{W}$. Since, $\Delta G = -RT \ln K_{eq}$, this base frequency at position "l" can be related to free energy by the equation $\Delta G_{(b,l)} = RT \ln Ka_{(b,l)}/Ka_{(ref,,l)}$, where $Ka_{(ref,,l)}$ is calculated from the base at position "l" in the consensus sequence. When considering an entire binding site (α), the term becomes $\Delta G(\alpha) = RT\, \Sigma_{(b,l)} \ln Ka_{(b,l)}/Ka_{(ref,l)}$. This calculation of $\Delta G(\alpha)$ will estimate a relative change in free energy in comparison to the consensus sequence. However, for hard numbers the binding constant of the consensus sequence would have to be experimentally determined.

The conflict between hunchback's activity as an activator in one instance and a repressor in others brings up an additional layer of regulation, the availability of coregulators. As an activator, Hunchback works together with Bicoid to express eve in the second stripe. The loss of either of these factors leads to the loss or reduction in the stripe 2 expression of eve. It appears that these factors have a synergistic effect on transcription. It has been demonstrated that a second bicoid site upstream of a reporter gene will increase the transcription by 10 fold in comparison to one bicoid site (Sauer et al., 1995a; Sauer et al., 1995b). A third bicoid binding site has little impact but a third bicoid site and a Hunchback site leads to a 65 fold increase in transcription. This appears to be due to the fact that hunchback and bicoid are able to interact with different TAF's of the TFIID complex. Bicoid is able to interact with both $TAF_{II}110$ and $TAF_{II}60$ using two separate domains (Sauer et al., 1995a). Hunchback is able to bind to $TAF_{II}60$ (Sauer et al., 1995b). When any of these interactions are disrupted the synergy is lost. Therefore, it appears that different combinations of hunchback and bicoid recruit multiple compontents of the TFIID complex and this leads to the synergistic assembly of the PIC at the neighboring promoter. It is likely that the activation of eve from the stripe 2 CRM, even at a low concentration of bicoid, is due to the synergistic recruitment of the PIC by the 3 potential pairs of factors on the CRM's 6 high affinity binding sites (5 bicoid and 1 Hb).

By contrast, the repressor activity of hunchback may be due to its interaction with another coregulator, dMi-2 (Tautz 1988). Hunchback is able to bind dMi-2 with a conserved D domain and this domain is required for its activity as a repressor. In addition, in the absence of dMi-2 Hunchback's repression of downstream HOX factors has been shown to be derepressed. In *Drosophila*, dMi-2 has been shown to interact with the Polycomb Group Proteins (Kehle et al, 1998) that lead to gene silencing by chromatin remodeling. In *Xenopus*, Mi-2 has been found associated with a histone deacetylase complex (Wade et al, 1999). Therefore, it is likely that Hunchback acts as a repressor through its interaction with dMi-2 and potentially other coregulators. It is unclear how it is determined which partner Hunchback will interact with. At least early in the embryo dMi-2 is ubiquitously expressed (Khattak et al, 2002), so the availability of dMi-2 is not the limiting factor in the central embryo where hunchback acts as an activator. There are likely to be other cofactors that are influencing where and when Hunchback will interact with dMi-2 or $TAF_{II}60$.

Thus far this discussion has focused on the regulatory networks involved in the expression of a single gene at a single time in development. However, this is just a small step in the program of development and the outcome of this network's control of eve's expression becomes just a single input in downstream networks. For example, one of eve's primary roles is to indirectly regulate the expression of engrailed, a segment-polarity gene, that will establish rigid borders between the stripes (Manoukian and Krause 1993; Kuhn et al, 2000). This regulation is dictated by the further refinement of eve's expression to the more anterior region of each stripe and the expression of fushi tarazu (ftz) in the more posterior (Manoukian and Krause, 1992). This results in14 narrow parasegments, 7 alternating eve and ftz stripes, that will coincide with the 14 segments of the adult body plan. The graded expression of eve, a repressor, from anterior to posterior will help to maintain engrailed as off. In the 7 alternating posterior segments the expression of engrailed will be activated by ftz. However, even this expression is only in the very posterior nuclei of each stripe, creating a stark contrast between the parasegments. In summary, the expression of a single gene is the result of a vast network of influences. The CRMs that regulate expression decode all of the available temporal and spacial information in order to determine a given outcome. However, it must not be lost that this is just a node, a single piece of information, in a higher network that may coordinate a cell's matabolism, a cell to cell signal, or even the development of an entire organism.

Identifying regulatory elements in metazoans

The identification of *cis*-regulatory sequences throughout the genome and their complementary transcription factors is obviously a powerful step in deconstructing the mechanism of spatial and temporal gene regulation. This has been eloquently demonstrated in the early segmentation of the *Drosophila* embryo. For the investigation of the TF-CRM relationship in other pathways, the majority of sequence-specific transcription factors in a genome can be readily identified by sequence homology to members of previously identified families of DBDs. This type of identification has revealed that typically 5 to 10 percent of the protein coding genes within a eukaryotic genome are TFs, with more complex organisms containing a higher proportion of TFs, presumably due to the requirement for more elaborate regulatory networks to control greater cellular complexity (Levine and Tjian, 2003). However, the number of CRMs within a genome that contain binding sites for these factors is hard to define due to the difficulty in identifying these elements based on sequence features alone (Ludwig et al., 1998) and, for more complex organisms, because this number presumably far exceeds the number of genes within the genome (Siepel et al., 2005). Therefore, defining the subset of functional interactions between these two groups - TFs and CRMs - is a complex problem in higher eukaryotes, where the vast majority of DNA is non-coding sequence. For example, the regulatory network that controls the anterior-posterior segmentation of the *Drosophila* embryo utilizes less than 40 transcription factors and took several decades of meticulous genetic study to deconvolute. By comparison, the human genome contains potentially 2,000 TFs or more that are involved in countless overlapping regulatory networks presenting an enormous challenge to the decoding of the networks that define gene expression in higher eukaryotes.

Fortunately, biochemical and computational methods for the identification of CRMs within the genome have been developed that are beginning to fulfill this goal, yet limitations remain. Biochemical methods based on genome-wide location analysis or "ChIP-chip" (Harbison et al., 2004; Lee et al., 2002; Zeitlinger et al., 2007), nuclease hypersensitive sites (Crawford et al., 2004; Sabo et al., 2004) and 5C (Dostie and Dekker, 2007; Dostie et al., 2006) allow the identification of functional elements throughout the genome. However, these techniques are limited typically to cell types that can be obtained in sufficient quantities for each protocol and, with the exception of ChIP-chip, cannot associate specific TFs with an identified CRM. The ChIP-chip approach, though able to make the TF-CRM association, requires an antibody specific to the TF of interest limiting the breadth of this protocols potential application. Furthermore, identification of genomic binding sites by ChIP-chip does not reveal whether those sites are functional; binding sites that are occupied *in vivo* may not contribute to organismal fitness, as long as they do not have negative consequences (Gao et al., 2004; Zeitlinger et al., 2007). An alternative is to take a gene centered approach whereby a set of genomic sequences can be used to screen a pool of TFs by a yeast one-hybrid assay (Deplancke et al, 2004; Deplancke et al, 2006). Because the yeast one-hybrid assay requires activation of a reporter gene, functional pairs determined by this method are likely to be representative of the functional interactions *in vivo*. However, how a TF is binding within a genomic sequence may be difficult to ascertain.

Computational methods to identify CRMs have focused primarily on two types of approaches: Phylogenetic footprinting and Binding site cluster analysis. Phylogenetic footprinting, where short conserved sequence blocks in non-coding sequence between species are used as surrogates for TF binding sites (McCue et al., 2002; Lenhard et al., 2003; Grad et al., 2004), can identify regions of a genome between closely related species that may be under stabilizing selection, a property of CRMs. However, this approach does not *a priori* associate TFs with identified CRMs. CRMs can also be computationally identified by searching for binding site clusters, groups of binding sites for TFs that function in a common transcriptional regulatory network, within sequence windows on the order of 500 bp (Berman et al., 2002; Lifanov et al., 2003; Markstein et al., 2002; Rajewsky et al., 2002; Sosinsky et al., 2003).. The accuracy of these predictions can be improved by incorporating phylogenetic comparisons between species separated by moderate evolutionary distances (Schroeder et al., 2004; Sinha et al., 2004).

The prediction of CRMs and their cognate factors via binding site cluster analysis has been most thoroughly studied in the context of the regulatory cascade driving anterior-posterior segmentation of the *Drosophila* embryo. The detail with which this pathway has been documented allows for a level of validation in the predictions a given method is able to make. However, even for this carefully studied regulatory network the architecture of CRMs neighboring genes within this regulatory cascade and the binding sites for the controlling factors are incompletely defined (Arnosti, 2003; Arnosti et al., 1996). The DNA-binding specificity of many prominent TFs involved in this process is imprecisely defined (*e.g.* Slp1, Kni, D & Tll), and though the majority of TFs can be identified by sequence homology alone, their DNA-binding specificities typically cannot be directly inferred based on amino acid sequence unless a direct homolog of the factor has been characterized. This dearth of DNA-binding specificities for factors involved in this regulatory cascade, has limited the scope of binding site cluster analysis for the identification of novel CRMs. Nonetheless, position weight matrices (PWMs) for subsets of the TFs involved in this regulatory network have been successfully utilized to identify novel CRMs (Berman et al., 2002; Berman et al., 2004; Rajewsky et al., 2002; Schroeder et al., 2004). A more complete set of specificities for factors in this pathway should provide a powerful database for the prediction of CRMs utilized by these early embryonic regulators.

Binding site cluster analyses followed by a phylogenetic filter will no doubt become a more and more powerful methodology as additional genomes are sequenced, but this approach is only possible if the DNA-binding specificities for the set of TFs of interest are defined. Therefore, the limited number of available TF DNA-binding specificities is clearly a major hurdle to deconstructing the network of TF-CRM pairings in a genome. This limitation could be addressed by the implementation of a high throughput system for determining TF specificity. Furthermore, the determination of specificities that require a biological activation are likely to be a closer representation of a TF functional *in vivo* specificity in comparison to purely *in vitro* methods. These "functionally" determined specificities could be used to filter which ChIP-recovered sequences are likely to be functional in a genome-wide location analysis. In addition, specificities for large sets of factors within a common DBD family might allow for the generation of a recognition code for that domain family. Such a code could allow for the prediction of specificities in other genomes that alleviate the need to experimentally characterize every like factor in other genomes. The combination of expression data, genome-wide location analysis and TF specificities would provide a powerful combination of data for the prediction of CRMs. A high throughput method for the determination of DBD specificities would therefore make a great contribution to our understanding of regulatory networks.

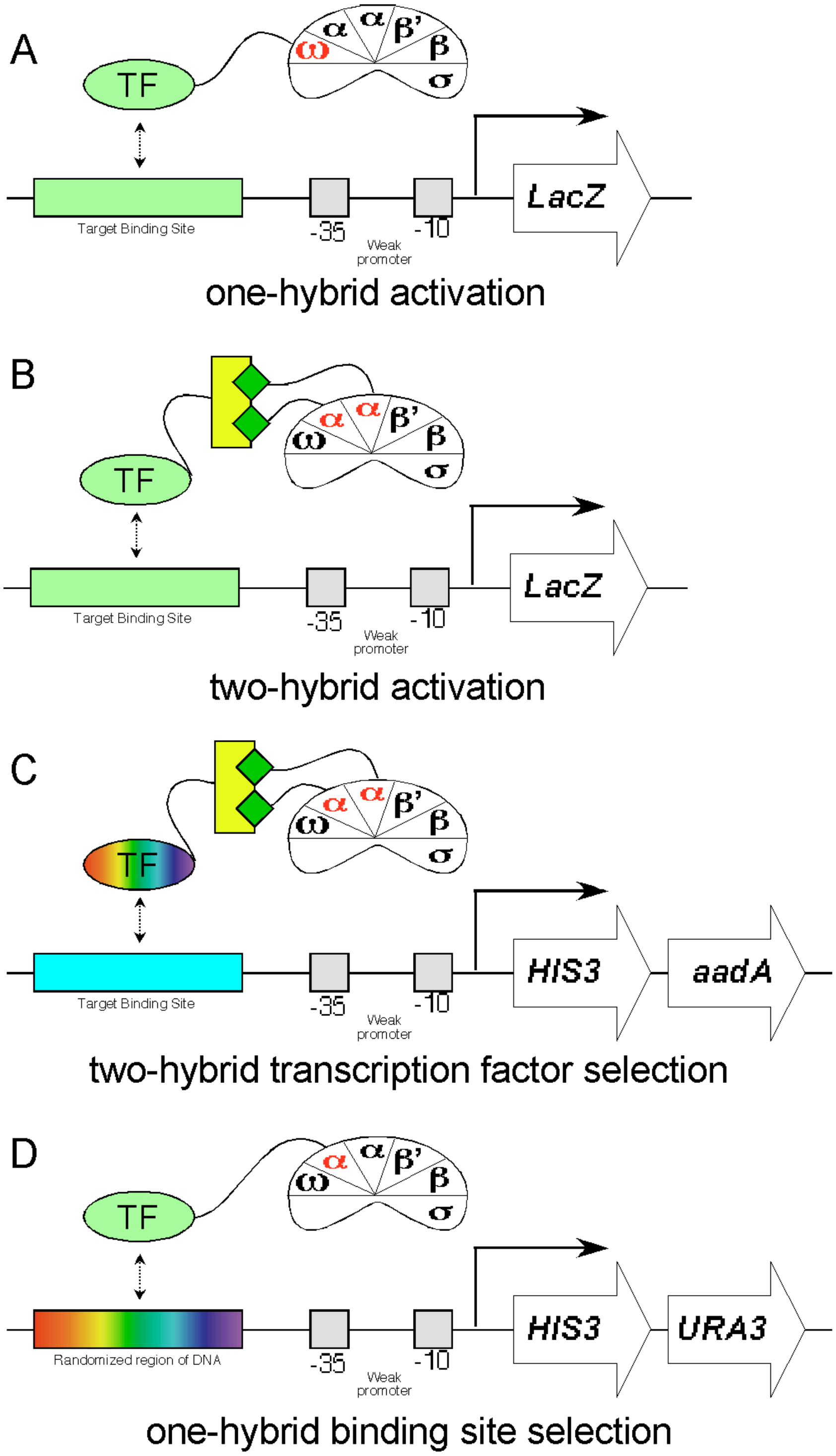
Characterizing Transcription Factor Specificity

The small proportion of TFs with well-characterized DNA-binding specificities is not limited to *Drosophila*; this state of knowledge is representative of the majority of eukaryotic genomes. This void reflects the absence of a readily accessible high-throughput method for characterizing the specificity of factors. Various methods have been developed for the characterization of sequence-specific DNA-binding domains. Traditional *in vitro* methods such as DNase I footprinting (Bergman et al., 2005) and

SELEX (Ellington and Szostak, 1990; Roulet et al., 2002; Tuerk and Gold, 1990; Wright and Funk, 1993), are cumbersome to employ at a genome-wide level for TFs in a complex eukaryotic genome such as *Drosophila*, which contains ~750 sequence specific TFs. Other methods such as ChIP-chip and DIP-chip have been employed on a case-by-case basis, but have not yet been scaled for high-throughput analysis of protein-DNA interactions (Harbison et al., 2004; Lee et al., 2002; Lieb et al., 2001; Zeitlinger et al., 2007). Protein binding microarrays provide one potential approach for characterizing a wide-variety of TFs (Berger et al., 2006; Bulyk et al., 2001; Linnell et al., 2004; Mukherjee et al., 2004). This platform, while powerful, has certain barriers that may limit its widespread adoption as a technology: it requires protein purification of each factor that is being characterized and the synthesis of custom microarrays used in the analysis.

Previous studies have demonstrated that bacteria provide an attractive platform to assay protein-DNA interactions that would not require protein purification. The Hochschild lab demonstrated that a reporter gene could be activated in bacteria by the DBD-mediated recruitment of the RNAP holoenzyme to a specified promoter (Joung et al., 1993; Joung et al., 1994; Dove et al., 1997; Dove and Hochschild, 1998). They also demonstrated that this recruitment could be mediated by a direct fusion to the omega subunit (one-hybrid interaction; Dove and Hochschild, 1998) or an indirect contact through an alternative protein-protein interaction such as the GAL4-GAL11P interaction (two-hybrid interaction; Dove et al., 1997; Dove and Hochschild, 1998) (Figure 1.4A, B).

Figure 1.4



one-hybrid activation

two-hybrid activation

two-hybrid transcription factor selection

one-hybrid binding site selection

Figure 1.4. Schematic of bacterial one and two-hybrid systems. A. One-hybrid activation. The Hochschild lab demonstrated that a transcription factor fused directly to a subunit of RNAP (here an omega fusion is shown) could activate transcription of a reporter gene if the recognition sequence for that TF was placed in an appropriate position upstream of the promoter. B. Two-hybrid activation. The Hochshild lab also demonstrated this same type of activation can be mediated by a protein-protein interaction such as the GAL4-GAL11P interaction. Here the TF is fused to GAL11P and the C-terminal domain of the alpha subunit has been replaced by the dimerization domain of GAL4. C. Two-hybrid selection. The Pabo lab was able to show that zinc fingers with novel specificity could be selected from pools of randomized zinc fingers through the GAL4-GAL11P mediated activation (the rainbow color indicates a library of random clones). D. One-hybrid binding site selection. The Wolfe lab demonstrated that activation mediated by the direct fusion of a transcription factor of interest to the alpha subunit could be utilized to select functional sequences from a library of randomized DNA upstream of a reporter.

Later, the Pabo lab demonstrated that DNA-binding domains with novel specificities could be selected from large libraries using a similar system (Joung et al., 2000) (Figure 1.4C). This later system demonstrates the advantages that the high transformation efficiency and fast growth rate of bacteria provide for the selection of library members with unique attributes from a large randomized pool. In addition, this bacterial two-hybrid system introduced the yeast *HIS3* gene as the reporter. In cells that have the bacterial *hisB* gene knocked out, this reporter makes survival dependant on its activation when plated on media that lacks histidine. This allows for the screening of libraries greater than 10^8 on a single petri dish. Finally, the system was modified by Meng et al. to determine DNA-binding specificities of different DBDs by selecting binding sites from a library of DNA sequences upstream of the promoter (Meng et al, 2005) (Figure 1.4D). Here, the DNA-binding domain of interest was expressed as a direct fusion to the α subunit of RNA polymerase. In addition, a second reporter, the yeast *URA3* gene, was installed downstream of the *HIS3* gene. This was necessary because the large library of DNA sequences could easily contain a large number self-activating sequence such as "up-elements". The *URA3* gene allows for a counter-selection in the presence of 5-fluoroorotic acid and therefore removal of sequences that activate independently of a DNA-binding domain.

We have previously described a bacterial one-hybrid (B1H) system for the rapid characterization of transcription factor specificity (Meng et al., 2005; Meng and Wolfe, 2006). This technology has certain attributes that make it suitable as a platform for the genome-wide analysis of DNA-binding domain specificities. The bacterial selection precludes the need to purify any given factor. Binding sites for a factor are isolated in a single round of selection where multiple selections can be performed in parallel, which provides an avenue for the high-throughput analysis of factors. Standard molecular biology and sequencing technologies are employed, making the technology accessible to most laboratories.

Herein we describe substantial improvements to the B1H system that increase its sensitivity and dynamic range, which make it amenable for the high-throughput analysis of sequence-specific TFs. Currently we have characterized 108 (14.3%) of the predicted TFs in *Drosophila*, demonstrating the feasibility of characterizing a large number of TFs using this technology. These characterized factors fall into a broad range of DBD families that are commonplace in eukaryotic genomes, including the five most common found in humans: Cys_2His_2 zinc finger, homeodomains, bHLH, bZip and Forkhead domains. Our overall success rate for characterization of individual factors using this system is currently >95%.

One of our groups of characterized factors is focused around those that play important roles in early A-P patterning with previously uncharacterized or poorly defined specificities. This dataset dramatically expands the set of defined specificities for these factors. In individual tests, the PWMs for the set of maternal, gap and terminal factors, for which a large number of target CRMs have already been identified, show strong correlations between the expression pattern of the TF and the enrichment of their binding sites in CRMs that regulate gene expression in neighboring or overlapping positions within the embryo. To fully exploit this large database of binding specificities, we have created a GBrowse-based search tool (Stein et al., 2002) that allows an end-user to examine the overrepresentation of binding sites for any number of individual factors as well as combinations of these factors in up to six Drosophila genomes (veda.cs.uiuc.edu/cgi-bin/gbrowse/gbrowse/Dmel4). Peaks of significant binding site overrepresentation for combinations of gap and maternal TFs are found in the majority of characterized CRMs regulated by these factors. This search tool has also been enabled to allow the end-user to search the fly genome for the most significant peaks of overrepresentation for combinations of factors. Using this tool with the PWMs for anteriorly expressed TFs, we find a remarkable correlation between the strongest hits in the genome and previously characterized anterior stripe CRMs, as well as a number of interesting predictions of potential novel CRMs neighboring genes with confirmed anterior expression patterns. Fundamental questions of how combinations of factors act at individual enhancers to produce diverse patterns of gene expression during development can potentially be addressed within the framework of this system. Thus, the combination of this tool and the dataset provided by our high throughput B1H system can be used as a test-bed for dissection of the TF/CRM combinations that control gene expression during fly development. As more PWMs for factors are added to the dataset, this should provide a framework for efforts to computationally map CRMs in the fly on a comprehensive scale.

To demonstrate the feasibility of characterizing an entire class of DBDs, we also characterized all of the independent homeodomains within the fly genome. This subset of factors provides a powerful basis set for predicting the specificities of homeodomains throughout eukaryotic genomes. Herein we provide a complete catalog of specificities for all 84 homeodomains in *Drosophila* that are not associated with an additional DNA-binding domain and use the rich experimental history of homeodomain studies in *Drosophila* to help interpret this dataset. The binding specificities of these factors can be clustered broadly into eleven specificity groups that encompass the majority of these factors. Homeodomains within these clusters typically share common recognition residues; coupled with previous structural and biochemical work on the homeodomain family, this data provides a global perspective on the specificity determinants within this family. Based on these observations we propose and test a detailed set of recognition rules for homeodomains and use this information to predict the specificities of the majority of homeodomains in the human genome. Furthermore, the Drosophila PWMs and the human homeodomain predictions are made available at http://ural.wustl.edu/flyhd/. The user is also able to input any homeodomain sequence from any genome and a specificity prediction will be made if by comparison to *Drosophila* factors in our set it is found similar enough with regards to both overall sequence similarity and the identity at key recognition residues. These predictions will become more robust as diverse homeodomain specificities from other genomes are characterized and as the number sequences from a given selection are increased. This could be done by deep sequencing pools of selected clones to provide a potential 10 to 100 fold increase in the number of sequences currently used in each PWM.

CHAPTER II: CHARACTERIZATION OF A BACTERIAL ONE-HYBRID SYSTEM FOR THE DETERMINATION OF DNA-BINDING SPECIFICITY

Introduction

The identification of *cis*-regulatory sequences throughout the genome and the complementary sequence-specific trans-acting factors that bind within these modules is an important step in deciphering the mechanism of spatial and temporal gene regulation in metazoans. The majority of sequence-specific transcription factors (TFs) in a eukaryotic genome can be readily identified by sequence homology to previously identified families of DNA-binding domains, where complex organisms usually contain a higher proportion of TFs (~5 to 10%) due to the requirement for more elaborate transcriptional regulatory networks (Levine and Tjian, 2003). However, identifying *cis*-regulatory modules (CRMs) within a genome is difficult due to the more dynamic nature of these sequences relative to coding sequences (Ludwig et al., 1998) and the fact that the vast majority of DNA in higher eukaryotes is non-coding sequence (Siepel et al., 2005). CRMs have been computationally identified by searching for overrepresented clusters of binding sites within the genome for groups of TFs that function in a common transcriptional regulatory network (Berman et al., 2002; Lifanov et al., 2003; Markstein et al., 2002; Rajewsky et al., 2002; Sosinsky et al., 2003). The accuracy of these predictions can be improved by incorporating phylogenetic comparisons between species separated by moderate evolutionary distances (Schroeder et al., 2004; Sinha et al., 2004). However, this approach is hindered by the general lack of transcription factor specificity data.

The small proportion of TFs with well-characterized DNA-binding specificities is not limited to *D. melanogaster*. This incomplete state of knowledge is representative of the majority of eukaryotic genomes and reflects the absence of high-throughput studies of factor specificities. *in vitro* methods for characterizing specificity include DNaseI footprinting (Bergman et al., 2005), SELEX (Ellington and Szostak, 1990; Roulet et al., 2002; Tuerk and Gold, 1990; Wright and Funk, 1993), and protein binding microarrays (Berger et al., 2006; Bulyk et al., 2001; Linnell et al., 2004; Mukherjee et al., 2004). All of these techniques require protein purification, limiting their application as a high-throughput methodology. To date, these methods have not been widely adopted for large-scale analysis of TF specificities. In addition, it remains unclear whether the off-rate measurements obtained via some of these *in vitro* methods are representative of the *in vivo* recognition properties of these factors (Berger et al., 2006).

TF specificities can also be identified as overrepresented motifs within DNA sequences identified in genome-wide TF ChIP datasets (Harbison et al., 2004; Lee et al., 2002; Lieb et al., 2001; Zeitlinger et al., 2007). When applied to the comparatively simple yeast genome, this approach successfully identified high confidence motifs for 65 of 203 (32%) TFs (Harbison et al., 2004). The inability to determine specificities for the majority of these factors may reflect the difficulty in identifying motifs within the larger sequence segments defined by ChIP experiments and the complications associated with TFs that bind DNA in complexes with one or more other TFs.

We have previously described a bacterial one-hybrid (B1H) system for the rapid characterization of TFs (Meng et al., 2005; Meng and Wolfe, 2006). This technology has certain attributes that make it suitable as a platform for the genome-wide analysis of DNA-binding domain specificities. Selections are performed *in vivo*, which precludes the need to purify any given factor. Moreover, binding sites are isolated based on their ability to activate a biological response in the context of competition from a pool of potential sites in the *E. coli* genome, which simulates the functional requirements in a eukaryotic genome. Binding sites for a factor are isolated in a single round of selection using standard molecular biology and sequencing technologies, making it accessible to most laboratories. Here, we describe substantial improvements to the B1H system that increase its sensitivity and dynamic range, and make it amenable for the high-throughput analysis of sequence-specific TFs (Figure 2.1). Using this system, we have determined specificities for 108 (14.3%) of the predicted TFs in *D. melanogaster*. These factors represent a broad range of DNA-binding domain families that are commonplace in eukaryotic genomes. Thus far, we have focused on two groups of factors. One group represents all of the independent homeodomains within the fly genome, which provides a basis for predicting the specificities of homeodomains throughout the eukaryotic kingdom (*see Chapter 4*). Members of the other group play prominent roles in early A-P patterning *(see Chapter 3*). Our dataset dramatically expands the set of defined specificities for these factors and these motifs are good predictors of CRMs throughout the genome.

Figure 2.1

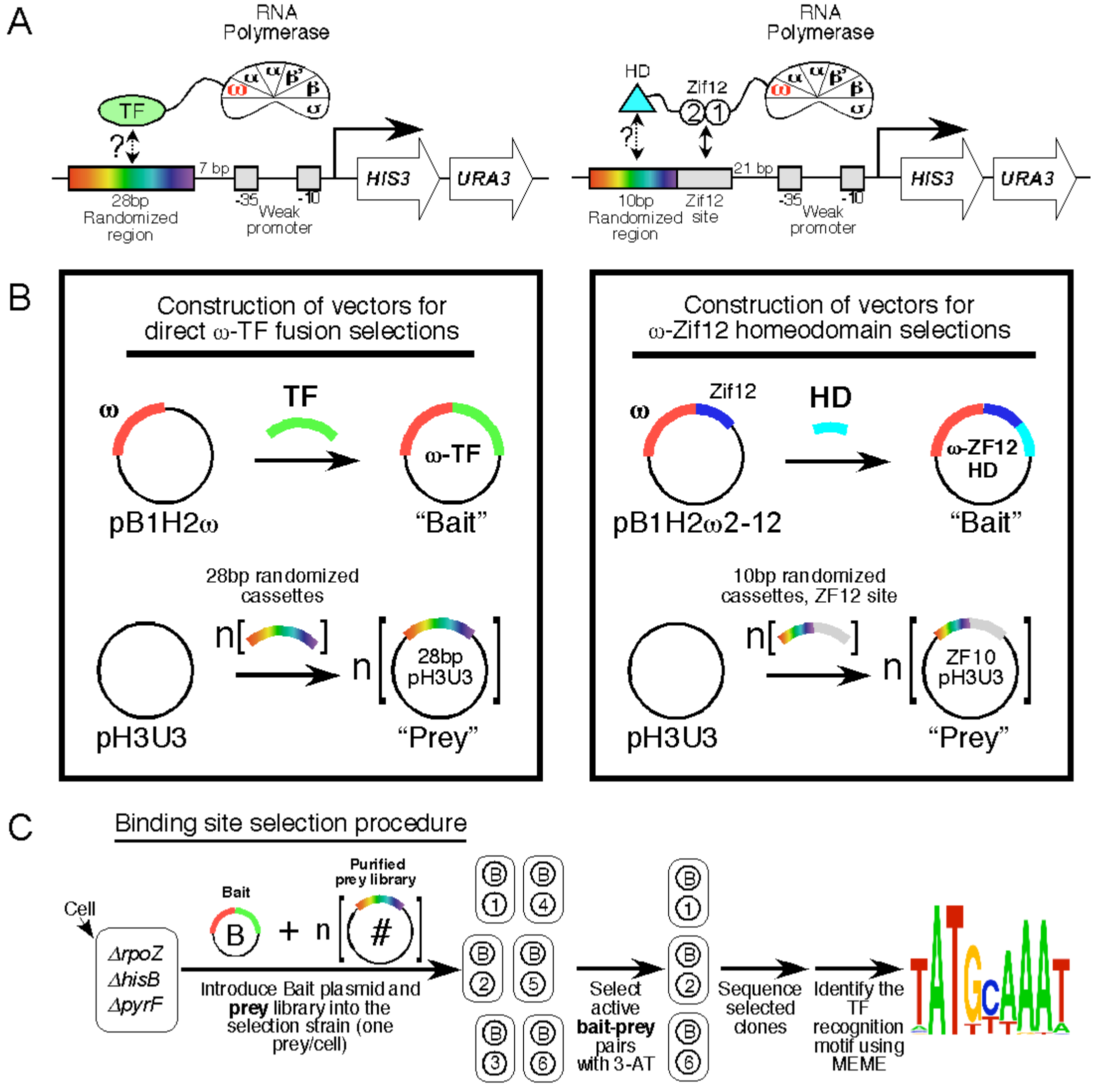

A
RNA Polymerase
TF
?
ω
α
α
β'
β
σ
7 bp
28bp Randomized region
-35
-10
Weak promoter
HIS3
URA3
RNA Polymerase
HD
Zif12
2
1
?
21 bp
10bp Randomized region
Zif12 site
-35
-10
Weak promoter
HIS3
URA3
B
Construction of vectors for direct ω-TF fusion selections
ω
TF
pB1H2ω
ω-TF
"Bait"
28bp randomized cassettes
n[ ]
n
28bp pH3U3
pH3U3
"Prey"
Construction of vectors for ω-Zif12 homeodomain selections
ω
Zif12
HD
pB1H2ω2-12
ω-ZF12 HD
"Bait"
10bp randomized cassettes, ZF12 site
n[ ]
n
ZF10 pH3U3
pH3U3
"Prey"
C
Binding site selection procedure
Cell
ΔrpoZ
ΔhisB
ΔpyrF
Bait
B
+ n
Purified prey library
#
Introduce Bait plasmid and prey library into the selection strain (one prey/cell)
B 1
B 4
B 2
B 5
B 3
B 6
Select active bait-prey pairs with 3-AT
B 1
B 2
B 6
Sequence selected clones
Identify the TF recognition motif using MEME
TATGCAAAT

Figure 2.1. Overview of the omega-based B1H system. A.) Cartoon depicting recruitment of direct omega fusions (left) and omega-Zif12-HD fusions (right) to the weak promoter driving the *HIS3* and *URA3* reporters used in this system. The 28 base pair library is positioned 7 bases upstream of the -35 box allowing the TF to bind to a recognition element up to 3 turns upstream of the promoter. The ZF10 library has the binding site for Zif12 (TGGGCGG) positioned 21 bases upstream of the promoter and the 10 base pair randomized region is located immediately 5' to this site. B.) Overview of Bait and Prey plasmids used in this system. Bait plasmids are constructed by cloning the TF of interest as a C-terminal fusion to omega (omega-TF hybrid). Homeodomains are cloned into a modified bait plasmid (pB1H2ω2-12) that results in their expression as an omega-ZF12-HD hybrid. C.) Binding site selection procedure. A bait plasmid and the appropriate prey plasmid are transformed into the selection strain. Transformants are grown on minimal media lacking histidine and challenged with various concentrations of 3-AT. Surviving colonies represent a complementary interaction between the bait plasmid (TF) and a single member of the prey library. The library region from approximately 20-25 surviving colonies are amplified by colony PCR and sequenced. The resulting sequences are analyzed by MEME (Bailey and Elkan, 1994) to recover the TF's recognition motif.

Results

Our original B1H system for characterizing DNA-binding specificity utilized TF fusions to the alpha-subunit of RNA polymerase (alpha-TF) (Meng et al., 2005; Meng and Wolfe, 2006). This system contained three components: the alpha-TF expression vector, a tandem HIS3-URA3 reporter cassette in a low copy number plasmid (pH3U3), and the selection strain with the bacterial homologs of the reporter genes inactivated (*ΔhisB*, *ΔpyrF*). The HIS3-URA3 reporter cassette is regulated by a weak promoter and consequently these genes, which provide a direct method for auxotrophic selection, are only weakly transcribed. However, when a functional binding site for the alpha-linked TF is present upstream of the weak promoter, RNA polymerase can be actively recruited to stimulate transcription of the reporter cassette (Dove et al., 1997). Thus, bacteria harboring a complementary interaction between the TF and reporter DNA can be selected under appropriate growth conditions, which permits binding sites complementary to a TF to be isolated from a randomized library introduced into the reporter vector. Our alpha-based system, while suitable for characterizing factors such as Cys_2His_2 zinc finger proteins, proved ineffective with several basic helix-loop-helix proteins (bHLH) and homeodomains. The origin of this limitation was unclear, but one potential source was insufficient sensitivity: Alpha is an essential gene, and as such, alpha-TF fusions are in competition with endogenous alpha for incorporation into RNA polymerase complexes.

Omega is the only conserved component of bacterial RNA polymerase ($\alpha_2\beta\beta'\omega$) that is not required for viability under laboratory growth conditions (Gentry and Burgess, 1989). Hochschild and colleagues demonstrated that in artificial interactions between a sequence-specific TF and the omega-subunit of RNA polymerase, like interactions with the alpha-subunit, could mediate activation of a nearby promoter (Dove and Hochschild, 1998). Because Omega is not required for viability, Omega-fusions have the potential advantage that selections might be performed in an omega-knockout (*ΔrpoZ*) strain, where omega-fusions could be uniformly incorporated into RNA polymerase without competition. Under these conditions the selection system should be more sensitive due to the higher cellular concentration of RNAP-TF complexes, allowing weaker protein-DNA interactions to be characterized.

To test this hypothesis we knocked-out the rpoZ gene in our selection strain (Figure 2.2) and examined the activity of an omega-Zif268 fusion with a reporter vector containing a Zif268 binding site. The fusion was expressed using three promoter strengths: the dual promoter used in the original alpha-based system (*lppC-lacUV5*), a *lacUV5* promoter and a mutant *lacUV5* promoter (*lacUV5m*) (Figure 2.3). Omega-Zif268 expressed via the weakest (*lacUV5m*) promoter displayed robust activity, allowing cells to survive at higher 3-AT concentrations than was tolerated by the alpha-Zif268 fusion under optimal expression conditions. Surprisingly, omega-Zif268 constructs expressed with either the dual promoter or the *lacUV5* promoter proved toxic. However, for other

Figure 2.2

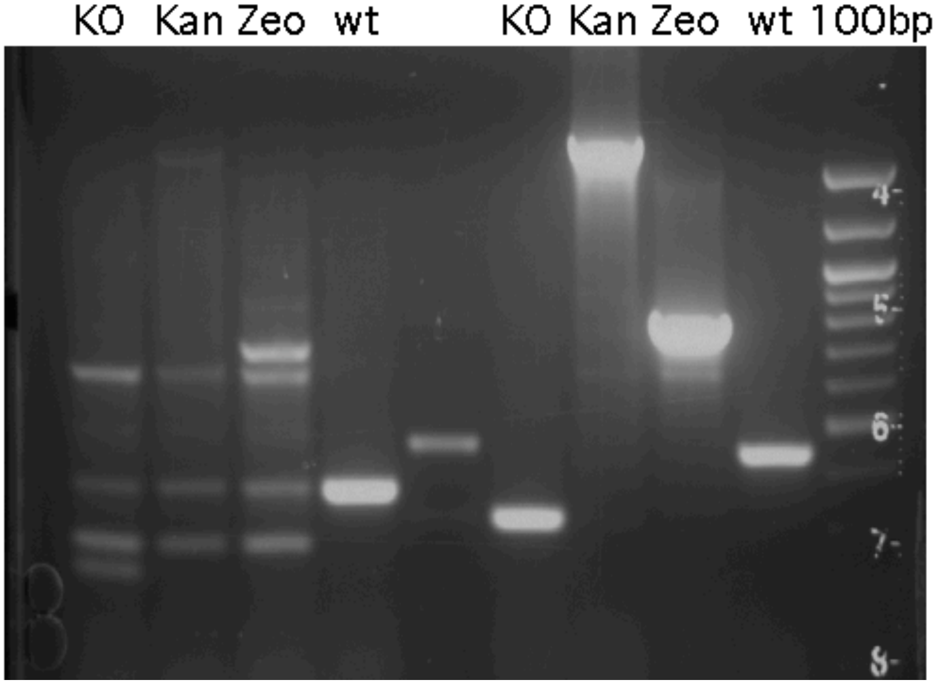
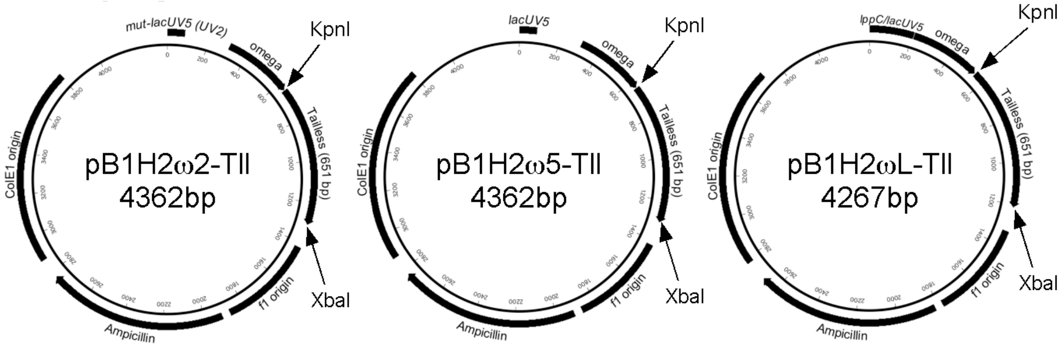


Figure 2.2._PCR products of rpoZ locus. These loci were disrupted by knock-in or knockout following the recombination method detailed by Wanner (Datsenko and Wanner, 2000). The PCR products generated from genomic DNA of our *rpoZ* cell types run on a 1.5% agarose gel. Above each lane the insertion at the *rpoZ* locus of the particular cell type is listed. Wt is the wild type *rpoZ* gene, Kan is the kanamycin cassette, Zeo is the zeocin cassette, and KO has the antibiotic cassette removed resulting in a complete removal of the *rpoZ* locus. On the left side of the gel are the PCR products from the *rpoZ 5’* primer and an internal omega primer (omega int 3’) that should only amplify the wt gene. On the right side of the gel are PCR products using the two external primers described in the text, *rpoZ 5’* and *rpoZ 3’*. These give an indication of the relative size of the insertion/deletion at the *rpoZ* locus.

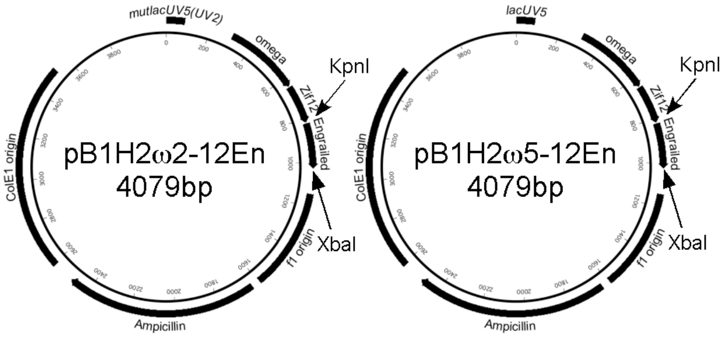
Figure 2.3



Linker between omega subunit and TF (69bp)

gcggccgcggactacaaggatgacgacgacaagttccggaccggttccaagacacccccccat**ggtacc**-*TF(TII)*-TAA**tctaga**

NotI **KpnI** **stop XbaI**



Linker between Zif12 and HD (15bp)

HisThrGlyThrGlyAsp

Zif 1,2 -cacacc**ggtacc**ggtgac-*HD(En)*-TAA**tctaga**

KpnI stop XbaI

Figure 2.3. Maps of the bait plasmids used in the omega-B1H system. The maps of the bait plasmids used to characterize the TF reported here are displayed with key features annotated. The plasmids allow a TF to be characterized at 3 different promoter strengths as direct fusions to omega or as fusions mediated by Zif268 fingers 1 and 2 (Zif12) by simply subcloning between the unique Kpn1 and Xba1 sites in each plasmid. The linkers between the omega subunit or Zif12 and the TF are indicated below the plasmid sets.

factors (Paired, Hunchback and Giant) higher expression levels achieved through the stronger promoters were required to fully activate the reporter system (Figure 2.4). The difference in promoter strengths used to drive expression of each factor was reflected in the relative protein expression levels of each factor within the cell (Figure 2.5). Thus, the availability of three different promoter strengths provides flexibility to characterize a wide variety of TFs that may differ in affinity, specificity and level of expression.

The omega-based B1H system is sensitive to changes in the strength of the interaction between a DNA-binding domain and its target site. The activity of omega-Zif268 with its consensus sequence was compared to three different variants of the binding site that have four to twenty-fold reduced affinity(Miller and Pabo, 2001). A clear correlation is observed between colony size and number with the quality of the binding site: cells containing the consensus sequence within the reporter displayed the highest rates of survival and the largest colonies relative to the survival rates and colony sizes for other sites with decreased affinity (Figure 2.6). Based on these results we expect that the distribution of sequences that are recovered from a binding site selection will be a function of the difference in affinity of the protein for these sites. As a result the recognition motif constructed from the selected sites should accurately reflect the specificity of the factor.. The optimal position of the Zif268 binding site was determined by examining the activity of reporters harboring sites positioned in various registers relative to the promoter. Two peaks of maximal activity were observed for sites positioned either 10 bp or 21 bp upstream of the -35 box (Figure 2.7).

Figure 2.4

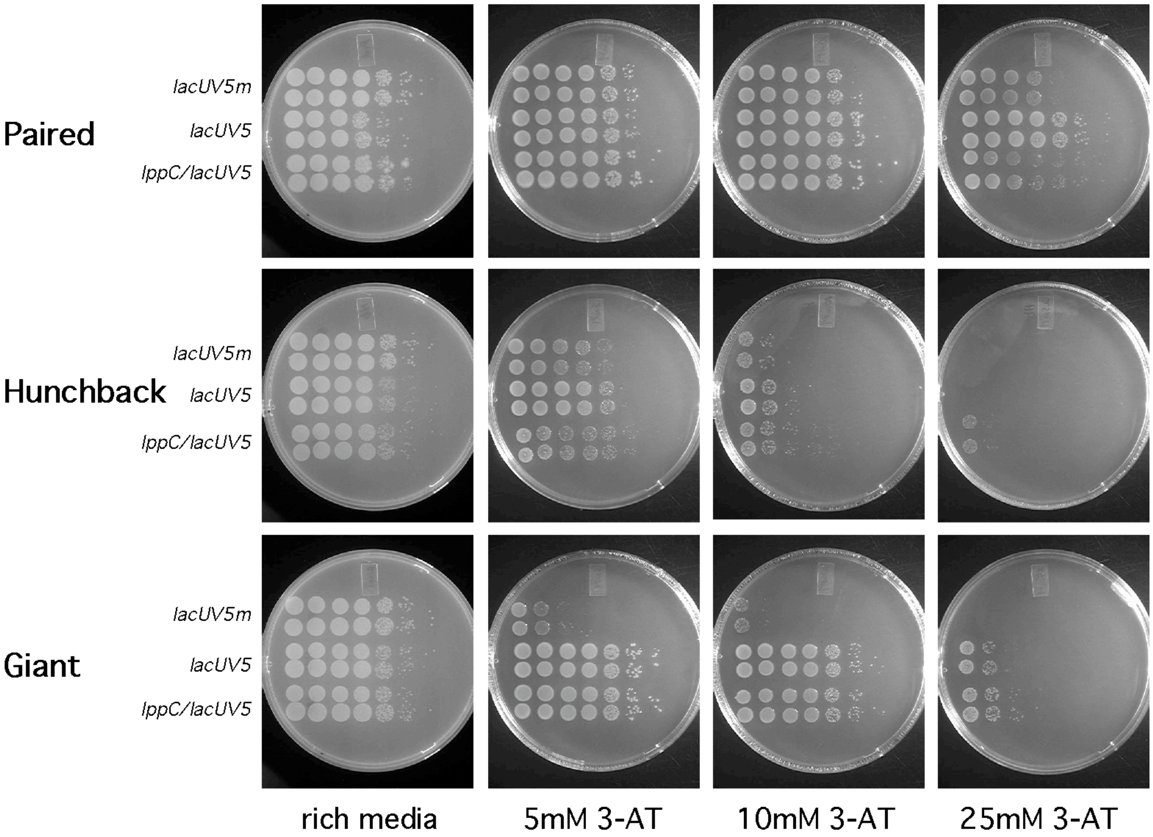

Paired
Hunchback
Giant
lacUV5m
lacUV5
lppC/lacUV5
lacUV5m
lacUV5
lppC/lacUV5
lacUV5m
lacUV5
lppC/lacUV5
rich media
5mM 3-AT
10mM 3-AT
25mM 3-AT

Figure 2.4. Comparison of the activity of three different factors at three different promoter strengths. The activity of three different TFs (Paired, Hunchback and Giant) that represent three different families of DNA-binding domains (a Paired motif, a Cys_2His_2 zinc finger protein, and basic leucine zipper motif, respectively) were characterized on consensus binding sites for each factor at three different promoter strengths. The optimal window of activity varied depending on the factor. Paired displays good activity when expressed from all three promoters (nearly 100% at 10 mM 3-AT), with the *lacUV5* promoter slightly superior to the other two. Hunchback displayed weaker activity than Paired at all three promoter strengths, with the strongest activity again with the *lacUV5* promoter. Giant displayed almost no activity when expressed under the *lacUV5m* promoter, which was optimal for Zif268, but strong activity when expressed from the other two promoters, with the dual promoter perhaps slightly superior. The reporter pH3U3 plasmids used in each case had the consensus binding site for each factor positioned at the most highly represented position relative to the -35 box based on B1H binding site selections. The activity of the omega-TF hybrids with these reporters was examined on minimal media plates containing 3 different concentrations of 3-AT where the 2xYT rich media plate serves as a control for the number of cells plated in each set. Each combination of expression plasmid and reporter plasmid was plated in duplicate, where each spot represents a 10-fold serial dilution of cells from left to right.

Figure 2.5

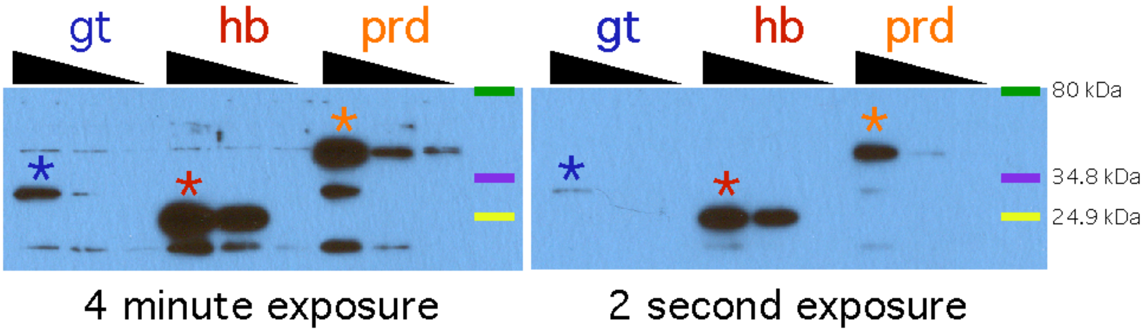
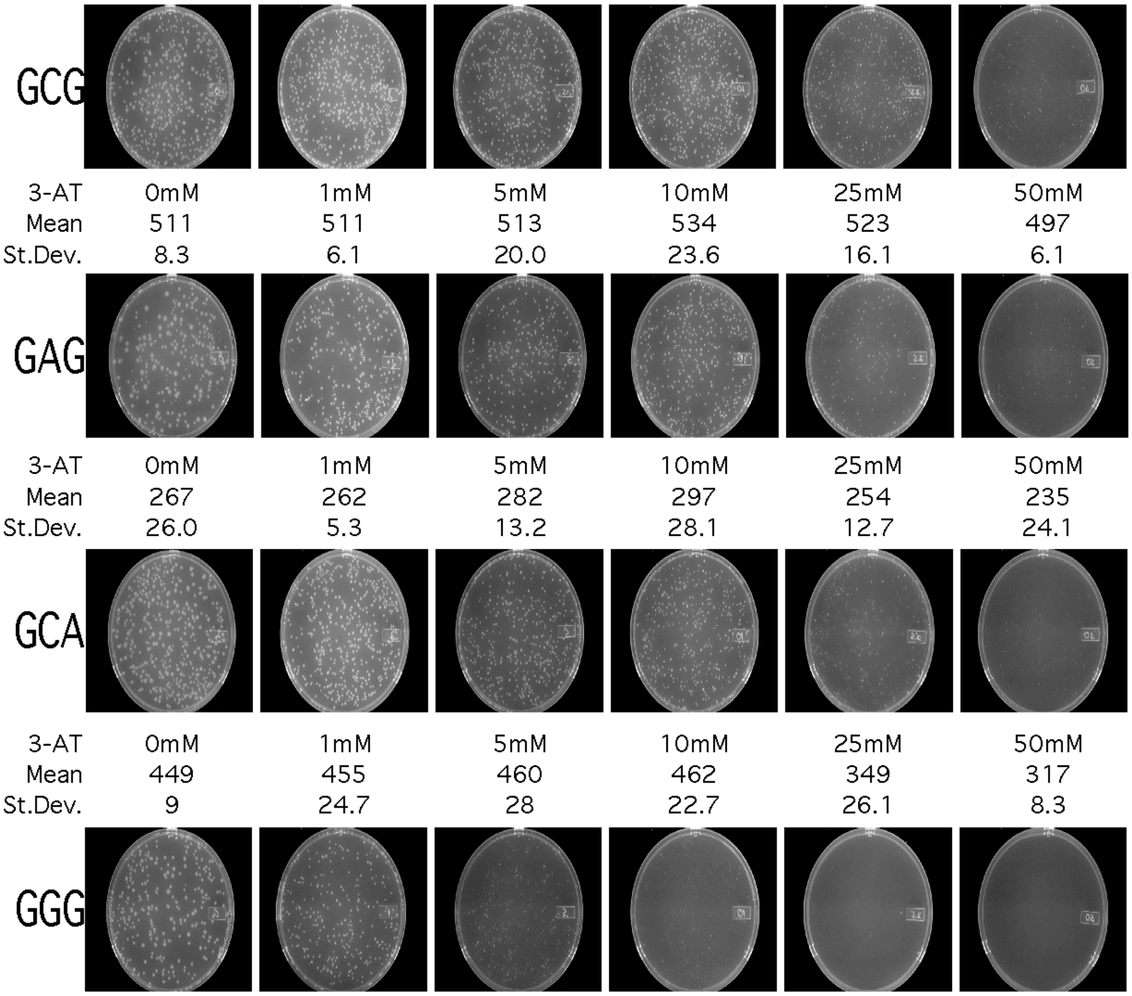

gt
hb
prd
*
*
*
gt
hb
prd
*
*
*
80 kDa
34.8 kDa
24.9 kDa
4 minute exposure
2 second exposure

Figure 2.5. Western blots of 3 ω-TF's at 3 different promoter strengths. Three factors, Prd, Hb, and Gt, each at 3 different promoter strengths were grown under inducing conditions (10 μM IPTG). For each lane on the gel a normalized amount of each bacterial culture was created based on its OD_{600} and these cells were solubilized in SDS-loading buffer. A identical fraction of this sample was run on an SDS gel and the flag tag in each construct was visualized by Western blot. Each factor is shown from left to right with decreasing promoter strength (*lppC*, *lacUV5*, and *lacUV5m*). The approximate molecular weights of the omega-TF constructs are: Prd = 41.58kDa, Hb = 25.74kDa, Gt = 29.7kDa.

Figure 2.6a



GCG

3-AT	0mM	1mM	5mM	10mM	25mM	50mM
Mean	511	511	513	534	523	497
St.Dev.	8.3	6.1	20.0	23.6	16.1	6.1

GAG

3-AT	0mM	1mM	5mM	10mM	25mM	50mM
Mean	267	262	282	297	254	235
St.Dev.	26.0	5.3	13.2	28.1	12.7	24.1

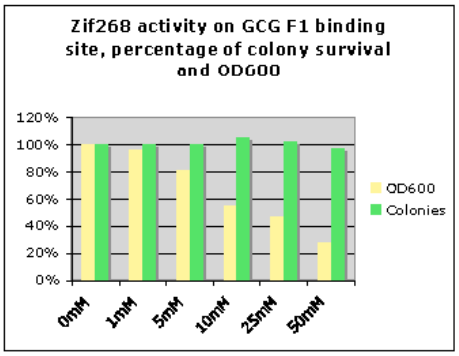
GCA

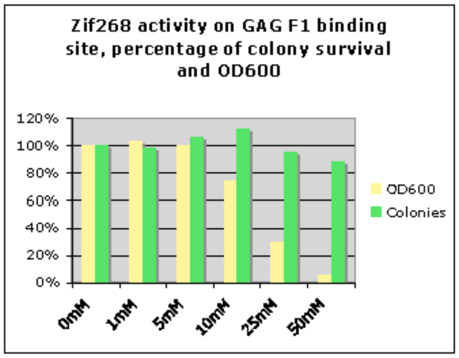
3-AT	0mM	1mM	5mM	10mM	25mM	50mM
Mean	449	455	460	462	349	317
St.Dev.	9	24.7	28	22.7	26.1	8.3

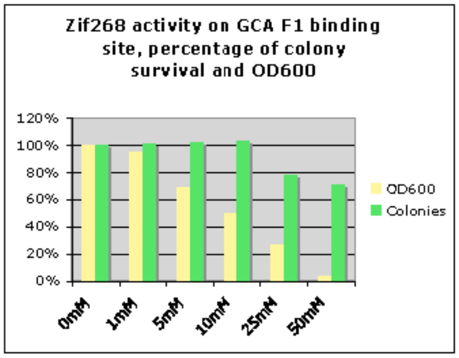
GGG

3-AT	0mM	1mM	5mM	10mM	25mM	50mM
Mean	303	304	281	261	17	0
St.Dev.	20.1	11.2	10.3	24.8	6	0

Figure 2.6b


Zif268 activity on GCG F1 binding site, percentage of colony survival and OD600
120%
100%
80%
60%
40%
20%
0%
0mM
1mM
5mM
10mM
25mM
50mM
OD600
Colonies


Zif268 activity on GAG F1 binding site, percentage of colony survival and OD600
120%
100%
80%
60%
40%
20%
0%
0mM
1mM
5mM
10mM
25mM
50mM
OD600
Colonies


Zif268 activity on GCA F1 binding site, percentage of colony survival and OD600
120%
100%
80%
60%
40%
20%
0%
0mM
1mM
5mM
10mM
25mM
50mM
OD600
Colonies

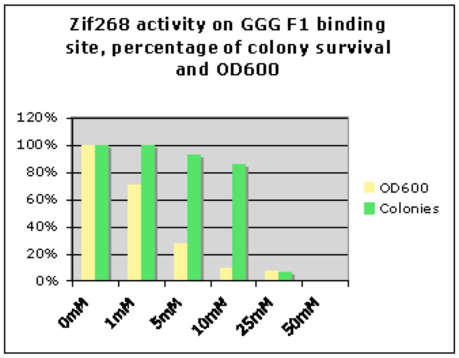

Zif268 activity on GGG F1 binding site, percentage of colony survival and OD600
120%
100%
80%
60%
40%
20%
0%
0mM
1mM
5mM
10mM
25mM
50mM
OD600
Colonies

Figure 2.6. Comparison of ω-Zif268 activity with 4 different binding sites. A. Cells harboring the pB1H2ω2-Zif268 plasmid and one of 4 different pH3U3 plasmids were challenged at several 3-AT concentrations. The pH3U3 plasmids contained one of 4 Zif268 binding sites (GCGTGG**GCG**) placed 10 bp upstream of the -35 box with modifications in the finger 1 binding sequence. The finger 1 sequence is listed to the left of each row and their previously determined relative K_ds are GAG=4.9-fold over GCG, GCA=4.7-fold over GCG, and GGG=20.3-fold over GCG. Cells were plated in triplicate and a representative plate is displayed here with the mean and standard deviation colony count is listed below. B. A histogram displays the percentage of colonies that survive at each 3-AT concentration in comparison to non-selective conditions (0mM 3-AT) in green. In yellow is the percentage OD_{600} that was measured from the cells pooled in liquid 2xYT media from each plate of different stringency.

Based on this analysis, a new 28 bp randomized binding site library was constructed to complement the omega-based B1H system. This library encompasses the region 8 to 35 bp upstream of the -35 box of the promoter, which spans these two peaks of activity. The constructed library contains ~2 x 10^8 unique clones, and though this clearly does not cover the hypothetical complexity of a randomized 28 bp region (7.2 x 10^{16}), it should encode nearly all of the possible 12 bp sites in each frame of the binding site window. Like the previously described 18 bp randomized library (Meng et al., 2005), a small proportion of the 28 bp randomized library consists of self-activating sequences. The population of self-activating sequences was reduced by performing a URA3 based 5-FOA counter-selection (Meng et al., 2006), which reduced the number of self-activating sequences to 1 in 10^6 at a typical selection stringency (10 mM 3-AT).

The utility of the 28 bp library in the omega-B1H system was assessed by determining the DNA-binding specificity of three well-characterized DNA-binding domains: Zif268, Mig1 and Rap1. Complementary sequences within the 28 bp randomized library to each TF were isolated under selective conditions (10 mM 3-AT). Each factor yielded a significant increase (>10-fold) in the number of surviving colonies over background when compared to a negative control (omega without a fusion partner). The recognition motif for each factor generated from approximately 20 selected sequences matches well with previously described specificities for these factors (Figure 2.8). Thus, the omega-based B1H system and the new 28-bp binding site library can be used to rapidly determine the DNA-binding specificity of a TF.

Figure 2.7

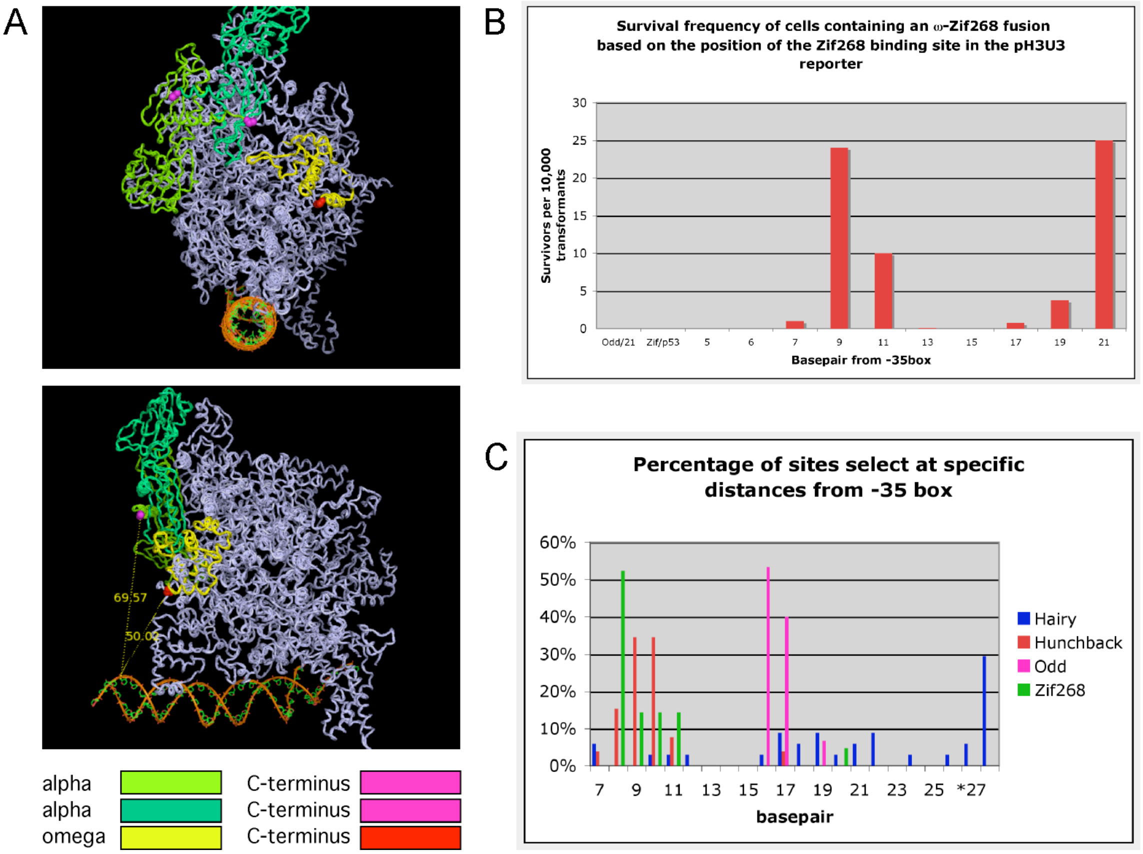

A
69.57
50.02
alpha
alpha
omega
C-terminus
C-terminus
C-terminus
B
Survival frequency of cells containing an ω-Zif268 fusion based on the position of the Zif268 binding site in the pH3U3 reporter
Survivors per 10,000 transformants
30
25
20
15
10
5
0
Odd/21
Zif/p53
5
6
7
9
11
13
15
17
19
21
Basepair from -35box
C
Percentage of sites select at specific distances from -35 box
60%
50%
40%
30%
20%
10%
0%
7
9
11
13
15
17
19
21
23
25
*27
basepair
Hairy
Hunchback
Odd
Zif268

Figure 2.7. The omega subunit and the positions utilized by omega-TF hybrids. A. The structure of RNA Polymerase bound to DNA. Top. An upstream view of the RNAP shows the relative positions of the C-termini of both alpha subunits (green and avocado) and the omega subunit (yellow). The position of the omega subunit is rotated about the DNA by about 30 degress relative to the nearest alpha subunit. Bottom. Measurements from the nearest alpha subunit and omega subunit C-termini to the DNA roughly one turn upstream of the -35 box reveals that the omega subunit is approximately 20Å closer to the DNA. B. The survival frequency of cells under stringent conditions harboring the pB1H2ω2-Zif268 plasmid and a reporter pH3U3 plasmid with the Zif268 consensus binding site placed at various distances from the -35 box reveals two clear peaks of activity at approximately 10 and 21 bp upstream. Ostermeier and colleagues observed a peak of activity for a site positioned 14 bp away from -35 box of the alpha-B1H system, which could be the result of differences in the positions of the polymerase subunits relative to the DNA (Durai et al., 2006) C. A comparison of the binding site positions of four factors characterized with the omega-B1H system. Some factors have a clear preference for a specific distance from the -35 box (Zif268, Hb, and Odd) and these distances differ from one factor to another (Hb sequences are centered 9 or 10 bp from the -35 box while Odd sequences are centered almost exclusively 16 or 17 bp away). There are other factors such as Hairy that utilize almost the entire library window. (*sequences that came through 27 and 28 basepair from the -35 box are utilizing a fixed region of the pH3U3 plasmid and could be a reason for the peak here. These sequences were not used for the construction of the Hairy logo).

Figure 2.8

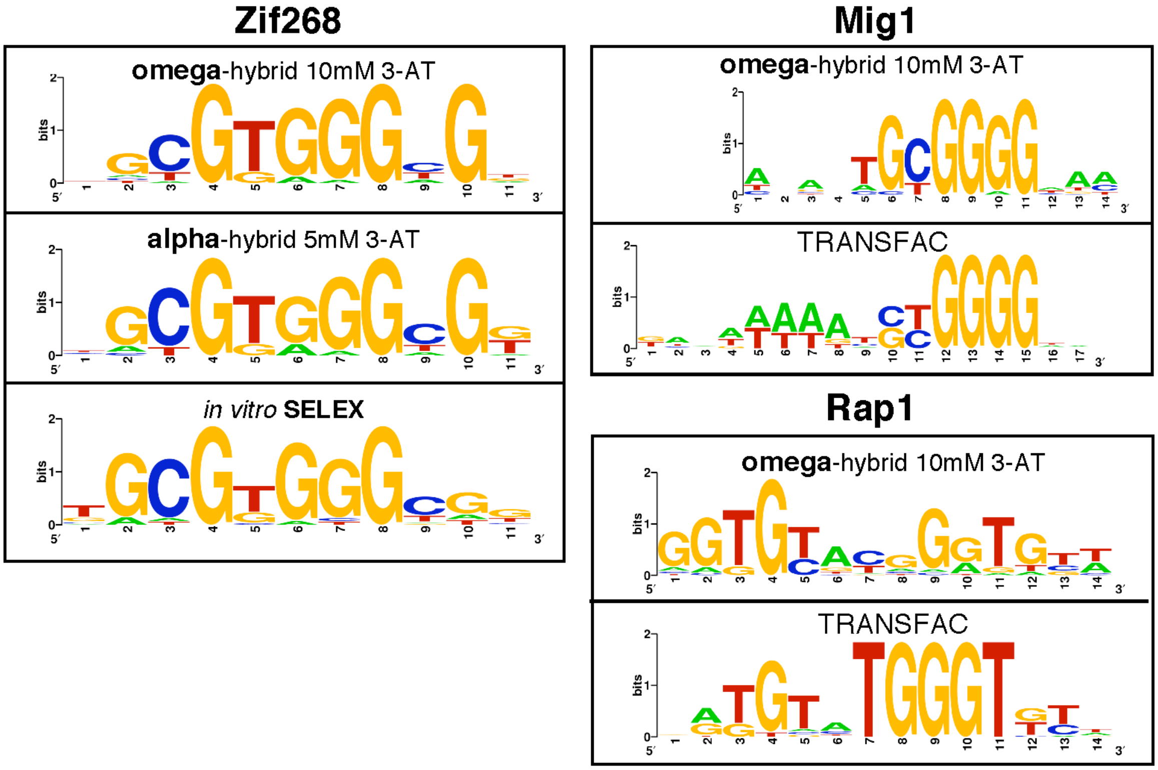

Zif268
omega-hybrid 10mM 3-AT
alpha-hybrid 5mM 3-AT
in vitro SELEX
Mig1
omega-hybrid 10mM 3-AT
TRANSFAC
Rap1
omega-hybrid 10mM 3-AT
TRANSFAC
bits
0
1
2
5′
3′

Figure 2.8. Comparison of Zif268, Mig1 and Rap1 logos. The DNA-binding domains of all three factors were expressed (*lacUV5m*) as C-terminal fusions to the omega subunit. Approximately 20 reporters were sequenced from each binding site selection and binding sites were determined by identifying overrepresented motifs within these sequences using MEME (Bailey and Elkan, 1994). The logos identified for each factor as an omega-TF fusion are compared to logos generated for these factors by other methods. The TRANSFAC logos were generated from the available PWMs at TRANSFAC (Matys et al., 2003). The Zif268 logo is compared to the logos from characterization as an alpha-Zif268 hybrid (Meng et al., 2005) and from *in vitro* SELEX data (Wolfe et al., 1999).

Despite these successes, binding site selections with Engrailed, a member of the homeodomain family, did not yield a motif. Since homeodomains represent the second largest family of TFs in the genomes of higher eukaryotes (Tupler et al., 2001), this failure represented a significant limitation for a comprehensive analysis of factor specificities. Because Engrailed recognizes a modest 6 bp element, we reasoned that competitive binding to the thousands of perfect recognition sequences that are present in the *E. coli* genome could act as a sink reducing the free pool of omega-Engrailed fusions. To increase the specificity of the homeodomain fusion protein we found inspiration in the studies by Pabo and colleagues, which demonstrated that increases in specificity are obtained when two DNA-binding domains are joined by a short linker (Klemm and Pabo, 1996). They developed synthetic a transcription factor - ZFHD1, a chimera of fingers 1 and 2 of Zif268 (Zif12) and the homeodomain of Oct1. This fusion displays superior specificity to either of its component parts (Pomerantz et al., 1995). Correspondingly, B1H binding site selections using an omega-Zif12-Oct1 hybrid yielded a pair of motifs consistent with the specificity of both DNA-binding domains (Figure 2.9). Thus, the omega-Zif12 scaffold can be used to select binding sites for homeodomains and presumably other factors with smaller recognition sequences. Since the Zif12 component would be constant in all selections using this scaffold, a complementary randomized library was constructed with the recognition motif for the Zif12 module positioned neighboring a 10 bp randomized library (Figure 2.1a). The "ZF10" library contains ~8 x 10^6 unique clones and is of adequate length to encompass the recognition sequence of a standard homeodomain with extra sequence diversity (unselected flanking sequence) to

Figure 2.9

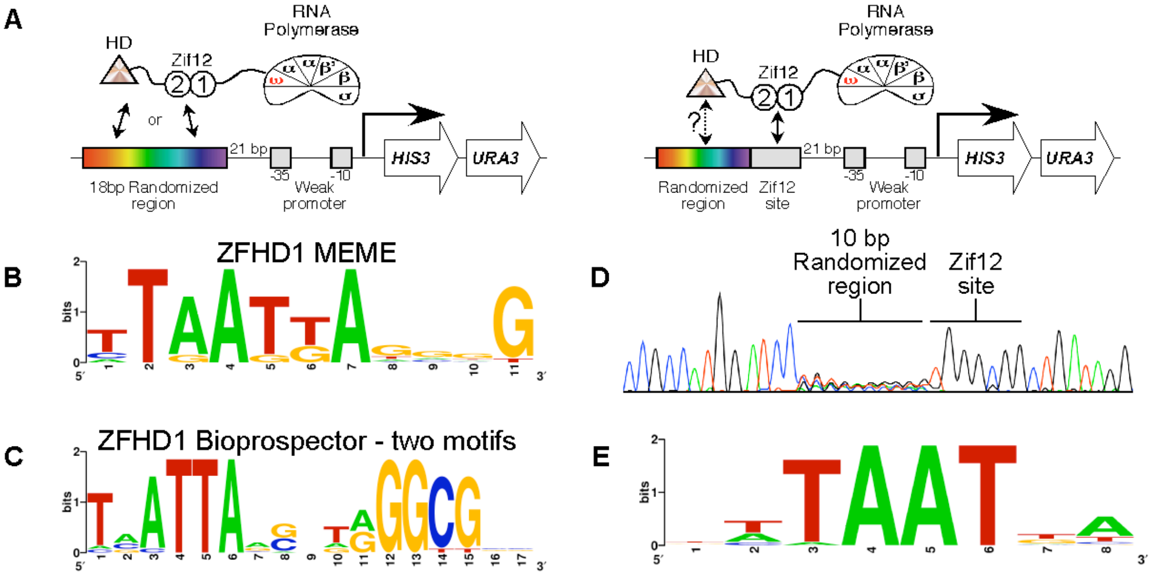

A
HD
Zif12
2
1
RNA
Polymerase
α
α
β'
β
σ
ω
or
21 bp
-35
-10
HIS3
URA3
18bp Randomized
region
Weak
promoter
HD
Zif12
2
1
RNA
Polymerase
?
21 bp
-35
-10
HIS3
URA3
Randomized
region
Zif12
site
Weak
promoter
B
ZFHD1 MEME
bits
5′
3′
C
ZFHD1 Bioprospector - two motifs
bits
5′
3′
D
10 bp
Randomized
region
Zif12
site
E
bits
5′
3′

Figure 2.9. The ZFHD1 model. A. (left) Cartoon depicting the omega-ZFHD1 construct interaction with a reporter with a 18 bp randomized window and (right) one that has a fixed binding site for the Zif12 DBD. B and C. The logos resulting from the ZFHD1 selection using the 28 bp library as determined by MEME (B. searches for one motif) and by BioProspector (Liu et al., 2001) (C. searching for 2 separate motifs). D. Chromatogram of the ZF10 library, the 10bp randomized region and the Zif12 binding sites are indicated. E. The logo resulting from the ZFHD1 selection on the ZF10 library. $5x10^7$ bacteria containing the ZF-10 library and the omega-Zif12-Oct1 expression vector were selected on minimal medium lacking histidine and containing 10 mM 3-AT. Approximately 2000 colonies survived the selection, which represented a > 100-fold increase over the number of surviving clones in the omega-Zif12 negative control. MEME analysis of 22 unique sequenced clones recovered a motif consistent with the specificity of Oct1 from 22 of 22 sequences (Verrijzer et al., 1992).

allow unique clones containing the same core binding site to be identified. Selections with omega-Zif12-Oct1 and the ZF10 library yielded a motif consistent with the specificity of Oct1 ((Verrijzer et al., 1992); Figure 2.9).

Large-scale analysis of *D. melanogaster* TFs

To demonstrate that this technology is sufficiently rapid and simple to perform a comprehensive characterization of the TFs we focused on two groups of factors from the *D. melanogaster* genome: all of the factors from a certain DNA-biding domain family (84 homeodomains, *see Chapter 4*) and all of the factors in a common regulatory network (36 A-P embryonic patterning, see *Chapter 3*). The former set provides a survey of the breadth of specificities that are observed for a particular family, while the later provides a basis set for testing the utility of B1H-generated PWMs for the prediction of CRMs throughout the genome.

Discussion

We have developed an omega-based B1H system that allows the high throughput determination of TF DNA-binding specificity. This system has several advantages over other techniques for characterizing DNA-binding specificity. First, the use of *E. coli* as our platform allows the isolation of complementary TF - binding site combination *in vivo* in a single round of selection using relatively simple techniques. Because *E. coli* demonstrate an extremely high transformation efficiency, randomized binding site libraries with complexity greater than 10^8 members can be utilized. Perhaps the greatest advantage realized by this system is the flexibility provided by utilizing omega-TF hybrids, as the absence of competition from endogenous omega has resulted in an extremely sensitive selection system with a much greater dynamic range than previous systems (Durai et al., 2006; Meng et al., 2005). This sensitivity has allowed us to successfully characterize TFs that failed to generate motifs in the alpha-based B1H system.

Using this system we have determined recognition motifs for ~14% of the predicted *D. melanogaster* TFs. For comparison the FlyREG database contains motifs for 53 TFs constructed from 5 or more identified binding sites (Bergman et al., 2005); thus our database doubles the number of specificities that are available, and in cases where these databases overlap, our data is typically of higher quality. The rate of successful TF characterization within this system (101 of 102) makes it amenable to perform comprehensive surveys of TF specificity in complex organisms: once cloned, ten or more factors can be analyzed in parallel in the B1H system in a matter of days. Our current dataset is focused primarily on monomeric DNA-binding domains, but also includes homodimers and heterodimers. This reductionist approach overlooks the potential for sets of factors to cooperatively recognize motifs that are not a simple composite sites formed from their individual motifs, such as the Exd-Hox combinations that play critical roles in specification during development (Pearson et al., 2005; Ryoo and Mann, 1999; Wilson and Desplan, 1999). These types of combinations can potentially be characterized using the B1H system, as complementary vectors for the characterization of heterodimers have been developed (Meng et al., 2005; Meng and Wolfe, 2006). However, some criteria for choosing sets of factors to be evaluated must be applied because of the combinatorial issues involved with testing all possible pair-wise combinations.

Experimental Procedures

Construction and genotype of the *ΔrpoZ* Selection Strain

The omega selection strain was created by knocking out the *rpoZ* gene from the selection strain used with the alpha-based B1H system (Meng et al., 2006): SB3930 *lac-,* Δ*hisB*463, Δ*pyrF* [F′ *proAB lacI*^{q}*Z*□*M15* Tn*10* (TetR)]. Gene inactivation was accomplished using the method detailed by Datse55nko and Wanner. To avoid using the Kanamycin or Chloramphenicol resistance markers for gene knock-outs that was developed by Datsenko and colleagues (Datsenko and Wanner, 2000), which are already utilized in the B1H system plasmids, we created a Zeocin resistant version of the pKD4 plasmid. This template has the Zeocin resistance cassette flanked by FRT sites with the P1 and P2 priming sequences 5’ and 3’, respectively, to the FRT flanked cassette. The following primers were used to amplify both the Zeocin and the Kanamycin recombination cassette using the P1 and P2 sites:

rpoZ 5’ (genomic homology italics, P1 site is bold)

5’*ATGCCCAGTCATTTCTTCACCTGTGGAGCTTTTTAAGTATGGCACGCGTAACTGTTCA***TGTGTAGGCTGGAGCTGCTTCG**-3’

rpoZ 3’ (genomic homology italics, P2 site is bold)

5’*ACAAGGGCGACCCGCTTTGTGATTAACGACGACCTTCAGCAATAGCGGTAACGGC***CATATGAATATCCTCCTTAGTTCCT**-3’

The resulting PCR products were introduced as recombination donors in conjunction with lambda red recombinase to inactivate the *rpoZ* locus. The resistance marker was removed following isolation of a confirmed knockout strain by FLP recombinase. PCR primers were designed to prime approximately 100 basepairs upstream and downstream from the *rpoZ* locus to confirm insertion of the resistance marker and its removal from this locus.

rpoZ 5’ PCR primer

5’-GCAGCGTCATGACGCTTTAA-3’

rpoZ 3’ PCR primer

5’-GATTTGGTCTTCCGGCAGG-3’

Following amplification of this locus, the PCR products were run on a 1.5% agarose gel to confirm the change in mobility of the products that should be associated with insertion of antibiotic resistance marker and their removal (Supplementary figure 1). These PCR products were sequenced to verify *rpoZ* replacement and deletion. The resulting strain containing the Zeocin insertion into the *rpoZ* gene, SB3930 *lac-,* Δ*hisB*463, Δ*pyrF,* Δ*rpoZ::*Zeo [F′ *proAB lacI*^{q}*Z*□*M15* Tn*10* (TetR)], was used in all of the B1H selections because it provides an additional selectable marker for identification and maintenance of the selection strain that is orthogonal to all of the plasmid-based markers and it demonstrated the same activity when compared to the complete *rpoZ* knockout.

Omega Constructs

Direct Omega fusions

The coding sequences for the wt omega subunit contains a BamHI and KpnI site, both of which are potential cloning candidates in our expression vectors. Therefore, the following primers were used to PCR amplify the omega subunit from *E. coli* genomic DNA in two fragments that would remove these target sequences.

Omega 5' (introduces a NcoI site at 5' end of gene)

5'-GCGGAATTCCATGGCACGCGTAACTGTTCAG-3'

Omega int 3' (introduces two silent mutations, removing restriction sites)

5'-TTCTTCCGGTACGAGCGGGTCCTTTCCGCC-3'

Omega int 5' (compliment to Omega int 5')

5'-GGCGGAAAGGACCCGCTCGTACCGGAAGAA-3'

Omega 3' (removes R91 and stop codon, introduces NotI site)

5'-TGCGCGGCCGCACGACCTTCAGCAATAGCGGT-3'

The first fragment was amplified using Omega 5' and Omega int 3'. This fragment introduced an NcoI site at the 5' end of the gene which contains the ATG start site that will be used for expression of the omega-TF hybrid. The second fragment was amplified with Omega int 5' and Omega 3' which removed the stop codon and the last residue

(R91) of the omega protein while introducing a NotI site. The two internal primers introduced two silent mutations that removed the BamHI and KpnI target sites. The final PCR was done using an equal molar mix of the first two fragments as templates for the external Omega 5' and Omega 3' primers. The resulting PCR product was recovered and digested with NcoI and NotI. The digestion product was ligated into the previously described pB1H2 expression plasmid (Meng et al., 2005). This resulted in a plasmid that could express omega-TF hybrids with exactly the same promoter and linker that had been used in the alpha system. This plasmid for expressing a transcription factor as a fusion to the C-terminus of omega was named pB1H2ωL where the L signifies the *lppC lacUV5* dual promoter that is driving expression of the omega-TF hybrid (Hu et al., 2000). The pB1H2ω2 and pB1H2ω5 expression plasmids were created by replacing a fragment of the pB1H2ωL plasmid that contains the promoter (EcoRI to NcoI) with a fragment that contains either the *lacUV5* promoter (pB1H2ω5) or a mutant version of the *lacUV5* promoter (*lacUV5m*) that has two mutations in the -10 box (pB1H2ω2). Once the three primary expression plasmids (pB1H2ω2, pB1H2ω5, and pB1H2ωL) had been constructed, a TF could be introduced into any of these plasmids for expression by designing a KpnI site in frame into the 5' primer and a stop codon and XbaI site into the 3' primer used to amplify the TF from a DNA template (genomic DNA or cDNA). The PCR product could then be digested with KpnI and XbaI and ligated into the pB1H2ω backbone DNA with the desired promoter strength (Supplementary figure 2). This universal cloning strategy allowed a TF to be easily moved between the plasmids by sub-cloning to examine the impact of changing the promoter strength. The linker between the omega-subunit and the TF was identical to the linker previously described in pB1H2, which contains a flag-tag allowing the expression of a TF to be verified by Western blot.

Omega-Zif12 fusions (Homeodomain constructs)

Homeodomains were expressed as omega fusions in combination with fingers one and two of Zif268 (Zif12) under control of the *lacUV5m* promoter plasmid (pB1H2ω2-12HD; Supplementary Figure 2). The KpnI site at the 5' end of TF construct (the beginning of Zif268 finger 1) was inactivated by converting the sequence from GGTACC to GGCACG, both sequences coding for Gly, Thr. A new KpnI site was then created 3' to Thr codon that is the first amino acid of the linker between fingers 2 and 3 of Zif268. Two additional amino acids were added after the Kpn1 site. The first amino acid was always glycine. In the majority of the homeodomains, the second amino acid was the -1 amino acid of the specific homeodomain being assayed, however for a subset we used the -1 residue of Oct1, arginine for purely historical reasons. The KpnI site and the inserted residues created a 5 amino acid linker between the 2nd His of Zif268 finger 2 and the beginning of the HD (Zif12-TGTGN-HD). Each homeodomain (with the additional two residues) was cloned between the KpnI site, which encodes the first set of TG residues, and the Xba1 site downstream with a stop codon introduced just prior to the Xba1 site. The *lacUV5* version of the Zif12-HD construct (pB1H2ω5-12HD) was created by moving the entire fragment encoding the linker-Zif12-HD fragment via NotI to XbaI digestion into the corresponding region of the pB1H2ω5 plasmid. All new expression constructs were sequence verified.

Library Construction

28 Basepair Library

The 28 basepair library was constructed in pH3U3 as previously described by Meng, Brodsky and Wolfe except that it was introduced between the NotI and EcoRI sites using the following oligonucleotides:

28 bp library oligonucleotide:

5’GGCGCGAATTCGNNNNNNNNNNNNNNNNNNNNNNNNNNNNGCGGCCGCA
AGGTAGCTGATTCCGTTCTCGC-3’

Library extension primer:

5’-GCGAGAACGGAATCAGCTACCTT-3’

This library places the randomized region 7 bp away from the 5’ edge of the -35 box of the weak promoter that controls expression of the HIS3/URA3 reporter genes. The 28 bp raw library contains ~2 x 10^8 independent clones based on titration of the initial transformants following electroporation of the ligated library after a 1 hour recovery in SOC medium.

ZF10 Library

The ZF10 library was created by annealing oligonucleotides with complimentary ends to the library oligonucleotide to make duplex DNA with a gap spanning the randomized region and appropriate overhangs for cloning into the pH3U3 plasmid between the NotI and EcoRI sites upstream of the promoter controlling the HIS3 and URA3 genes.

ZF10 library oligonucleotide:

5’-GGCCGCCATGGATCCNNNNNNNNNNTGGGCGGCTGATAGGCGCGCCG-3’,

5 prime complimentary oligonucleotide:

5’-GGATCCATGGC-3’

3 prime complimentary oligonucleotide:

5’-AATTCGGCGCGCCTATCAGCCGCCCA-3’

The oligonucleotides were 5’ phosphorylated as a mixture by combining 200 pmol of each oligonucleotide in 100μl of 1xT4 polynucleotide kinase buffer (NEB) with 1mM ATP and 20 Units T4 polynucleotide kinase (NEB). This reaction was incubated at 37°C for 30 minutes and then boiled for 5 minutes before annealing by a gradual reduction in temperature to 4°C. 1 μl of the phosporylated, annealed oligonucleotides (2μM) was ligated into 1 μg of gel purified pH3U3 plasmid backbone that had been digested with NotI and EcoRI in a 30 μl reaction containing 1xT4 DNA Ligase buffer and 1 μl of T4 DNA Ligase (400 units, NEB) overnight at 16 °C. Following completion, the ligation reaction was ethanol precipitated, and the DNA pellet was resuspended in 2 μl of H_20, and transformed into 80 μl of electrocompetent XLI-Blue cells. The transformed cells were recovered in 50 ml SOC for 1 hour at 37°C. Following the recovery, a 200ul sample was titrated by 10-fold serial dilution on 2xYT plates containing 25 μg/ml Kanamycin to determine the total number of transformants. The number of transformants in a dilution normalized to the fraction of library culture should reflect the constructed library size. Kanamycin (25 μg/ml) was then added to the remaining culture and the cells were expanded for an additional hour at 37°C. After expansion, the cells were plated on

10 large 2xYT plates (150mm round) containing 25 μg/ml kanamycin and grown overnight at 37 °C. 200μl of the culture was again titrated by 10-fold dilutions on 2xYT plates containing kanamycin to determine the degree of expansion that occurred during the additional hour of growth. After these large plates had grown overnight, cells were harvested from these plates by resuspending the colonies in 10ml 2xYT per plate. The resuspensions were pooled and cells pelleted by centrifugation for 15 minutes at 3000 rpm. The plasmid DNA was recovered from this pellet by scaling 20-fold the procedure for DNA isolation using the QIAGEN plasmid Miniprep Kit.

Counterselection of the libraries

Counterselections were performed on each library to remove self-activating sequences. 250 ng (28bp) or 1 μg (ZF10) of raw library material was transformed into 80 μl of the *rpoZ* positive version of the selection strain (US0 Δ*hisB*, Δ*pyrF*). These cells were recovered in SOC for 1 hour at 37°C while rotating. The cells were then pelleted by centrifugation for 15 minutes at 3000 rpm and the resulting pellet was resuspended in 5 ml of YM medium (a type of minimal medium). The cells were acclimated to the YM medium for 1 hour at 37°C while rotating. Following recovery the cells were pelleted, washed 2 times with YM medium (by pelleting and resuspension) and then resuspended in a final volume of 1ml YM medium. 20 μl of this final resuspension was titrated by 10-fold serial dilution on rich media plates (2xYT + 25 μg/ml Kanamycin) to determine the total transformants. The titration plates were grown overnight at 37°C while the remaining 980 μl cell resuspension was stored at 4 °C. The following day cells were counted from the rich media titrations and approximately 5x10^{6} transformants were plated on each of the 150mm round YM plates containing 2.5mM 5-FOA (27 plates used for the 28bp library, 10 plates for the ZF10 library). The plates were wrapped with parafilm and incubated at 37°C for 24 to 36 hours. The cells were then harvested from the plates and the plasmid DNA recovered as described for the ZF10 library construction. A second round of counterselection was performed on the 28 basepair library to further reduce the background of self-activating clones, which remained higher than desired following the first counterseleciton. This additional counterselection was performed as described above except that approximately 2.5x10^{8} transformed cells were split over 20 5-FOA/YM plates.

Selection Overview (Figure 2.1)

Two different libraries were constructed to assess the specificity of each TF: a 28 bp randomized library for general use in characterizing factors, and a 10 bp library containing a neighboring binding site for fingers 1 and 2 of Zif268 (ZF10), which is for characterizing factors with low affinity or specificity. The 28 bp library, which was constructed for the characterization of direct omega-TF fusions has a relatively high background of false positive clones when compared to the previously described 18 pair library. The majority of this background results from a population of self-activating sequences in the randomized region that are coupled to a deactivated *URA3* reporter gene that allows these clones to persist even under stringent counterselection. This subpopulation is sensitive to the absence of uracil during the positive selection of binding sites, and consequently uracil was omitted from binding site selections using this library in the description of general procedure that follows. The false positive rate within the ZF10 library following counterselection was quite low and did not change significantly when uracil was omitted, consequently all selections using the ZF10 library were performed in the presence of uracil (200 μM uracil was added to each NM selection plate), which provides slightly greater growth rates under selective conditions.

General selection procedure (Figure 2.1c)

Approximately 2 μg of the bait plasmid (pB1H2ω2/5/L or pB1H2ω2/5-12HD) and 50 ng of the library plasmid were electroporated into 80 μl of the selection strain. The cells and the two plasmids were mixed on ice and moved to a pre-chilled 1 mm cuvette. The cell and plasmid suspension was electroporated at 4°C and immediately resuspended in 10 ml pre-warmed SOC. The cells were then recovered while rotating at 37 °C for 1 hour. Next, the cells were pelleted by centrifugation at 3000 rpm for 15 minutes and resuspended in 5 ml NM medium that was supplemented with 200 μM uracil, and 0.1% histidine. These cells were acclimated to the NM medium while rotating for 1 hour at 37°C. Cells were pelleted, washed 4 times in NM medium (no supplement) by sequential pelleting and resuspension and then resuspended to a final volume of 1 ml NM medium. 20 μl of this final resuspention was titrated by 10-fold serial dilutions on rich media plates (2xYT + 25 μg/ml Kanamycin, and 100 μg/ml Carbenicillin) to determine the total number of transformants. The titration plates were grown overnight at 37 °C while the remaining 980 μl cell culture was stored at 4 °C. The following day cell counts were determined from the rich media titrations and based on the total number of transformants, between 1x10^7 and 1x10^8 cells were placed on one 5 mM 3-AT and one 10mM 3-AT NM selective plate (150mm diameter rounds). Cells were spread on the plates with sterile glass beads and allowed to dry. The plates were then wrapped individually with parafilm and grown at 37°C for 36 to 48 hours. Typically five to fifteen binding site selections were performed in parallel with positive and negative controls included to allow a qualitative assessment of the success of each experiment. Surviving colonies on each plate were counted and the fraction of surviving clones was determined based on the number of cells that were plated. A 10-fold increase in the fraction of surviving clones compared to the negative control (omega without a tethered TF) was highly correlated with a successful selection. Selections with ratios lower than a 10-fold increase were successful in many instances, but in this range there was more variability in the number of sequenced clones that contained binding sites.

Failed Selections

A low fold increase (or no increase) in the number of surviving clones on selective media relative to the background when normalized to the number of cells plated was a reliable indicator of a failed selection. This outcome was predominantly the result of improper expression of the TF (too low or too high). Factors that had too high an expression level displayed toxicity on rich media plates (reduced colony number relative to selections performed with other factors in parallel and/or extreme variability in colony size), which allowed us to gauge whether the expression level of a TF from a failed selection should be increased or decreased. If the expression plasmid for a particular factor appeared toxic, the TF was moved to a version of the pB1H2ω plasmid that contains a weaker promoter (*lacUV5m* or *lacUV5*) and the selections were repeated. If the expression plasmid for a particular factor did not appear to be toxic to cells based on the rich media titrations but the selection failed, the TF was moved to a version of the pB1H2ω plasmid with a stronger promoter (*lacUV5* or *lppC*) and reselected. This approach proved successful in almost every example. In principle the concentration of the inducer IPTG could also be increased or decreased to further modify the expression levels, but we found adjusting the overall promoter strengths on the expression plasmid to be the most reliable path to success. In the case of one homeodomain, Eve, where our initial selections failed, we found that the removal of a small string of hydrophobic residues from the C-terminus of the protein just after the end of the homeodomain resulted in a significant improvement in activity. Hydrophobic residues at the C-terminus of a protein can lead to lower levels of functional expression in bacteria (Parsell et al., 1990). Croc is the single example of a TF that did not generate a binding site motif using the omega-based B1H system. This factor produced colony numbers following selection that were between 20 and 100-fold over the background when expressed form either a *lacUV5* and *lppC* promoter at either 5 or 10mM 3-AT, however computational analysis of the selected binding sites was unable to identify a significantly overrepresented motif within this population. It is not clear why Croc was able to achieve a high fold over the background without selecting a specific binding site; one possible explanation is recognition of a binding site present in a fixed region of the pH3U3 plasmid outside the library window, which would return sequences that were not related to sequence-specific binding.

Medium for Selection procedure

SOB medium

Purchased from Becton, Dickinson and Company, cat. No 244310. Add 28.086g SOB powder to 1L purified water. Heat while stirring. Allow to boil for 1 minute. Autoclave at 121^{o}C for 15 minutes. The SOB medium contains the following components per liter:

Typtone	20.000g
Yeast Extract	5.000g
Sodium Chloride	0.500g
Magnesium Sulfate, anhydrous	2.400g
Potassium Chloride	0.186g

SOC medium

Contains a final concentration of 0.5% filter sterilized glucose in autoclaved SOB (above).

2xYT medium

Purchased from Becton, Dickinson and Company, cat. No 244020. Add 31.0g 2xYT powder to 1L purified water. Heat while stirring. Allow to boil for 1 minute. Autoclave at 121^{o}C for 15 minutes. The 2xYT medium contains the following components per liter:

Pancreatic Digest of Casein	16.0g
Yeast Extract	10.0g
Sodium Chloride	5.0g

YM medium

Prepare the following solution: 1xM9 Salts, 4 mg/ml glucose, 200μM Uracil, 0.1% Histidine, 0.01% Yeast Extract, 1 mM MgSO4, 10 μg/ml thiamine, 10 μM ZnSO4, 100 μM CaCl2, 25ug/ml kanamycin, and 10 μM IPTG. Filter sterilize through a 0.22 μm filter.

NM medium

Prepare the following solution: 1xM9 Salts, 4 mg/ml glucose, 200 μM adenine-HCl, 1x Amino acid mixture (below), 1 mM MgSO4, 10 μg/ml thiamine, 10 μM ZnSO4, 100 μM CaCl2, 25ug/ml kanamycin, 100 μg/ml carbenicillin, and 10 μM IPTG. Filter sterilize through a 0.22 μm filter.

YM counter-selective plates

Conter-selective plates contain: 1.5% autoclaved agar, 1xM9 Salts, 4 mg/ml glucose, 200μM Uracil, 0.1% Histidine, 0.01% Yeast Extract, 1 mM MgSO4, 10 μg/ml thiamine, 10 μM ZnSO4, 100 μM CaCl2, 25ug/ml kanamycin, and 10 μM IPTG and the desired concentration of 5-FOA.

NM selective plates

Selective plates contain: 1.5% autoclaved agar, 1xM9 Salts, 4 mg/ml glucose, 200 μM adenine-HCl, 1x Amino acid mixture (below), 1 mM MgSO4, 10 μg/ml thiamine, 10 μM ZnSO4, 100 μM CaCl2, 25 μg/ml kanamycin, 100 μg/ml carbenicillin, 10 μM IPTG and the desired concentration of 3-AT.

Amino acid mixture (33.3x)

Contains 17 of the 20 amino acids, omitting His, Met, and Cys.

Prepare the following six solutions (all percentages are wt/vol):

Solution I (200x): dissolve 0.99g Phe (0.99%), 1.1g Lys (1.1%) and 2.5g Arg (2.5%) in water to a final volume of 100ml.

Solution II (200x): dissolve 0.2 g Gly (0.2%), 0.7 g Val (0.7%), 0.84 g Ala (0.84%) and 0.41 g Trp (0.41%) in water to a final volume of 100ml.

Solution III (200x): dissolve 0.71g Thr (0.71%), 8.4 g Ser (8.4%), 4.6 g Pro (4.6%) and 0.96 g Asn (0.96%) in water to a final volume of 100ml.

Solution IV (200x): add 1.04g Asp (1.04%) and 18.7g potassium-Glu (18.7%) to water, bring to a final volume of 100ml.

Solution V (200x): add 14.6g Gln (14.6%) and 0.36g Tyr (0.36%) to roughly 90ml of water. Add solution V to solution IV. Add NaOH pellets slowly until all amino acids go into solution. Bring final volume to 200ml.

Solution VI (200x): dissolve 0.79 g Ile (0.79%) and 0.77 g Leu (0.77%) in water to a final volume of 100ml.

Mix solutions I to VI together and filter sterilize through a 0.22 μm filter and store at 4C. This results in a 33.3x amino acid mixture.

Colony PCR and Sequencing

The binding sites from successful selections were recovered by PCR amplification of the corresponding pH3U3 Library window from individual surviving colonies picked from each selection plate. These PCR products were sequenced to generate the desired data for computational analysis. PCR reactions were done in a 96-well plate format where 25μl of the PCR mix (1μM HU100 primer, 1μM OK181 primer, 300μM Denville dNTP mix, 1x NEB ThermoPol Buffer and 1 unit of NEB Taq polymerase) was added to each well of the plate. For each of the 96 wells, a single colony was picked from a selection plate with an autoclaved toothpick to inoculate the 25ul PCR mix. For each set of PCRs from a given selection plate a negative control reaction (no inoculation) was run in parallel in one well to insure that the appearance of an amplified product was not due a contaminating DNA source. Once each well had been inoculated, the plate was covered with aluminum film and placed in the thermocycler. The PCR reaction initiated with a single, 2 minute denaturation step at 94^{o}C. This was followed by 35 reaction cycles consisting of one minute denaturation at 94°C, a 1.5 minute annealing step at 56°C, and a 2 minute extension at 68°C. After these 35 cycles were complete, there was a final extension for 5 minutes at 68°C. The plate was held at 4°C from that point until being removed from the block. To confirm successful PCR reactions, 5µl of each 25µl PCR reaction was run out on a 1.5% agarose gel and the mobility of product in each well was compared to a DNA ladder standard (NEB). Successful reactions were sequenced (Agencourt) using HU100 as the sequencing primer.

HU100

5’-GAAATATGTATCCGCTCATGAC-3’

OK181

5’-CCAGAGCATGTATCATATGGTCCAGAAACCC-3’

Motif discovery and alignment

The chromatogram from each sequence read was inspected for quality and accuracy, and if judged interpretable, the sequence between the NotI site and the EcoRI site (the library window) was confirmed by visual inspection of the peaks. The group of all unique library sequences recovered for a TF was then analyzed using MEME motif discovery tool (http://meme.sdsc.edu/meme/intro.html). Overrepresented motifs were recovered using the following settings in MEME. Selections employing the 28 bp library: zero or one motif per sequence could be discovered, a motif could be discovered on either strand of the DNA and the search width for a motif was set from 3 to 28 bases. Selections employing the ZF10 library were identical except that the search width for a motif was set from 3 to 10 bases. In all instance where a binding site motif was successfully identified, it was the top hit recovered from the MEME search and had an expectation value that was <e^{-5}. The aligned sequences that compose each motif recovered by MEME were used to generate a Sequence logo using the WebLogo tool (http://weblogo.berkeley.edu/logo.cgi).

Omega-TF activity assays

Target sites in pH3U3 for Zif268, Paired, Giant and Hunchback

The following binding sites were designed for the promoter strength activity assays based on the preferred consensus sequence and the preferred position of that site relative to the promoter that were determined for each factor from an omega-B1H binding site selection. For the Zif268 mutant sites, mutant target sequences were designed based on previously determined binding constants for variants of the finger 1 binding site.

Giant

5'-GGCCGCAGACCGGAG**ATTACGTAAC**TATAAGACACG-3'

Hunchback

5'-GGCCGCCTACCGGAGCGATA**CACAAAAAACA**TGCG-3'

Paired

5'-GGCCGCGAGTCTCACATAC**ATCCGTCACGCT**ACCCG-3'

Zif268-F1 wt (GCG)

5'-GGCCGCTGCGTGG**GCG**GGACG-3'

Zif268-F1 GAG

5’-GGCCGCTGCGTGG**GAG**GGACG-3’

Zif268-F1 GGG

5’-GGCCGCTGCGTGG**GGG**GGACG-3’

Zif268-F1 GCA

5’-GGCCGCTGCGTGG**GCA**GGACG-3’

These oligonucleotides were annealed to a complementary oligonucleotide that contains a 5’ AATT overhang and were truncated by four bp on the 3’ end to leave a GGCC overhang on the complementary strand. These overhangs are complimentary to the overhangs created by a NotI/EcoRI digestion of pH3U3. Each complementary pair of oligonucleotides were annealed and ligated between the Not1 and EcoR1 site of the pH3U3 plasmid. Cloned binding sites were verified by sequencing.

Zif268 activity on mutant finger 1 binding sites

100 ng of pB1H2ω2-Zif268 and 50 ng of a reporter pH3U3 plasmid containing one of four different Zif268 finger 1 binding sites (GCG, GCA, GAG, or GGG) were transformed into 80 μl of the selection strain by electroporation. The cells were recovered in 1 ml of SOC for 1 hour at 37 °C. After recovery, the cells were plated on rich media plates (2xYT + 25 μg/ml Kanamycin, and 100 μg/ml Carbenicillin) and grown overnight at 37 °C. A single colony was selected from each plate and a 5 ml 2xYT culture with Kanamycin and Carbenicillin was grown overnight at 37 °C. The following day, a second 5 ml culture in identical medium was inoculated with 50 μl of the overnight, saturated culture and grown to an OD600 of approximately 0.2 to 0.4. The cells in each culture were pelleted by centrifugation at 3000 rpm for 15 minutes. The

pellet was then resuspended in 5 ml NM media supplemented with 200 µM uracil. The cells were acclimated to NM media for 1 hour at 37 °C while rotating. This cells in this culture were pelleted by centrifugation at 3000 rpm for 15 minutes. The cells were then washed 4 times by sequential pelleting and resuspension in NM media supplemented with 200 µM uracil and finally resuspended in a volume of 1ml. 20 µl of this final resuspension was titrated by 10-fold serial dilutions on rich media plates (2xYT + 25 µg/ml Kanamycin, and 100 µg/ml Carbenicillin) to determine the total number of cells. The titration plates were grown overnight at 37 °C while the remaining 980 µl cell culture was stored at 4 °C. The following day cell counts were determined from the rich media titrations and a fraction of the culture was diluted to roughly 500 cells per 750 µl. Next, 750 µl of each diluted culture was spread on small NM selective plates with various concentrations of 3-AT. Each 3-AT concentration challenge was performed in triplicate for each binding site. Plates were wrapped in parafilm and grown at 37 °C for 36 hours. The number of surviving colonies was counted on each plate. Next, the colonies on each plate were resuspended in 5 ml of 2xYT and the cell suspension from the three duplicate plates for each 3-AT concentration were pooled. These cells were pelleted by centrifugation at 3000 rpm for 15 minutes and then the pellet was resuspended in 100 ml of 2xYT and an OD600 reading was measured in triplicate to determine the approximate number of cells recovered form each group of selection plates. For easy interpretation, the mean colony counts from each condition for each binding site were normalized by the colony count from non-selective conditions (NM medium with no 3-AT and 0.1% Histidine) to provide a fractional colony count. This fractional colony count for each binding site at each stringency describes the percentage of surviving colonies relative to the number observed at non-selective conditions (maximum survival rate). The measured OD600's were normalized by colony count to provide quantitative basis for evaluating the colony size that was recovered for each binding site at each stringency. To remove the complication of the number of colonies contributing to the OD600 at a stringency, the mean OD600 measurement at each stringency was multiplied by the fraction of the non-selective colony count divided by the colony count at a that specific stringency. This product represents OD600's at equivalent cell densities. Once the OD600's were normalized by colony count, the percentage of the OD600 from non-selective conditions (0mM 3-AT with 0.1% Histidine) was calculated for each binding site, 3-AT combination.

Paired, Giant and Hunchback Activity Assay

100 ng of pB1H2ω2, 5, and L versions of each TF was cotransformed with 50 ng of a reporter pH3U3 plasmid containing the optimal binding site for the corresponding TF into 80 μl of the selection strain by electroporation.   The cells were recovered in 1ml of SOC for 1 hour at 37 °C. After recovery, the cells were plated on rich media plates (2xYT + 25 μg/ml Kanamycin, and 100 μg/ml Carbenicillin) and grown overnight at 37°C. The following day, a single colony was selected from each plate and a 5ml 2xYT culture with Kanamycin and Carbenicillin was grown overnight at 37 °C. A second 5 ml culture in identical medium was inoculated with 50 μl of this overnight, saturated culture and grown to an OD600 of approximately 0.2 to 0.4. The cells in this culture were

pelleted by centrifugation at 3000 rpm for 15 minutes. The pellet was resuspended in 5 ml NM medium supplemented with 200 μM uracil, and 0.1% histidine. The cells were acclimated to NM media for 1 hour at 37 ºC while rotating. This culture was pelleted by centrifiugation at 3000 rpm for 15 minutes. The cells were then washed 4 times by sequential pelleting and resuspension in NM medium supplemented with 200 μM uracil and then resuspended in a final volume of 1ml. 20 μl of this final resuspention was titrated by 10-fold serial dilution on rich media plates (2xYT + 25 μg/ml Kanamycin and 100 μg/ml Carbenicillin) to determine the total number of cells. The titration plates were grown overnight at 37 ºC while the remaining 980 μl cell culture was stored at 4 ºC. The following day cell counts were determined from the rich media titrations. The cell concentration in each sample was normalized based on the titration results to achieve uniform cell densities at each cell dilution. The normalized cell cultures were then titrated by 10-fold serial dilution on rich media and NM selective plates supplemented with 200μM uracil and various 3-AT concentrations. The NM selective plates were wrapped with parafilm and grown at 37 ºC for 36 hours whereas the rich media plates were grown overnight at 37 ºC.

Western blot assay

100 ng of pB1H2ω2, 5, and L versions of each TF assayed was cotransformed with 50 ng of a reporter pH3U3 plasmid containing the optimal binding site for the corresponding

TF into 80 μl of the selection strain by electroporation. The cells were recovered in 1ml of SOC for 1 hour at 37 °C. After recovery, the cells were plated on rich media plates (2xYT + 25 μg/ml Kanamycin, and 100 μg/ml Carbenicillin) and grown overnight at 37°C. The following day, a single colony was selected from each plate and a 5ml 2xYT culture with Kanamycin and Carbenicillin was grown overnight at 37 °C. A second 5 ml culture in identical medium with the addition of 10μM IPTG was inoculated with 50 μl of this overnight, saturated culture and grown to an OD600 of approximately 0.4 to 0.6. 1.5ml of each culture was pelleted by centrifugation at 18,000g for 2 minutes. The pellets were resuspended in a volume of 1xSDS loading buffer that was normalized by the OD600 of each sample in comparison to the sample with the lowest OD600, which was resuspended in 50μl. The samples were promptly boiled for 10 minutes. After boiling, the samples were then centrifuged briefly to pellet debris and diluted 1:10 in 1xSDS loading buffer. 10μl of each of the diluted samples were loaded on to a 14% SDS-polyacrylamide gel along with 5μl of Kaleidoscope Prestained Standard (BioRad). The samples were run on the gel for approximately 2.5 hours. The gel was then transferred to 0.45 micron polyvinylidene difluoride (PVDF) membrane and the membrane was blocked overnight with gentle rocking in binding buffer (5% nonfat dry milk resuspended in wash buffer). After blocking the membrane was washed four times in wash buffer (20mM Tris, 150mM NaCl, and 0.05% Tween 20). The membrane was then labeled for 2 hours with 2.5μl of anti-FLAG per 10ml of binding buffer. The membrane was again washed four times with wash buffer. The secondary labeling was done for 2 hours with 1ml anti-murine-hc-HRP per 15ml binding buffer. The membrane was washed four times with wash buffer. Finally, the HRP was reacted using the Millipore Immobolin reagents for HRP substrates. The membrane was bathed with this Peroxide Solution/Luminol Reagent, 1:1 mix for 1 minute before exposing film. Film exposures were taken from 2 seconds to 5 minutes.

ACKNOWLEDGEMENTS

We would like to thank the Drosophila Genomics Resource Center (DGRC) for providing cDNA collections and Mark Stapleton and Susan Celniker for access to their unpublished transcription factor clone collection and Robin Smith for technical support. S.A.W. M.B.N. and X.M. were supported by NIH grants 1R01GM068110 from NIGMS and 1R21HG003721 from NHGRI, A.W. was supported in part by NIH grant 1R21HG003721 from NHGRI. M.H.B. and A.W. were supported in part by a New Scholar in Aging Award from the Ellison Medical Foundation and American Cancer Society grant RSG-05-026-01-CCG.

CHAPTER III: THE SYSTEMATIC CHARACTERIZATION OF DNA-BINDING SPECIFICITIES FOR FACTORS INVOLVED IN THE ANTERIOR – POSTERIOR SEGMENTATION OF THE *DROSOPHILA* EMBRYO

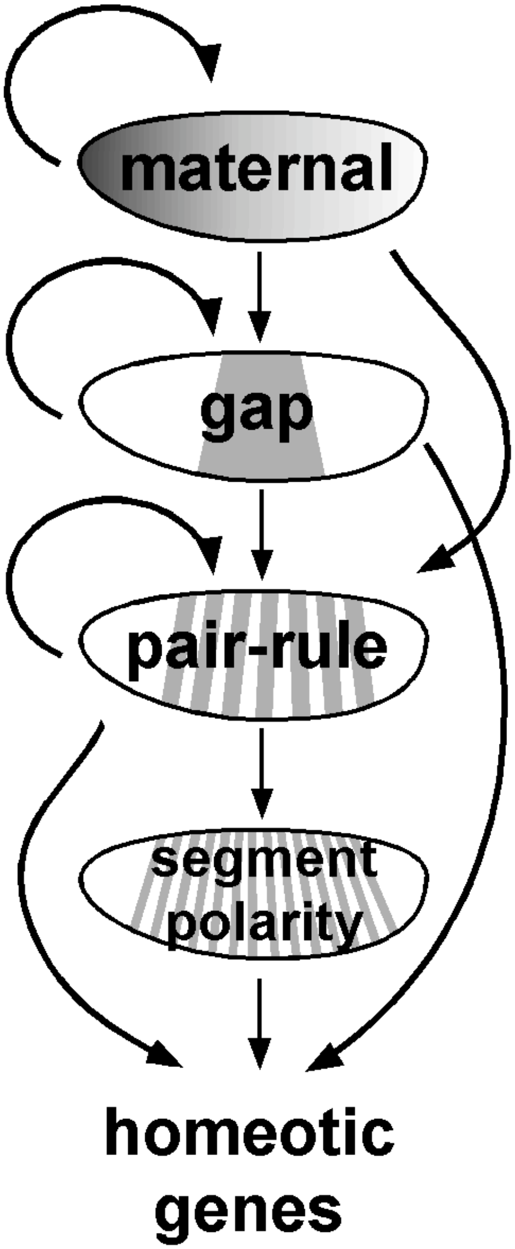
Introduction

The identification of *cis*-regulatory sequences throughout the genome and the complementary sequence-specific trans-acting factors that bind within these modules is an important step in deciphering the mechanism of spatial and temporal gene regulation in metazoans. The majority of sequence-specific transcription factors (TFs) in a eukaryotic genome can be readily identified by sequence homology to previously identified families of DNA-binding domains, where complex organisms usually contain a higher proportion of TFs (~5 to 10%) due to the requirement for more elaborate transcriptional regulatory networks (Levine and Tjian, 2003). However, identifying *cis*-regulatory modules (CRMs) within a genome is difficult due to the more dynamic nature of these sequences relative to coding sequences (Ludwig et al., 1998) and the fact that the vast majority of DNA in higher eukaryotes is non-coding sequence (Siepel et al., 2005).

Biochemical and computational methods for the identification of CRMs within the genome have been developed, yet limitations remain. Biochemical methods based on ChIP-chip (Harbison et al., 2004; Lee et al., 2002; Zeitlinger et al., 2007), nuclease hypersensitive sites (Crawford et al., 2004; Sabo et al., 2004) and 5C (Dostie and Dekker, 2007; Dostie et al., 2006) allow the identification of functional elements throughout the genome. However, these techniques are limited typically to cell types that can be obtained in sufficient quantities for each protocol. A second limitation is that identification of genomic binding sites by ChIP does not reveal whether those sites are functional; binding sites that are occupied *in vivo* may not contribute to organismal fitness, as long as they do not have negative consequences (Gao et al., 2004; Zeitlinger et al., 2007). CRMs can be computationally identified by searching for overrepresented clusters of binding sites within the genome for groups of TFs that function in a common transcriptional regulatory network (Berman et al., 2002; Lifanov et al., 2003; Markstein et al., 2002; Rajewsky et al., 2002; Sosinsky et al., 2003). The accuracy of these predictions can be improved by incorporating phylogenetic comparisons between species separated by moderate evolutionary distances (Schroeder et al., 2004; Sinha et al., 2004). In combination with ChIP experiments, computational analysis of evolutionary conservation provides an approach to identify functional TF binding sites.

The prediction of CRMs and their cognate factors via binding site cluster analysis has been most thoroughly studied in the context of the regulatory cascade driving anterior-posterior (A-P) pattern formation during embryogenesis in *D. melanogaster*. A hierarchy of genes responsible for the systematic subdivision of the embryo into 14 segments has been defined through exhaustive genetic studies (Jaeger and Reinitz, 2006; Peel et al., 2005; Pick, 1998). These genes are expressed in four sequential cascades - maternal, gap, pair-rule and segment polarity - with genes in each tier of the hierarchy cooperating with the previous group of factors to coordinate expression of the next set. This cascade also activates the homeotic genes in distinct zones that define the initial body plan (Figure 3.1). The majority of genes within this regulatory cascade are TFs that coordinate patterned expression of the next tier of genes by binding to clusters of sites within their

Figure 3.1



maternal genes

bcd, cad, hb

gap genes

head:
btd, D, Gsc, oc, ems, Optix
trunk:
hb, Kr, Ttk, gt, nub, kni, cad
terminal:
tll, Hkb, fkh

pair-rule genes

h, slp1, prd, opa, odd,
Blimp-1, eve, ftz, run*

segment polarity genes

en, inv

homeotic genes

lab, Dfd, Scr, pb, Antp,
Ubx, Abd-A, Abd-B

Figure 3.1. Transcription factors involved in A-P patterning. TFs involved in A-P patterning function in a hierarchical network to subdivide the embryo into 14 segments. The early maternal factors are expressed in broad gradients, with subsequent TF groups expressed in patterns that are increasingly refined. TFs involved in early segmentation or expressed in early patterns that were characterized in this study are grouped according to their initial stage of expression and they are color-coded to indicate the type of DBD (Cys_2His_2 Zinc fingers = Blue, homeodomains = Green, bHLH = Gray, bZip = Red, Winged helix = Pink, Nuclear Hormone Receptor = Orange, POU motif= light Blue, Paired motif = Yellow, and HMG = lavender. Runt is black and was characterized as an alpha fusion (Meng et al., 2005)).

CRMs; the unique expression patterns of the activators and repressors that bind each CRM determine its spatial activity (Arnosti, 2003; Arnosti et al., 1996). However, even within this carefully studied network, the location and number of CRMs that regulate many genes within this segmentation cascade are unknown. One major obstacle is the limited specificity data that is available for even some of the central factors involved in this process (*e.g.* Gt & Kni) despite their identification about 20 years ago (Berman et al., 2004; Schroeder et al., 2004; St Johnston and Nusslein-Volhard, 1992). Position weight matrices (PWMs) for subsets of these TFs have been utilized to identify new CRMs (Berman et al., 2002; Berman et al., 2004; Rajewsky et al., 2002; Schroeder et al., 2004), but a more complete description of specificities would provide a powerful tool to predict CRMs in this transcriptional regulatory network. A comprehensive characterization of *D. melanogaster* TF specificities would provide an invaluable tool for the systematic prediction of CRMs in all regulatory networks.

Here, we describe substantial improvements to a B1H system that increase its sensitivity and dynamic range, and make it amenable for the high-throughput analysis of sequence-specific TFs (*see Chapter 2*). Using this system, we have determined specificities for 108 (14.3%) of the predicted TFs in *D. melanogaster*. These factors represent a broad range of DNA-binding domain families that are commonplace in eukaryotic genomes. Members of one of the groups that we've focused on are those factors that play prominent roles in early A-P patterning. Our dataset dramatically expands the set of defined specificities for these factors and these motifs are good predictors of CRMs throughout the genome. To facilitate utilization of these specificities, we have created a GBrowse-based visualization tool (Stein et al., 2002) that allows an end-user to examine the overrepresentation of binding sites for any number of individual factors as well as combinations of these factors throughout the *D. melanogaster* genome (veda.cs.uiuc.edu/cgi-bin/gbrowse/gbrowse/Dmel4). The genome browser interface is coupled to a genome-wide search tool to identify the most significant peaks of binding site overrepresentation for any combination of factors. The combination of a large database of factor specificities coupled with web-based tools for the rapid analysis of any combination of transcription factors provides the community with a readily accessible tool to analyze and discover CRMs genome-wide. Eventually, the combination of computational predictions and experimental techniques for genome-wide CRM identification (e.g. ChIP-chip) should allow a comprehensive annotation of the CRMs throughout the genome and the TFs that function through these elements.

Results

Large-scale analysis of *D. melanogaster* TFs

To demonstrate that this technology is sufficiently rapid and simple to perform a comprehensive characterization of the TFs we focused on two groups of factors from the *D. melanogaster* genome: all of the factors from a certain DNA-biding domain family (homeodomains, *see Chapter 4*) and all of the factors in a common regulatory network (A-P embryonic patterning). The former set provides a survey of the breadth of specificities that are observed for a particular family, while the later provides a basis set for testing the utility of B1H-generated PWMs for the prediction of CRMs throughout the genome.

.

The early anterior-posterior patterning network in the *D. melanogaster* embryo contains representative members of a wide variety of DNA-binding domain families that are present in higher eukaryotes (Schroeder et al., 2004). Included within this set of factors are members of the five most represented DNA-binding domain families (Tupler et al., 2001): Cys_2His_2 zinc fingers, homeodomains, bHLH, bZIP and winged helix as well as other less represented domains (Figure 3.1). All told, ~80% of the sequence specific TF's in the fly genome utilize one of the DNA-binding domains represented in this group (Adryan and Teichmann, 2006). Some of these TFs, such as the gap gene Kruppel (Kr), have very well defined genetic roles and DNA-binding specificities. Others, such as the gap genes Giant (Gt) and Knirps (Kni) have well defined genetic roles, but their specificities are only roughly described by a handful of binding sites mapped by DNaseI footprinting (Bergman et al., 2005). Therefore, this set of factors provides an opportunity not only to supplement and improve the existing specificity data for this network but also to assess the ability of our technology to characterize a wide variety of DNA-binding domain families.

We characterized the specificity of 35 different factors involved in the A-P pathway, which represent 9 different DNA-binding domain families (Figure 3.2 and *see appendix*

Table A.1). The specificity determined for these factors using the omega-based B1H system is in most cases consistent with previously determined specificity data, where available. For example the existing DNase-based specificities of Bicoid (Bcd), Kr and Tailless (Tll), which represent 3 different families of DNA-binding domains, are quite similar to the specificities obtained from the B1H system (Figure 3.3). Moreover, the stringency of the selection can be varied to recover binding sites with different ranges in affinity as demonstrated by motifs generated for Bcd from sites collected at two different selection stringencies (5 and 10 mM 3-AT, *see Chapter 4*). Both of these motifs display the same core DNA-binding specificity that is consistent with previously published data (Bergman et al., 2005; Wilson et al., 1996), but the higher stringency selection yielded a more constrained motif, due to a greater enrichment of the highest affinity sites. Thus, where good specificity data previously exists, there is excellent concordance between the B1H data and other datasets for these factors.

There are also a number of factors in the dataset with previously poorly defined specificity. Some of these factors, such as Caudal (Cad), Gt and Kni, were originally identified described ~ 20 years ago and play critical early roles in segmentation (St Johnston and Nusslein-Volhard, 1992), yet have poorly defined specificity (Figure 3.3). For example, our data recognition motif for Cad is similar, but much better defined, than the specificity of Cad determined by Dearolf and colleagues (TTTATG) based on sites in the *ftz* zebra stripe element (Dearolf et al., 1989) or than SELEX data available on the Chicken Cad homolog, CdxA (Margalit et al., 1993). The existing DNaseI footprinting

Figure 3.2

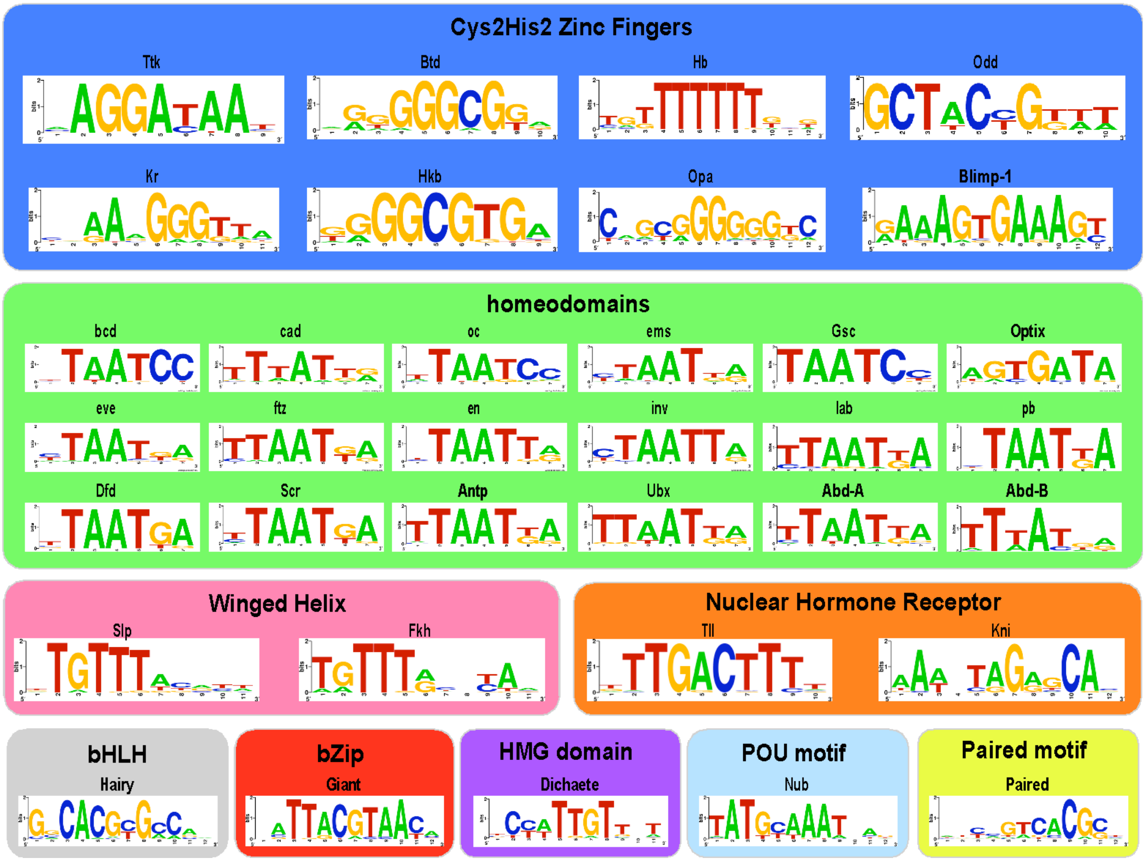

Cys2His2 Zinc Fingers
Ttk
Btd
Hb
Odd
Kr
Hkb
Opa
Blimp-1
homeodomains
bcd
cad
oc
ems
Gsc
Optix
eve
ftz
en
inv
lab
pb
Dfd
Scr
Antp
Ubx
Abd-A
Abd-B
Winged Helix
Slp
Fkh
Nuclear Hormone Receptor
Tll
Kni
bHLH
Hairy
bZip
Giant
HMG domain
Dichaete
POU motif
Nub
Paired motif
Paired

Figure 3.2. Specificities of the 35 TF in the A-P regulatory pathway characterized by the omega-B1H system. These TFs are grouped by DNA-binding domain family and color coded as described in Figure 3.1.

Figure 3.3

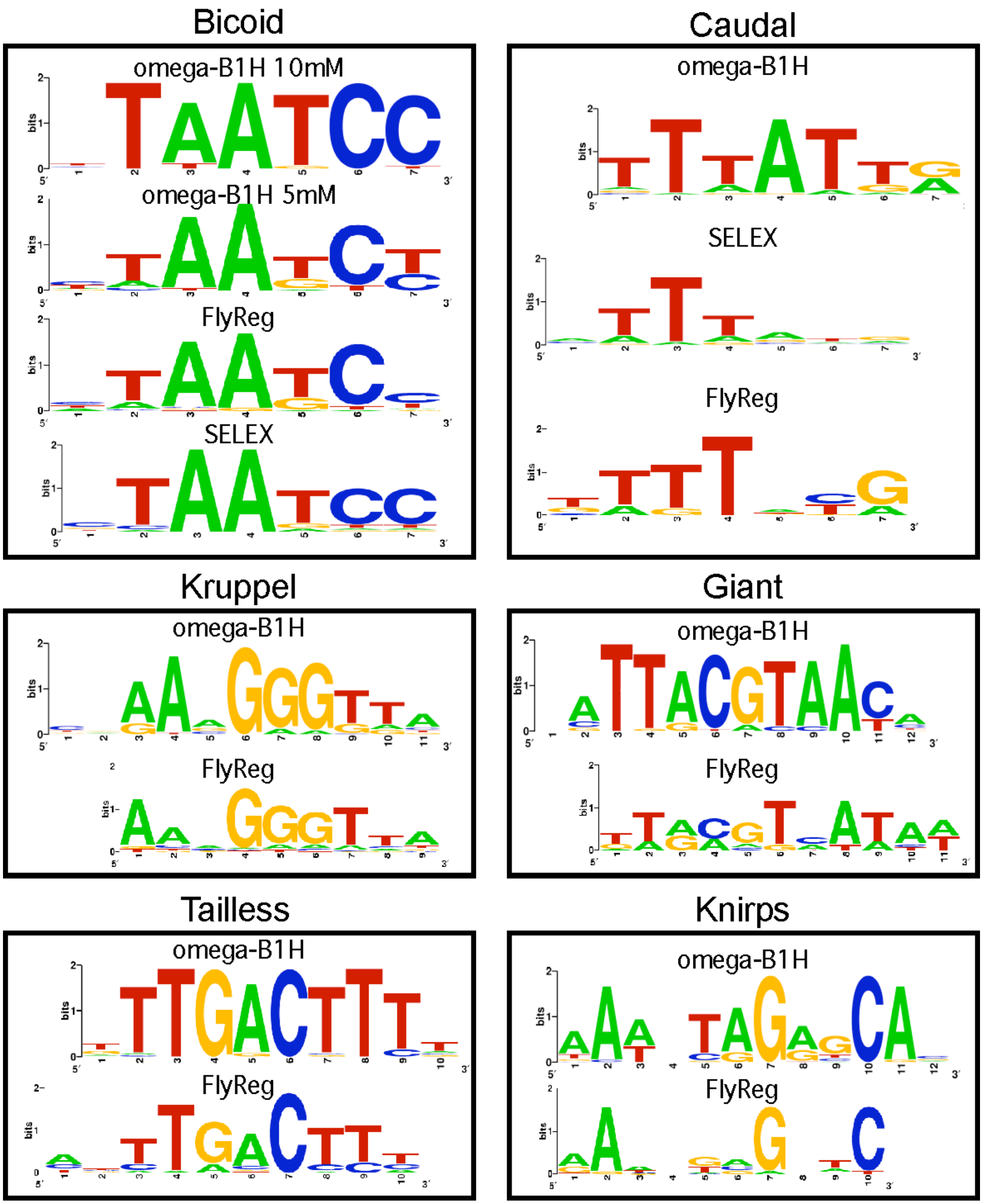

Bicoid
omega-B1H 10mM
omega-B1H 5mM
FlyReg
SELEX
Caudal
omega-B1H
SELEX
FlyReg
Kruppel
omega-B1H
FlyReg
Giant
omega-B1H
FlyReg
Tailless
omega-B1H
FlyReg
Knirps
omega-B1H
FlyReg
bits
5′
3′

Figure 3.3. Comparison of B1H-generated recognition motifs to previously published data. (Left panels) The B1H recognition motifs for Bcd, Kr and Tll are very similar to the motifs generated from DNase footprinting data (FlyREG) (Bergman et al., 2005). In the case of Bcd, the high stringency data (10 mM 3-AT) is most similar to the previously described SELEX data (Wilson et al., 1996), whereas the lower stringency data is more similar to the FlyREG data. (Right panels) The B1H recognition motifs for Cad, Gt and Kni differ significantly from the FlyREG data. For Cad, the B1H-generated data is similar to SELEX data on the chicken homolog (Margalit et al., 1993), but provides better definition of the overall sequence preference.

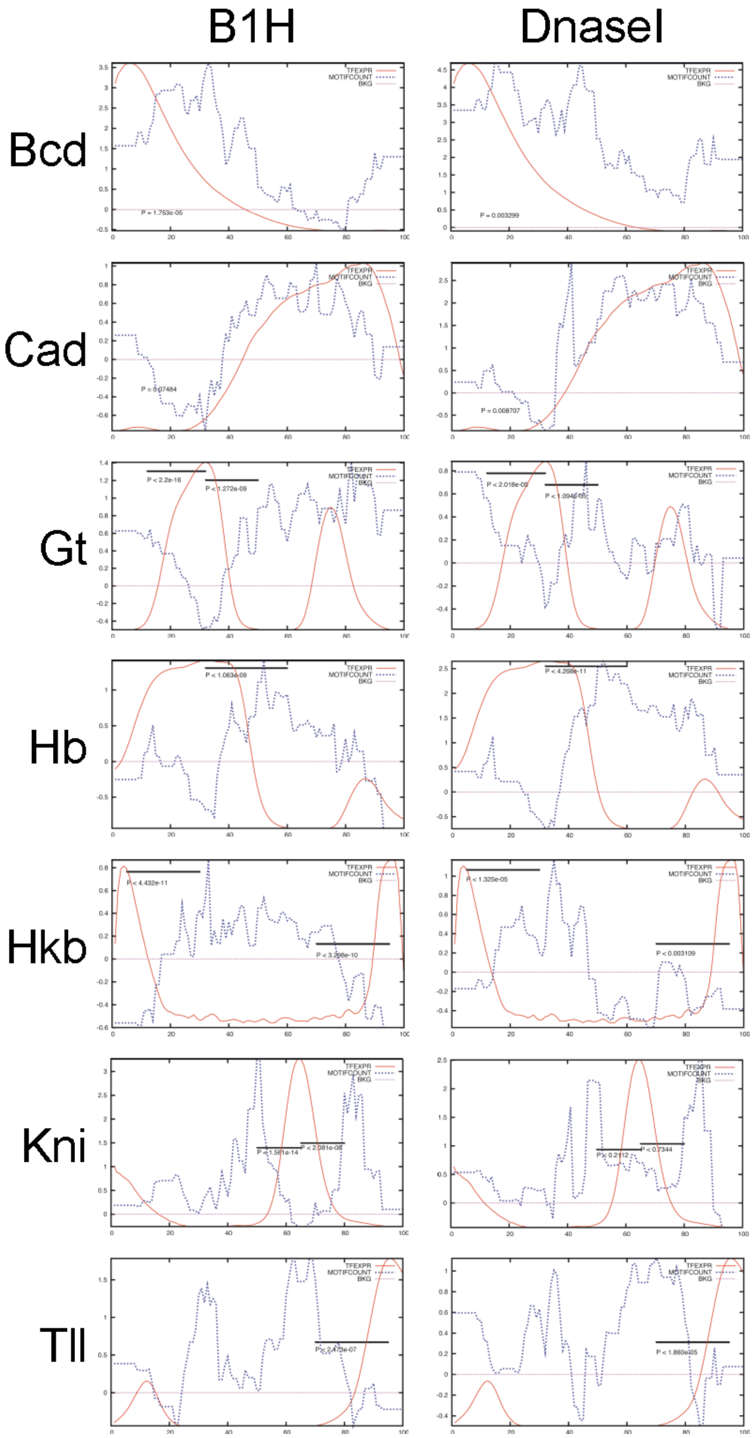
data for Cad appears to misrepresent its specificity in the 3' end of its recognition sequence (Bergman et al., 2005). Two other notable examples are Gt and Kni. The existing DNase motifs for both of these factors contain only limited information about their sequence preferences. By comparison the B1H data for these factors provides a detailed description of their recognition motifs.

In total, we successfully determined the specificity of 101 *D. melanogaster* TFs between this and the homeodomain sets. Data for these factors is available at labs.umassmed.edu/WolfeLab. Only a single factor (Croc) attempted within these sets of factors failed to produce a recognition motif using the omega-B1H system, resulting in a 99% success rate (101/102). This flexibility suggests that the system will be suitable for the high-throughput characterization of the majority of sequence specific transcription factors present in the *D. melanogaster* as well as other eukaryotic genomes.

Assessing the predictive value of the B1H-generated motifs

As an initial assessment of the utility of our binding site motifs for identifying CRMs, we examined the correlation between the expression profile of each transcription factor (TF) and the occurrence of its binding sites in 48 CRMs from *D. melanogaster* that drive patterned gene expression in the early embryo as previously described (Schroeder et al., 2004). When a TF functions as an activator, one would expect an overrepresentation of its binding sites in CRMs that drive gene expression in the same spatial and temporal domains. Conversely, when a TF functions as a repressor that defines a spatial boundary for the expression of a CRM, there should be an anticorrelation between the expression profile of the TF and of CRMs that contain its binding sites. We focused on a set of eight TFs that play prominent roles in early patterning for which we could compare our characterized recognition motifs ("B1H") with existing motifs previously utilized for CRM discovery ("DNaseI") (Schroeder et al., 2004). We used Stubb (Sinha et al., 2003) to calculate, for each CRM, a "Motifcount" score that describes the number of binding sites for each TF and their quality based on its PWM. The average of the Motifcount scores for all of the CRMs contributing to gene expression at each position along the A-P axis was plotted along with the expression profile of each TF (Figure 3.4). Strong correlations or anticorrelations are observed within these plots. For some TFs, such as Bcd and Kr, we find that there are very similar Motifcount profiles for both the DNaseI and B1H PWMs, which is consistent with the similarity between their motifs (Figure 3.3). Bcd displays a strong correlation between its Motifcount and expression profile, as would be anticipated for an activator, whereas Kr displays a strong anticorrelation between its Motifcount and expression profile, as would be anticipated for a repressor. For the majority of these comparisons, the significance of the observed correlation or anticorrelation is greater for the B1H PWMs (indicated by the P-value; Figure 3.4). The most striking difference is observed for Kni, where the P-value improves from 0.2 to e^{-14}. Cad is the one exception; although the B1H motif is more consistent with the existing specificity data, the DNaseI recognition motif displays a somewhat better correlation with the expression data. The improved correlations observed for most of the B1H motifs are

Figure 3.4



	PWM P-values	
	B1H	DnaseI
Bcd	**$1.753e^{-05}$**	0.003299
Cad	0.07484	**0.008707**
Gt	**$2.000e^{-16}$**	$1.094e^{-05}$
Hb	$1.063e^{-09}$	**$4.268e^{-11}$**
Hkb	**$4.432e^{-11}$**	$1.325e^{-05}$
Kni	**$1.561e^{-14}$**	0.2112
Tll	**$2.473e^{-07}$**	$1.860e^{-05}$

Figure 3.4. Motifcount plots over 48 CRMs from *D. melanogaster* that drive early patterned expression. Comparison of the Motifcount plots based on the DNase (Schroeder et al., 2004)and B1H PWMs for Bcd, Cad, Hb, Hkb, Gt, Kni, Kr & Tll. In each plot the red line indicates the TF expression profile over the embryo length (x-axis, 0 = anterior pole; y-axis, arbitrary units). The average number of binding sites for each factor over the CRMs that drive expression in each region (the Motifcount) is indicated by a blue line plotted as a function of the z-score, where 0 is the genome-wide mean (indicated by the magenta line). Solid bars at the top of each graph indicate the window regions over which correlations between the factor expression profile and the Motifcount profile were calculated with the exception of Bcd and Cad, which were calculated over the entire region. P-values for strongest correlation/anticorrelation between the TF expression profile and its Motifcount along the A-P axis are listed to the right of the plots where the bold value indicates the most significant correlation.

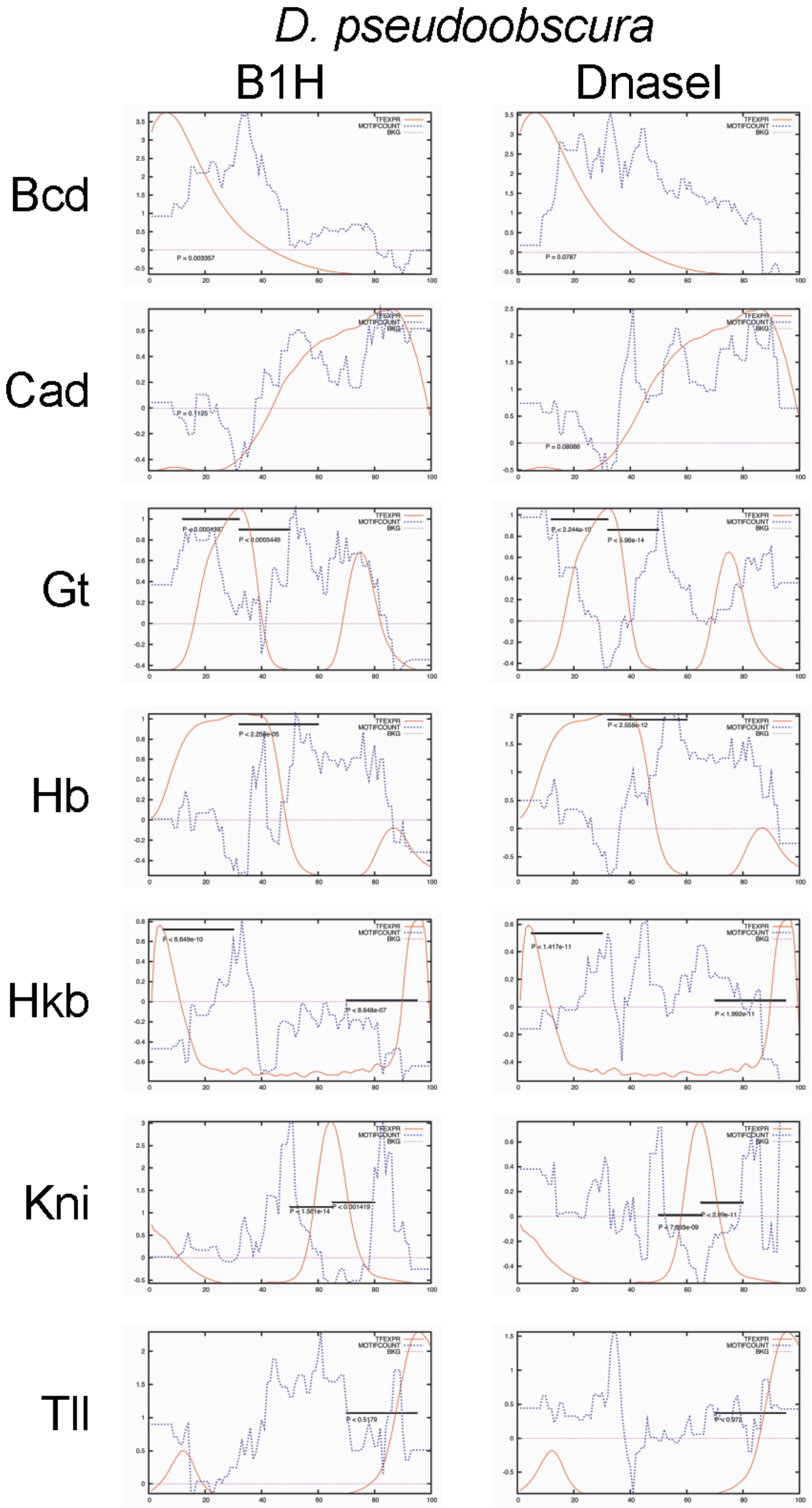
particularly noteworthy given that the majority of the DNaseI data is obtained from bindings sites footprinted within these CRMs.

One other observation from these plots is noteworthy. For some of the repressors, *e.g.* Gt, Hb and Hkb, there is a strong underrepresentation of binding sites in CRMs that have overlapping expression profiles. Selective pressure against the presence of these binding sites may play an important role in shaping the sequence composition of the CRM just as there is selective pressure to maintain binding sites for factors that participate in gene regulation (Ludwig et al., 2000). Overall these results suggest that our B1H-generated PWMs have favorable properties for the prediction of CRMs, and for many factors our motifs appear to be superior to the previously employed PWMs for CRM discovery (Berman et al., 2004; Schroeder et al., 2004). A Motifcount analysis on syntenic regions to these CRMs within the *D. pseudoobscura* and *D. mojanvensis* genomes generates similar plots indicating that our PWMs should have utility for the prediction of CRMs within related species (Figure 3.5).

Genome Surveyor: a new tool for identifying CRMs

We developed a new genome analysis tool, Genome Surveyor, to rapidly search for putative CRMs based on the presence of overrepresented binding sites for a combination of TFs. A simple scoring function was chosen based on its ability to readily identify known CRMs amongst a large population of random intergenic sequences (Table 3.1): Putative CRMs are identified by computing the average of the overrepresentation score

Figure 3.5



D. pseudoobscura
B1H
DnaseI
Bcd
Cad
Gt
Hb
Hkb
Kni
Tll
TFEXPR
MOTIFCOUNT
BKG
P = 0.003357
P = 0.0787
P = 0.1125
P = 0.08086
P < 0.0004397
P < 0.0005449
P < 2.244e-10
P < 5.96e-14
P < 2.258e-05
P < 2.558e-12
P < 6.649e-10
P < 8.648e-07
P < 1.417e-11
P < 1.992e-11
P < 1.561e-14
P < 0.001419
P < 7.033e-09
P < 2.09e-11
P < 0.5179
P < 0.972
0
20
40
60
80
100

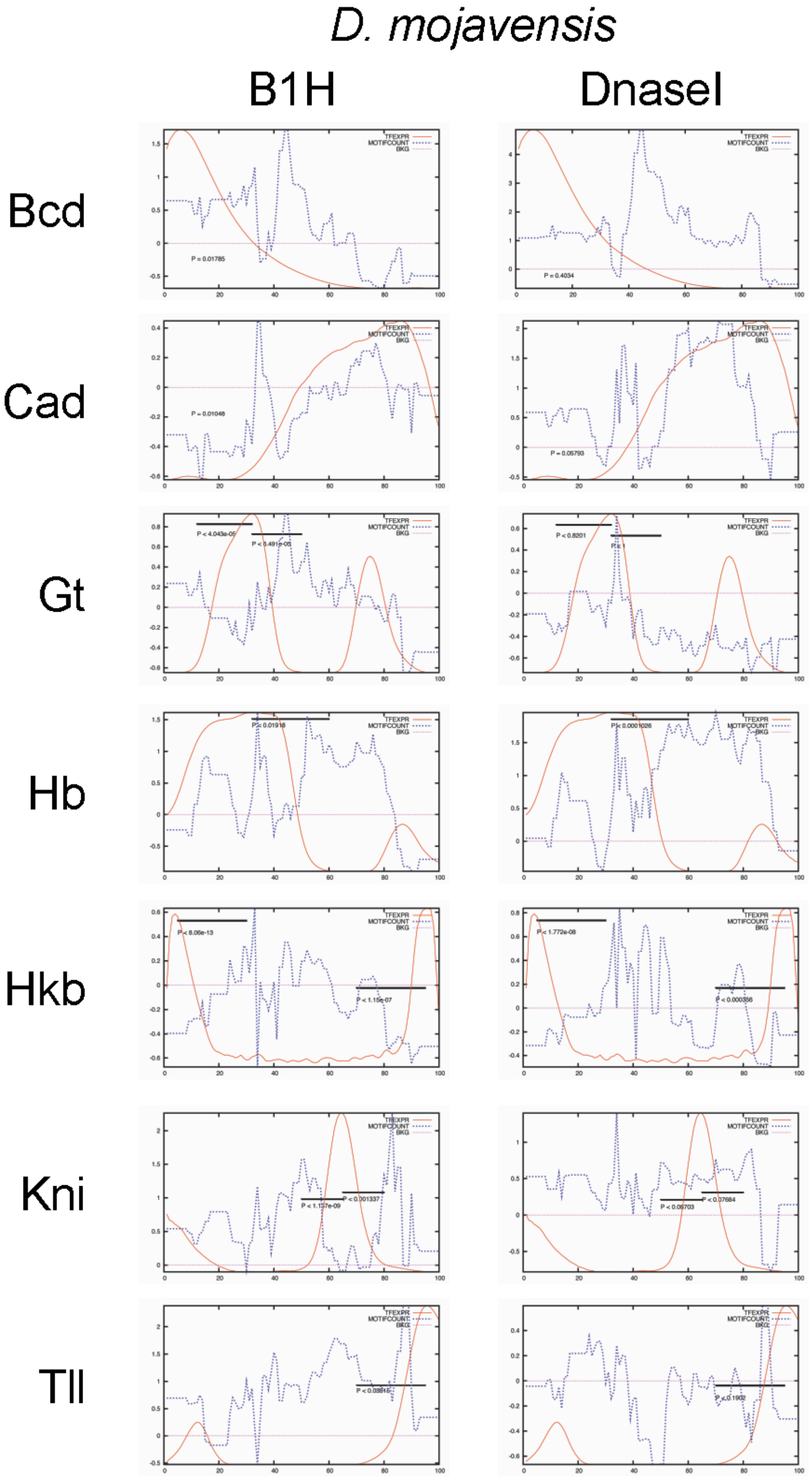

D. mojavensis
B1H
DnaseI
Bcd
Cad
Gt
Hb
Hkb
Kni
Tll
TFEXPR
MOTIFCOUNT
BKG
P = 0.01785
P = 0.4034
P = 0.01048
P = 0.05793
P < 4.043e-05
P < 5.491e-05
P < 0.8201
P < 1
P < 0.01918
P < 0.0001026
P < 8.06e-13
P < 1.15e-07
P < 1.772e-08
P < 0.000356
P < 1.137e-09
P < 0.001337
P < 0.06703
P < 0.07684
P < 0.03815
P < 0.1902

Figure 3.5. Motifcount plots for the DNase and B1H PMWs over syntenic regions for the 48 CRMs within the *D. pseudoobscura* and *D. mojavensis* genomes. Overall these plot look similar to the *D. melanogaster* plots in Figure 5, although some of the feature are more poorly defined. The p-values are significant for a number of factors and in most cases superior for the B1H PWMs. Otherwise, these plots have the same features as described in Figure 5.

(z-score) for a group of TFs over 500 bp windows tiled across the genome. Using our PWMs, this scoring function distinguishes CRMs in our test set with an accuracy that is similar to that of Stubb (Sinha et al., 2003). Importantly, this scoring function provides an enormous advantage in speed over Stubb, as the z-scores for each factor can be calculated once across the genome and then this stored information may then be used in all combination searches that include a particular TF. Our method differs from that of ecis-analyst (Berman et al., 2004) in that we value each site according to its PWM score, which allows both strong and weak sites to contribute to the overall score for each 500 bp window. Then, a z-score is determined for each TF to reflect how the score in that window compares to the overall genomic distribution. By contrast ecis-analyst employs a user-defined threshold (P-value) to determine if a site will be scored as present, and if defined as present, all sites contribute equally to the score.

We developed a flexible user interface that operates through the GBrowse software package (Stein et al., 2002) to allow a user to utilize our scoring function and library of PWMs to search for CRMs in the *D. melanogaster* genome (Figure 3.6). This interface allows gene-specific browsing or genome-wide searching for CRMs. For gene-specific browsing, tracks that indicate the scores for individual factors, along with their significance values, can be displayed across a genomic region of interest (up to 500 kb). Combination tracks can also be generated to identify peaks of binding site overrepresentation for any collection of factors. For example, in the genomic region

Table 3.1 Evaluation of CRM scoring function using 48 defined CRMs and 4800 random regions.

	Number of CRMs of 48 sampled above threshold		
Scoring Function	P≤0.005	P≤0.01	P≤0.05
Stubb, all PWMs together	13	14	24
Sum-of-PWMs, dict	9	15	23
Sum-of-PWMs, dict, z,	13	17	25
Sum-of-PWMs, dict, z + thresholding	**16**	**19**	**29**
Sum-of-PWMs, dict, z + thresholding, Mel-Pse	17	20	33
Sum-of-PWMs, dict, z + thresholding, Mel-Moj (43 CRMs)	16	20	31
Above values extrapolated to 48 CRMs*	18	22	35
Sum-of-PWMs, dict, z + thresholding, Mel-Pse-Moj (43 CRMs)	14	21	28
Above values extrapolated to 48 CRMs*	16	23	31

*Syntenic regions for only 43 CRMs are available in *D. mojavensis*

Figure 3.6

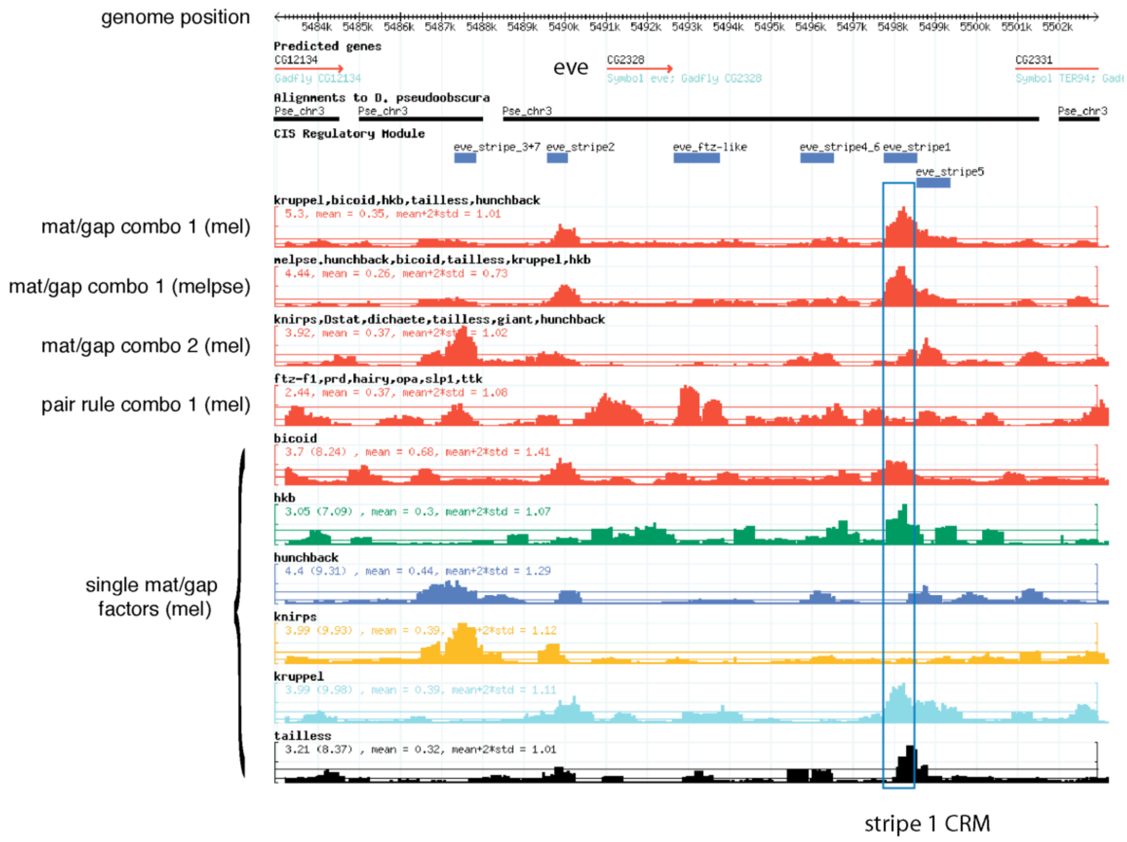

genome position
5484k 5485k 5486k 5487k 5488k 5489k 5490k 5491k 5492k 5493k 5494k 5495k 5496k 5497k 5498k 5499k 5500k 5501k 5502k
Predicted genes
CG12134
Gadfly CG12134
eve
CG2328
Symbol eve; Gadfly CG2328
CG2331
Symbol TER94; Gadf
Alignments to D. pseudoobscura
Pse_chr3
Pse_chr3
Pse_chr3
Pse_chr3
CIS Regulatory Module
eve_stripe_3+7
eve_stripe2
eve_ftz-like
eve_stripe4_6
eve_stripe1
eve_stripe5
mat/gap combo 1 (mel)
kruppel,bicoid,hkb,tailless,hunchback
5.3, mean = 0.35, mean+2*std = 1.01
mat/gap combo 1 (melpse)
melpse,hunchback,bicoid,tailless,kruppel,hkb
4.44, mean = 0.26, mean+2*std = 0.73
mat/gap combo 2 (mel)
knirps,Dstat,dichaete,tailless,giant,hunchback
3.92, mean = 0.37, mean+2*st = 1.02
pair rule combo 1 (mel)
ftz-f1,prd,hairy,opa,slp1,ttk
2.44, mean = 0.37, mean+2*std = 1.08
single mat/gap factors (mel)
bicoid
3.7 (8.24) , mean = 0.68, mean+2*std = 1.41
hkb
3.05 (7.09) , mean = 0.3, mean+2*std = 1.07
hunchback
4.4 (9.31) , mean = 0.44, mean+2*std = 1.29
knirps
3.99 (9.93) , mean = 0.39, +2*std = 1.12
kruppel
3.99 (9.98) , mean = 0.39, mean+2*std = 1.11
tailless
3.21 (8.37) , mean = 0.32, mean+2*std = 1.01
stripe 1 CRM

Figure 3.6. Genome Surveyor display interface. A 20 kb region surrounding the *eve* locus is displayed. Annotations for the *D. melanogaster* genome are shown at the top of the browser window. The predicted transcripts and genes in the *D. melanogaster* genome are indicated within the genomic region. Immediately below is a line indicating the regions where a high confidence alignment with the *D. pseudoobsura* genome has been assembled onto the *melanogaster* scaffold. Annotations for identified CRMs (downloaded from REDfly (Gallo et al., 2006)) can also be displayed within this region. The user-configurable tracks for individual factors or groups of factors are displayed below the annotations. Multiple different factor combination tracks can be displayed simultaneously. These tracks represent the average of the z-scores for each factor plotted over this genomic region for the combination of TFs selected by the user, where the factors included are indicated above each track (*i.e.* Kr, Bcd, Hkb, Tll and Hb, which were the anterior factor search set used to generate the list of hits in Table 1). The numbers in the upper left hand corner indicate the maximum value (z-scores) for each plot, the genome-wide mean and the mean plus two standard deviations, respectively. The positions of the genome-wide mean and the mean plus two standard deviations are also indicated on the plot by horizontal lines of the same color that transect the plot. In this view the two combination tracks (red) for the anterior factor search set are shown across *D. melanogaster* genome (mel) and the average over the *D. melanogaster* and *D. pseudoobsura* genomes (melpse). Both of these factor combinations contain a strong peak within the *eve* stripe 1 CRM. Two other Combination tracks for other groups of factors (a different gap set and a pair-rule set) are also shown. These groups display

Figure 3.6 cont.

significant peaks that overlap with other CRMs. Below the five Combination tracks are a number of tracks for individual factors. These tracks provide a rapid assessment of the individual factors that are potentially contributing to each combination track. For example significant peaks for Bcd, Hkb, Kr and Tll all overlap with the stripe 1 CRM (blue box).

surrounding *eve* the tracks for individual maternal and gap factors (e.g. Bcd, Hkb, Hb, Kr and Tll) display small peaks indicating overrepresentation of sites at various positions, but when certain groups of these factors are combined, strong peaks of binding site overrepresentation are evident that correspond to known *eve* pair-rule stripe CRMs (Figure 3.6). The accuracy of these CRM predictions can be increased by cross-species comparisons to identify peaks that are present in the *D. melanogaster* genome and in a syntenic region of the *D. pseudoobscura* genome (Berman et al., 2004; Sinha et al., 2004). Using our scoring function, the identification of CRMs in a population of intergenic sequences is improved if scores from two genomes are combined (Table 3.1). These comparisons are implemented in Genome Surveyor by calculating z-scores for each TF within the *D. pseudoobscura* genome and mapping the homologous regions onto the *D. melanogaster* genome. The Gbrowse window can be used to display individual and combination tracks for TFs in the *D. pseudoobscura* genome as well as "two-species tracks" that average the z-scores of each factor or group of factors between the two genomes (Figure 3.6). This cross-species analysis over syntenic windows evaluates the total number of sites in each window, not the conservation of individual sites, as individual sites in a CRM may not be conserved but the entire element should be under stabilizing selection (Ludwig et al., 2000). These features allow a user to define significant clusters of binding sites for a group of factors in each genome independently, as well as within both genomes.

We also created a Genome Search Tool within Genome Surveyor that allows a user to perform genome-wide searches for the highest scoring windows using any combination of factors. This page can be accessed via a link in the Gbrowse webpage wherein users can select the combination of factors that they want to employ in their search, the number of top hits that they want returned, and the option to search in the *D. melanogaster* genome alone, or in combination with the *D. pseudoobscura* genome. To avoid recovering peaks that are primarily the result of a strong peak for a single factor, an additional filter can be enabled that requires the combination peak score to be composed of a certain number of factors with individual scores above a desired significance threshold. Each search returns a table of positions within the *D. melanogaster* genome with the highest average z-scores listed in descending order (Table 3.2). The z-scores for each hit are listed in the *D. melanogaster* and *D. pseudoobscura* genomes as well as the combination score across both genomes. The output also includes a list of factors that are contributing significantly to the score within each region, as well as the nearest neighboring genes and their distances from the center of the binding site cluster. The location of each hit is linked back to the Gbrowse tool to enable visualization of the surrounding genomic region for more detailed inspection of the contributing factors.

The effectiveness of these tools and database is evident in the top hits that are returned from a combined *D. melanogaster* and *D. pseudoobscura* genome search using TFs that are involved in anterior patterning (Bcd, Hb, Hkb, Kr & Tll; Table 3.2 and Figure 3.7a-d).

Table 3.2 Top twenty matches in a genome-wide search with Bcd, Hb, Hkb, Kr & Tll.

							known	
Rank	**Location**	**Dmel**	**Dpse**	**DmelDpse**	**Motifs**	**gene**	**CRM**	**expression**
1	2R 5498250	5.3	3.59	**4.44**	bcd, hkb, tll, kr	*eve*	eve_stripe1	stripe1
2	3L 8645450	5.65	1.6	**3.62**	bcd, hkb, tll, kr	*h*	h_stripe1	stripe1
3	2L 3611150	2.51	4.22	**3.31**	bcd, tll kr	*odd*	odd_-5	ant+post
4	3R 2694100	3.47	3.07	**3.24**	bcd, tll, kr	*ftz*	ftz_ftzDE	stripe1+5
5	3L 20630500	3.33	2.93	**3.13**	tll, kr	*kni*	kni_KD	ant+post
6	X 20534450	3.28	2.98	**3.11**	hkb, tll, kr	*run*		stripes
7	X 9535750	3.35	2.73	**3.04**	bcd, tll, kr	*btd*	btd_Ss-Bg	head
8	2L 12682100	3.29	2.66	**2.91**	hkb, kr, hb	*pdm2*		stripes
9	3R 4527100	2.73	2.96	**2.78**	kr, hb	*hb*	hb_HZ526	post
10	3R 675650	1.81	3.47	**2.64**	bcd, tll, hb	*opa*		stripes
11	2L 12689800	2.85	2.43	**2.61**	hkb, kr, hb	*pdm2*		stripes
12	2L 3834050	3.22	2.09	**2.59**	bcd, kr	*slp2*	slp2_-3	ant
13	3R 15955950	2.32	2.99	**2.58**	bcd, tll			
14	X 7500350	3.2	1.89	**2.54**	tll, kr, hb	*cut*		CNS
15	2R 20730400	2.19	3	**2.52**	bcd	*Kr*	Kr_CD1	ant + central
16	X 20462750	2.85	2.38	**2.52**	tll, kr			
17	3L 14138800	2.65	2.33	**2.49**	bcd, hkb, kr	*D*	D_(+5)	central
18	2R 20744500	1.99	2.89	**2.44**	kr, hb	*Kr*		stripes
19	2R 5490050	2.25	2.53	**2.38**	bcd, tll, kr	*eve*	eve_stripe2	stripe 2
20	3L 6090300	2.16	2.76	**2.37**	hkb, tll, kr, hb	*Ets65A*		CNS

Table 3.2 Top twenty matches in a genome-wide search for sequences with overrepresented binding sites for TFs that regulate anterior gene expression during early embryogenesis (Bcd, Hb, Hkb, Kr & Tll). Genomic sites were ranked based on dual genome z-scores (DmelDpse; where genome-wide mean + 2*stdev = 0.74). TF motifs with significant individual scores (>mean + 2*stdev) are shown for each segment. Flanking genes and overlapping CRMs with anterior-posterior specific expression are shown. Because several of the factors are also expressed during CNS development, two flanking genes with CNS specific expression are also indicated.

Figure 3.7a Hairy

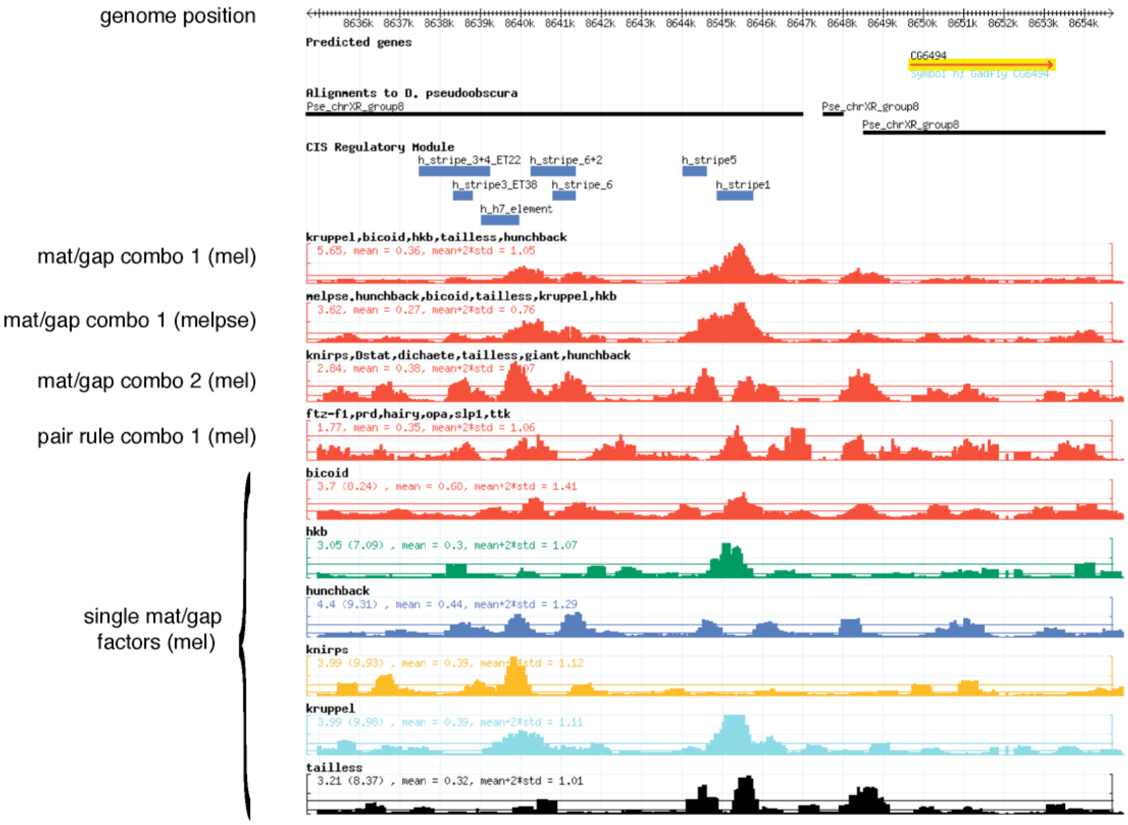


Figure 3.7b Odd

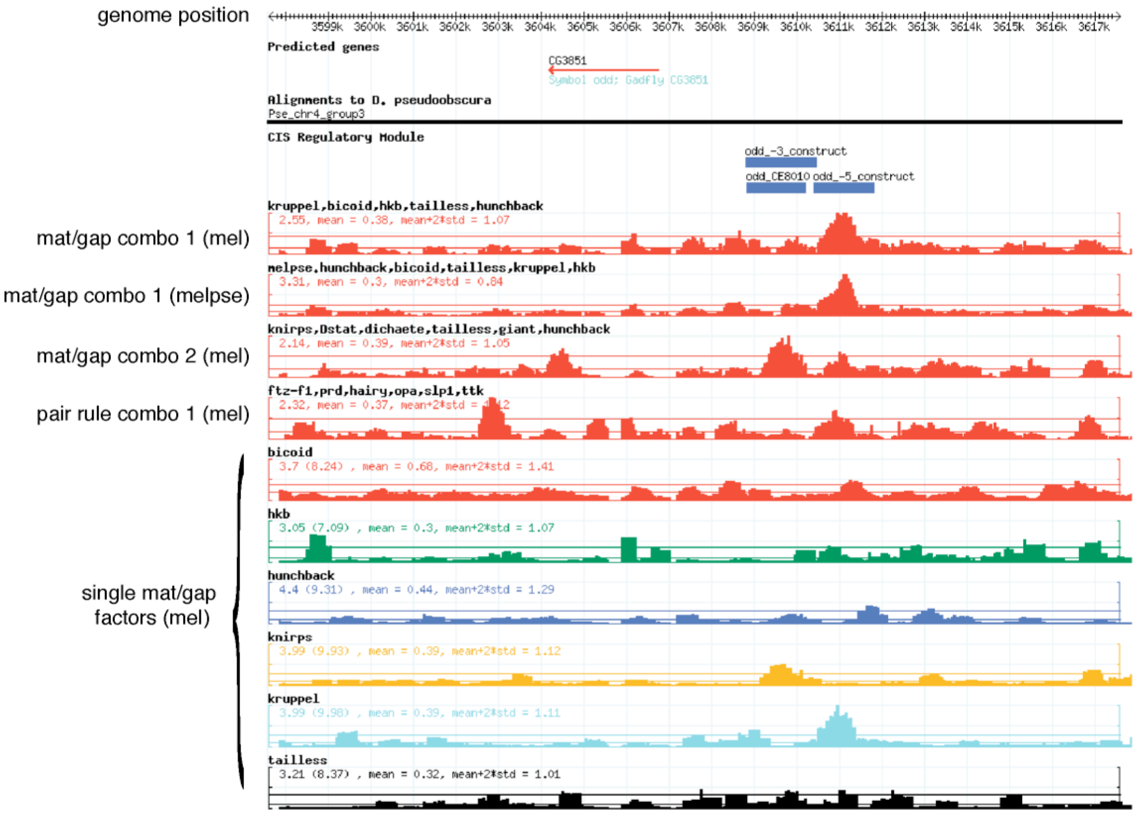


Figure 3.7c Ftz

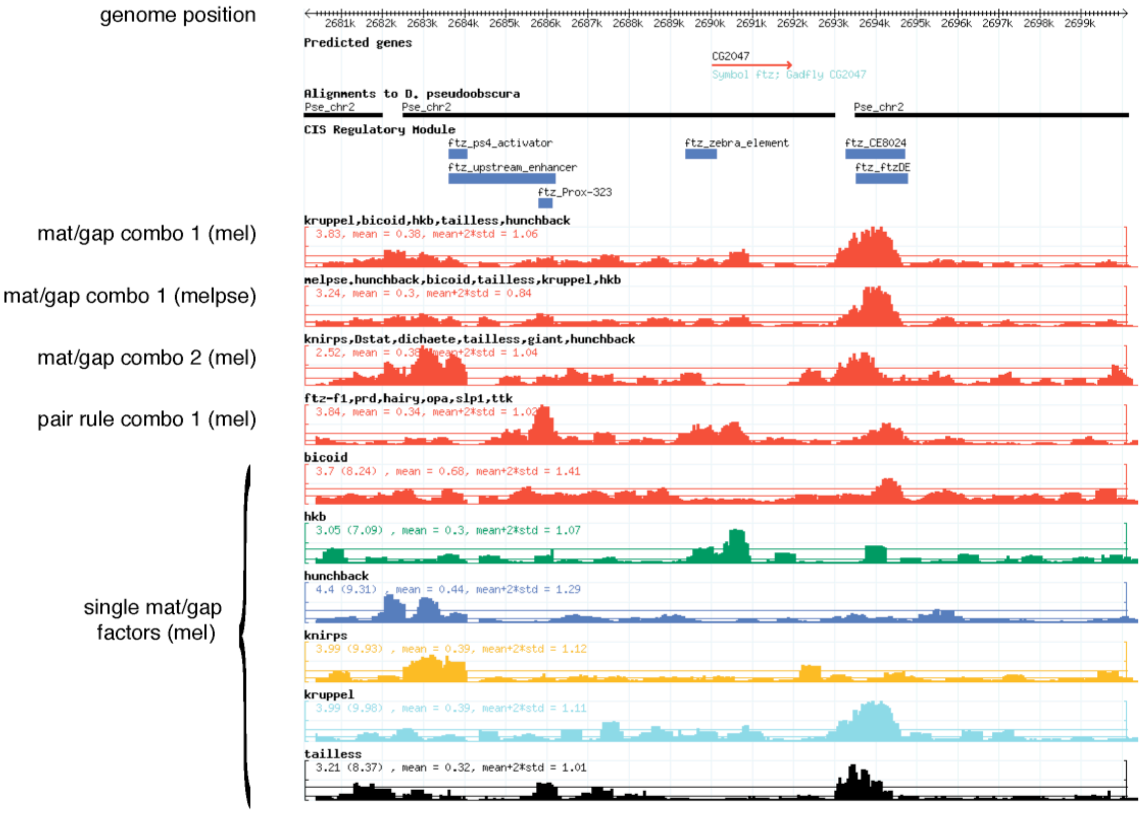


Figure 3.7d knirps

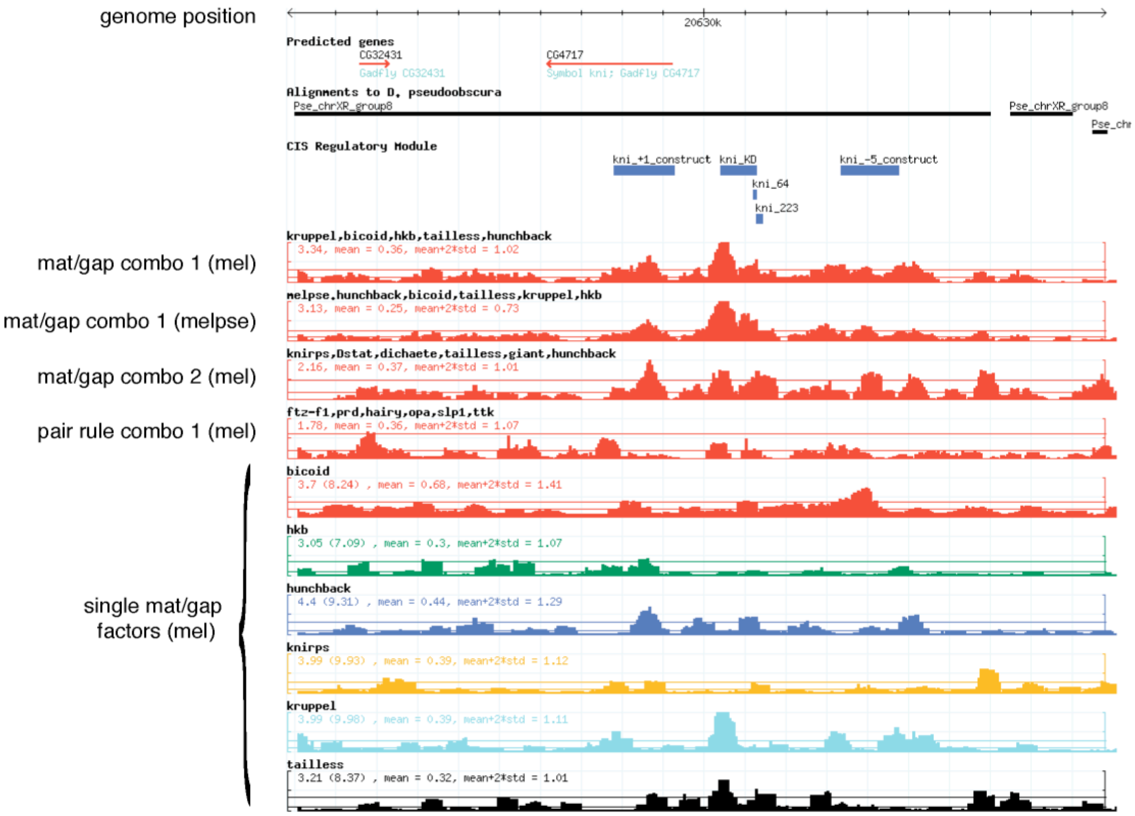


Figure 3.7. Hits two through five from the anterior TF search. Loci surrounding hits neighboring A) *hairy*, B) *odd*, C) *ftz* and D) *knirps*. Strong peaks of binding site overrepresentation are observed in all four of these loci and they overlap with previously identified CRMs that control patterned gene expression in the early embryo. For the three pair-rule genes, these peaks overlap with CRMs that control expression of the most anterior stripe (stripe 1) in their patterned expression. Otherwise the tracks displayed in these images are identical to those in Figure 3.6.

This search produces a remarkable number of strong hits that neighbor genes with early anterior expression patterns: 13 of the 15 top hits are in genes that display early anterior expression and 8 of these 13 are in previously annotated CRMs. The top hit from this search falls within *eve* stripe 1 (Figure 3.6). Bcd, Hkb, Kr, and Tll all contribute robustly to the composite peak at this position as is evident from their individual factor traces, which are all well in excess of 2 standard deviations above the genomic mean. The next four hits within this search neighbor genes with gap or pair-rule patterns of expression (*h*, *odd*, *ftz* & *kni*; Table 3.2). The three hits neighboring the pair-rule genes are all in known CRMs that control expression of "stripe 1", as might be anticipated for the anterior TFs set (Figure 3.7). Performing the search using two genomes significantly increased the number of top hits near genes that are involved in early segmentation. A search with the same set of factors using only the *D. melanogaster* genome yielded a subset of the CRMs that were found in the two-species search (8 of the top 15 hits neighbor genes that display anterior expression, as opposed to 13 of 15 with the "two-species" scores). Thus the dual genome search has enriched the validated positives recovered by the genomic search consistent with previous studies that have utilized of cross-species comparisons in CRM identification (Berman et al., 2004; Sinha et al., 2004).

Discussion

We have developed an omega-based B1H system that allows the high-throughput determination of TF DNA-binding specificity. This system has several advantages over other techniques for characterizing DNA-binding specificity. First, the use of *E. coli* as our platform allows the isolation of complementary TF - binding site combination *in vivo* in a single round of selection using relatively simple techniques. Because *E. coli* demonstrate an extremely high transformation efficiency, randomized binding site libraries with complexity greater than 10^8 members can be utilized. Perhaps the greatest advantage realized by this system is the flexibility provided by utilizing omega-TF hybrids, as the absence of competition from endogenous omega has resulted in an extremely sensitive selection system with a much greater dynamic range than previous systems (Durai et al., 2006; Meng et al., 2005). This sensitivity has allowed us to successfully characterize TFs that failed to generate motifs in the alpha-based B1H system.

Using this system we have determined recognition motifs for ~14% of the predicted *D. melanogaster* TFs. For comparison the FlyREG database contains motifs for 53 TFs constructed from 5 or more identified binding sites (Bergman et al., 2005); thus our database doubles the number of specificities that are available, and in cases where these databases overlap, our data is typically of higher quality. The rate of successful TF characterization within this system (101 of 102) makes it amenable to perform comprehensive surveys of TF specificity in complex organisms: once cloned, the DBDs of ten or more factors can be analyzed in parallel in the B1H system in a manner of days. Our current dataset is focused primarily on monomeric DNA-binding domains, but also includes homodimers and heterodimers. This reductionist approach overlooks the potential for sets of factors to cooperatively recognize motifs that are not a simple composite sites formed from their individual motifs, such as the Exd-Hox combinations that play critical roles in specification during development (Pearson et al., 2005; Ryoo and Mann, 1999; Wilson and Desplan, 1999). These types of combinations can potentially be characterized using the B1H system, as complementary vectors for the characterization of heterodimers have been developed (Meng et al., 2005; Meng and Wolfe, 2006). However, some criteria for choosing sets of factors to be evaluated must be applied because of the combinatorial issues involved with testing all possible pair-wise combinations.

The PWMs generated from our B1H data when used in combination with Genome Surveyor provide a fast, flexible and accessible platform for user-guided prediction of CRMs in the fly genome. This type of PWM-guided CRM discovery has been previously accomplished with a set of maternal and gap TFs by several groups (Berman et al., 2002; Berman et al., 2004; Rajewsky et al., 2002; Schroeder et al., 2004) using different computational approaches. Both demonstrated that known CRMs and novel CRMs could be successfully identified within the genome based on the presence of clusters of binding sites for factors that function in a common regulatory pathway. These studies demonstrated that even relatively crude representations of the DNA-binding specificity of a TF, typically constructed from DNaseI footprinting on a limited number of sites (Adryan and Teichmann, 2006), could help identify CRMs and that these predictions could be improved by using two related fly genomes (Berman et al., 2004; Sinha et al., 2004). These computational approaches, as well as an additional method (Sosinsky et al., 2003), differ in the tactics used for CRM identification, but share a common strategy with Genome Surveyor of identifying clusters of overrepresented binding sites.

The key features of Genome Surveyor are that it evaluates the quality of each binding site as well as over-representation of binding sites relative to the genome average, but is still rapid enough to allow genome-wide searches to be performed on a web-server. Thus, CRM Surveyor, which is integrated within the GBrowse software interface, provides a particularly powerful platform for gene-specific or genome-wide searches for CRMs regulated by a combination of factors. Users can rapidly perform genome-wide searches with any combination of 100 factors over the *D. melanogaster* and *D. pseudoobscura* genomes and then investigate the locations of peaks of interest within the genome using the GBrowse tools. Peaks that overlap with previously identified CRMs can be easily identified by uploading annotations for these elements from the REDfly website (redfly.ccr.buffalo.edu)(Gallo et al., 2006). The number and quality of PWMs available for these searches will increase with the adoption of new, high-depth sequencing and barcoding technologies such 454 (Hoffmann et al., 2007; Margulies et al., 2005) and SOLEXA-based sequencing (Barski et al., 2007; Johnson et al., 2007) for the analysis of the B1H-selected binding sites.

As the number of factors with high quality PWMs increases, it should be feasible to annotate most potential CRMs using combinations of factors that function in common regulatory networks. Cooperating TFs could be identified based on common expression patterns, phenotypes, or physical interactions. Because Genome Surveyor is built into the GBrowse webtool format (Stein et al., 2002), it will also be possible to incorporate other corroborating datasets into these tools, such as genome wide ChIP analysis of TF binding or chromatin structure. The combination of these experimental and computation approaches for the identification of CRMs should provide the most robust method for the functional annotation of these elements throughout eukaryotic genomes.

Experimental Procedures:

Factor information

The amino acid sequence for each factor used and all of the sequences of the binding sites recovered in the individual selections are provided in appendix Table A.1 with the exception of the majority of the homeodomain sequences and selected binding sites, which are available in appendix Table A.2 and described in detail in Chapter 4. Sequence logos (Schneider and Stephens, 1990) for each factor were created by WebLogo (Crooks et al., 2004) using the aligned motifs defined by MEME (Bailey and Elkan, 1994) identified within the B1H-selected sequences.

Motifcount analysis

First, the "expression profile" of a TF is determined from available data on the in situ hybridization of the TF's mRNA (Schroeder et al., 2004; Tomancak et al., 2007), which is a real-valued measurement of the TF's expression level in each of 100 equally-spaced intervals ("bins") along the anterior-posterior axis of the Stage 4-6 (blastoderm) embryo. Then calculate the "discrete expression profile" for a set of 48 CRMs that drive anterior-posterior gene expression in a defined pattern in the blastoderm embryo (Schroeder et al., 2004). For each CRM, determine whether it drives gene expression in each of the 100 bins along the A-P axis by imposing a fixed threshold on the real-valued expression levels. For each CRM, "count" the number of binding sites for the TF, using its PWM and Stubb (Sinha et al., 2003) as described in (Sinha et al., 2006). Then for each of the 100 bins along the A-P axis, collect the set of CRMs that are "expressed" in that bin, and compute the average of the binding site counts for these CRMs (from Step 3). This average is the TF's "MOTIFCOUNT", which is plotted along with the TF's expression profile along the A-P axis.

P-values for this analysis were computed as follows:

1. For a repressor, regions of influence were chosen around the boundaries of its domain of expression, and for each such region of influence the correlation coefficient between the TF expression profile and its MOTIFCOUNT was tested by calculating the Pearson correlation coefficient (rho). We tested the null hypothesis of rho = 0 (with the alternative hypothesis rho < 0, which represents anticorrelation).

2. For an activator, CRMs were classified as either "positive" or "negative" depending on whether the CRM's region of expression overlapped predominantly with the TF's expression domain, or not. A two-sample t-test was performed on the

MOTIFCOUNT in these classes of CRMs to test for a difference of means in these two classes.

Gbrowser web tool

Single motif tracks: For each PWM, scan the genome with a sliding window of 500 bp, in shifted in 50 bp increments, and count the number of occurrences of the PWM in each window, using the Stubb program (Sinha et al., 2003) and the method described in (Sinha et al., 2006) to generate the "DICT" score. The resulting profile of DICT scores is then plotted as a "track" in GBrowse (Stein et al., 2002). These tracks are shown for each PWM in *D. melanogaster* and *D. pseudoobscura* in genomic coordinates of the former. A "two-species" track is also plotted, combining the DICT scores of homologous windows from the two genomes. For this, each species' DICT score is first converted to a "z-score", by subtracting the genome-wide mean and then dividing by the genome-wide standard deviation, and the z-scores of the homologous windows are averaged. For *D.melanogaster* windows in which the syntenic region could not be properly defined using the "liftover" tool (genome.ucsc.edu), the *D. melanogaster* z-score is halved.

Motif combination tracks: Any combination of two or more PWMs can be used to create a "motif combination track" that is dynamically plotted as follows: For each 500 bp window, the z-score of each PWM's DICT score is computed as above, set to zero if it is negative, and an average over the chosen combination of PWMs is regarded as the score of this window. The resulting score profile is plotted as a track. Such tracks may be created for each of the two genomes separately. A “two-species” motif combination track may also be created by averaging the scores from homologous windows. The mean and standard deviation of a combination track is computed from 1 Mbp sequence on either side of the region currently displayed by the browser.

ACKNOWLEDGEMENTS

We would like to thank the Drosophila Genomics Resource Center (DGRC) for providing cDNA collections and Mark Stapleton and Susan Celniker for access to their unpublished transcription factor clone collection and Robin Smith for technical support. S.A.W. M.B.N. and X.M. were supported by NIH grants 1R01GM068110 from NIGMS and 1R21HG003721 from NHGRI, A.W. was supported in part by NIH grant 1R21HG003721 from NHGRI. M.H.B. and A.W. were supported in part by a New Scholar in Aging Award from the Ellison Medical Foundation and American Cancer Society grant RSG-05-026-01-CCG.

CHAPTER IV: A COMPREHENSIVE CATALOG OF HOMEODOMAIN DNA-BINDING SPECIFICITIES FROM *D. MELANOGASTER*

Introduction

In humans, as well as many other metazoans, homeodomains comprise the second largest class of sequence-specific transcription factors (TFs) (Tupler et al., 2001). Homeotic genes were first identified in *D. melanogaster* where the altered activity of a particular gene can result in the morphological transformation of one segment into that of its neighbor, leading to dramatic phenotypes such as the appearance of a second pair of wings (Lewis, 1978). Cloning of these genes led to the landmark observation that they contain a common sequence motif that encodes a DNA-binding domain (Gehring et al., 1994a). Subsequent studies have identified a large number of additional homeodomain proteins in *Drosophila* that regulate diverse processes including appendage formation, organogenesis and cell fate determination. A remarkable number of these genes have mammalian homologs with conserved developmental functions and biochemical properties, contributing to the now widespread appreciation of the conserved molecular mechanisms that regulate signaling and development in metazoans (Banerjee-Basu and Baxevanis, 2001; Mukherjee and Burglin, 2007).

Insights into the mechanisms of sequence-specific DNA binding by homeodomains have been provided by the three-dimensional structures of individual protein-DNA complexes coupled with directed mutagenesis and biochemical analysis (Ades and Sauer, 1995; Gehring et al., 1994b; Wolberger, 1996). The homeodomain consists of approximately 60 amino acids that fold into a stable 3-helix bundle preceded by a flexible

N-terminal arm. Interactions with a 5 to 7 base pair DNA binding site are formed by positioning a single “recognition” helix in the major groove and an N-terminal arm in the minor groove (Figure 4.1, A and B). Despite a common DNA-binding architecture, there is significant variation in the sequence composition within the homeodomain family; for example the two superclasses of homeodomains, denoted as typical and atypical (Banerjee-Basu and Baxevanis, 2001; Gehring et al., 1994a; Mukherjee and Burglin, 2007), share low sequence identity and generally recognize substantially different DNA sequences. Nonetheless, the docking of typical and atypical homeodomains with the DNA is nearly identical, (Kissinger et al., 1990; Wolberger et al., 1991) likely facilitated by common sets of contacts to the phosphodiester backbone. This conserved binding geometry allows differences in amino acids sequence and DNA-binding specificity for various homeodomains to be interpreted within a common structural framework. Residues at positions 2, 3 and 5-8 on the N-terminal arm and at positions 47, 50, 51, 54 and 55 on the recognition helix can all contribute to DNA-binding specificity (Ades and Sauer, 1995; Damante et al., 1996; Ekker et al., 1994; Fraenkel et al., 1998; Hanes and Brent, 1989; Hovde et al., 2001; Passner et al., 1999; Percival-Smith et al., 1990; Piper et al., 1999; Treisman et al., 1989; Wolberger et al., 1991) (Figure 4.1, B and C). The roles of some of these residues in specificity are clear; for example, in the vast majority of homeodomains, Asn is present at position 51 and defines a preference for Ade at binding site position 3 through a bidentate interaction with the N6 and N7 positions of the base. However, at other recognition positions, a precise role in DNA-binding specificity is more difficult to define. For example, Gln is present at position 50 in most homeodomains, but extensive biochemical and structural studies have not defined an obvious role for Gln50 in DNA recognition (Ades and Sauer, 1994; Fraenkel et al., 1998; Grant et al., 2000).

How variations in homeodomain sequence result in differences in binding site specificity has been defined for a number of specific residues. Seminal experiments demonstrated the importance of Lys50 in the recognition of TAATCC by the Bicoid class of homeodomains instead of TAAT(T/G)(A/G) by the Antp and En classes (Hanes and Brent, 1989; Percival-Smith et al., 1990; Treisman et al., 1989). Beachy and colleagues mapped differences in binding site position 2 specificity of the posterior HOX protein Abd-B (TTATGG) and more anterior HOX family members (TAATGG) to amino acids at positions 3, 6 and 7 in the N-terminal arm (Ekker et al., 1994). Interestingly, substitutions at overlapping amino acid positions (6-8) are sufficient to switch the specificity of an NK-2 type homeodomain (CAAGTG) to the specificity of an Antp-type homeodomain (TAAGTG) at the *neighboring* base, binding site position 1 (Damante et al., 1996). This complexity is not limited to the N-terminal arm, as artificial combinations of residues at positions 50 and 54 in the recognition helix fail to interact with anticipated binding sites (Pellizzari et al., 1997). Moreover, residues at different amino acid positions can contact the same base pair. For example, residues at positions 47 and 54 can both interact with complementary bases at binding site position 4, which could create competition between these amino acids to define alternate specificities (Fraenkel et al., 1998; Gruschus et al., 1997; Wolberger et al., 1991). This recognition complexity may

Figure 4.1

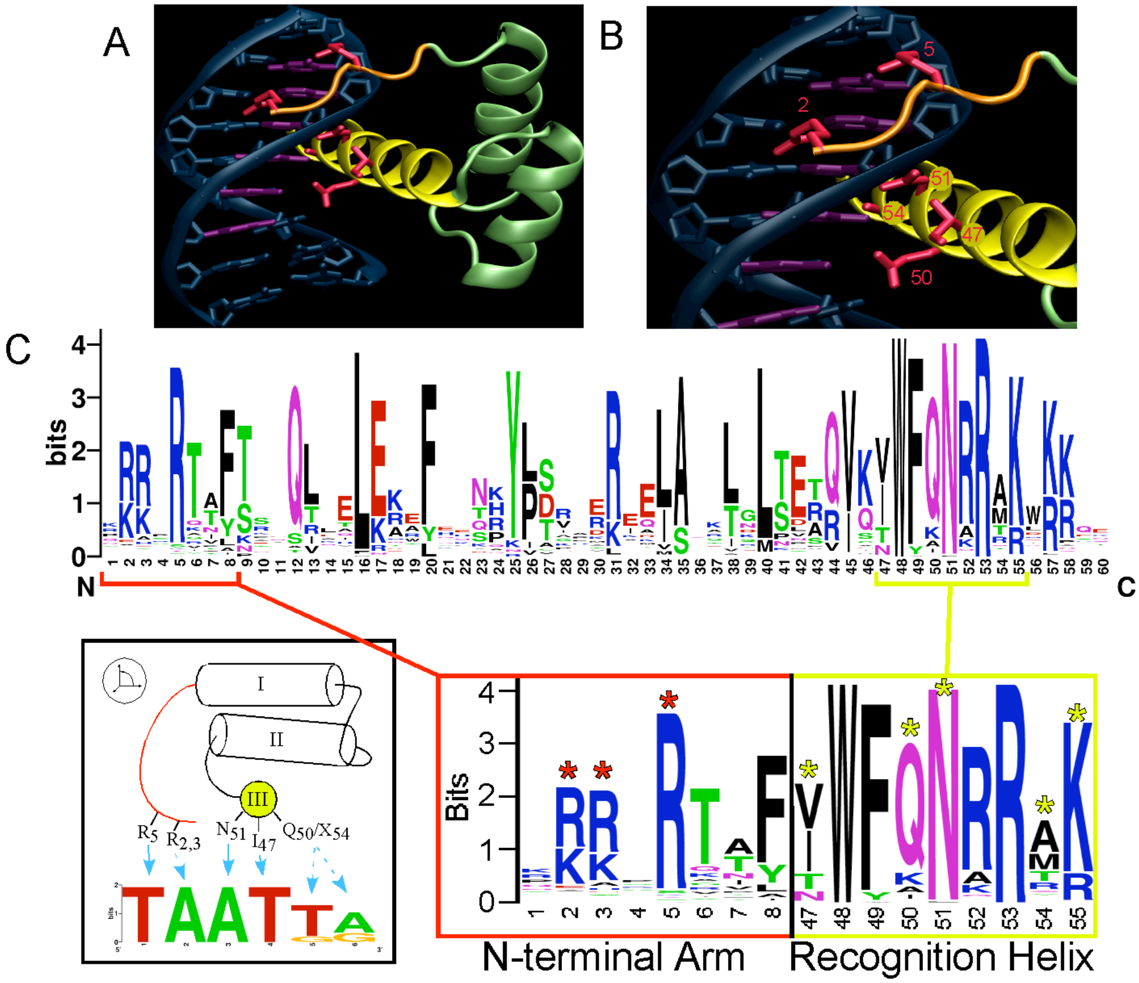

A
B
5
2
51
54
47
50
C
bits
0
1
2
3
4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
N
C
I
II
III
R5
R2,3
N51
I47
Q50/X54
TAATTA
Bits
1 2 3 4 5 6 7 8 47 48 49 50 51 52 53 54 55
N-terminal Arm
Recognition Helix

Figure 4.1. DNA recognition by the homeodomain family. A) The structure of Msx-1 bound to DNA is representative of homeodomain-DNA interactions (Hovde et al., 2001). The N-terminal arm (orange) wraps around the 5' end of the DNA recognition sequence (magenta) interacting through the minor groove while additional contacts to the 3' end of the sequence are achieved by the recognition helix (yellow) through the major groove. Residues involved in base-specific interactions are shown (red). B) Detailed view of the recognition contacts, where residues at positions 2 and 5 interact with bases in the minor groove and residues at positions 47, 50, 51 and 54 are positioned to make contacts in the major groove. C) (Top) Sequence logo representation of the diversity in our set of 84 homeodomains (TALE insertion in atypical homeodomains omitted for clarity). (Bottom) Windows corresponding to the DNA-recognition regions - the N-terminal arm (red) and recognition helix (yellow) – have been expanded to highlight the diversity or conservation at the key DNA-recognition positions (indicated with asterisks). For reference a schematic representation of homeodomain-DNA recognition is shown to the left of the logo.

have hindered some efforts to reengineer homeodomain binding specificity. For example, an attempt to convert the specificity of Matα2 (TTACA) to the specificity of Engrailed (TAATTA) through directed substitutions failed (Mathias et al., 2001).

Likewise, more unbiased approaches to establish alternate specificities, such as selection via phage display, have enjoyed little success (Connolly et al., 1999). Consequently, while roles for specific homeodomain residues in binding site recognition have been defined, a comprehensive description of the determinants of homeodomain DNA-binding specificity remains an important goal.

A comprehensive survey of DNA-binding specificity on a large, diverse family of DNA-binding domains has not been previously attempted. We have recently described a bacterial one-hybrid (B1H) system that allows the specificities of a DNA-binding domain to be rapidly characterized with sufficient ease that multiple factors can be characterized in parallel (Meng et al., 2005; Meng and Wolfe, 2006). We have made further modifications to this selection system that facilitate the characterization of DNA-binding domains, such as homeodomains, that bind short recognition elements. Using this system, we analyze the DNA-binding specificities for all 84 homeodomains in *D. melanogaster* that are not associated with an additional DNA-binding domain. We use the rich experimental history of homeodomain studies in *D. melanogaster* to help interpret this dataset. Our analysis reveals that homeodomain proteins exhibit a diverse array of DNA-binding specificities, with eleven specificity groups in *D. melanogaster* that encompass the majority of these factors. Certain specificities are highly represented, with fifty-one percent of the members falling in the En and Antp groups. The remaining forty-nine percent display subtle to drastic differences in specificity from the common En and Antp groups. Homeodomains within a given specificity group typically share common recognition residues. Combining this data with previous structural and biochemical work on the homeodomain family, we propose and evaluate a detailed set of recognition determinants for homeodomains and use this information to broadly and accurately predict the specificities of homeodomains in the human genome.

Results

Analysis of homeodomains using a modified bacterial one-hybrid (B1H) system

We have made two modifications to our B1H system for characterizing the DNA-binding specificity of a homeodomain (Meng et al., 2005; Meng and Wolfe, 2006): one-hybrid fusions are made to the omega subunit of RNA-polymerase instead of the alpha subunit (Dove and Hochschild, 1998) and homeodomains are expressed as fusions to an accessory DNA-binding domain (Figure4.2A). Omega is not required for viability under laboratory growth conditions (Gentry and Burgess, 1989). Consequently, omega-based one-hybrid selections performed in an omega knockout strain are more sensitive than the corresponding alpha-based system due to the absence of competition from the endogenous subunit (Noyes, Meng, Wakabayashi, Brodsky and Wolfe, *unpublished results*). However, even this more sensitive selection system is unable to determine a recognition motif for several homeodomains. Since homeodomains may recognize a modest 4 to 6 bp element, this failure is likely the result of competitive binding to the thousands of perfect recognition sequences in the *E. coli* genome. Therefore, homeodomains were characterized as fusions to fingers 1 and 2 of the zinc finger protein Zif268 (Zif12; Figure 4.2A). ZFHD1, a fusion between Zif12 and the homeodomain of Oct1, displays superior specificity compared to either of its component parts (Pomerantz et al., 1995). Omega-based binding site selections performed on ZFHD1 yielded two motifs separated by variable spacing consistent with the specificity of both DNA-binding domains (Figure 4.2). A library with10 randomized basepairs adjacent to a Zif12 binding site (ZF10) was created to analyze the DNA-binding specificity of homeodomains fused to Zif12 (Figure 4.2D).

This system was used to determine DNA-binding specificities for all 84 of the homeodomains in the *D. melanogaster* genome that are not associated with an auxiliary DNA-binding domain (Figure 4.3 and appendix Table A.2). The specificity of each homeodomain was obtained in the absence of any other domains present in the endogenous protein. These fly homeodomains cluster into previously described families (Banerjee-Basu and Baxevanis, 2001; Mukherjee and Burglin, 2007) based on their amino acid similarity (Figure 4.4), where homeodomains falling in the typical superclass represent the largest fraction of the population (71 of the 84). A diverse set of amino acids are present at previously defined DNA-recognition positions, implying that a range of DNA-binding specificities may be represented within this population (Figure 4.1C).

Figure 4.2

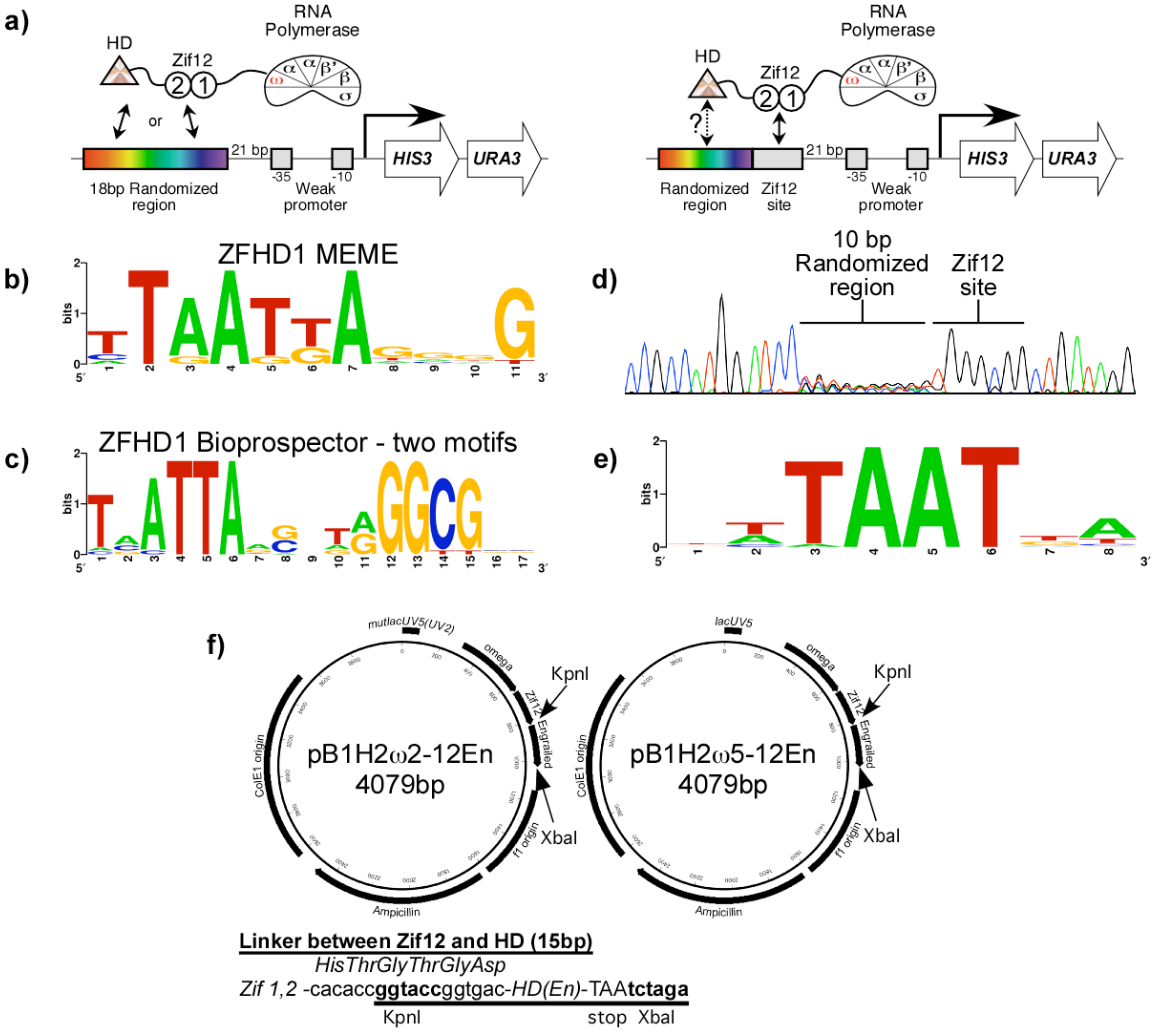
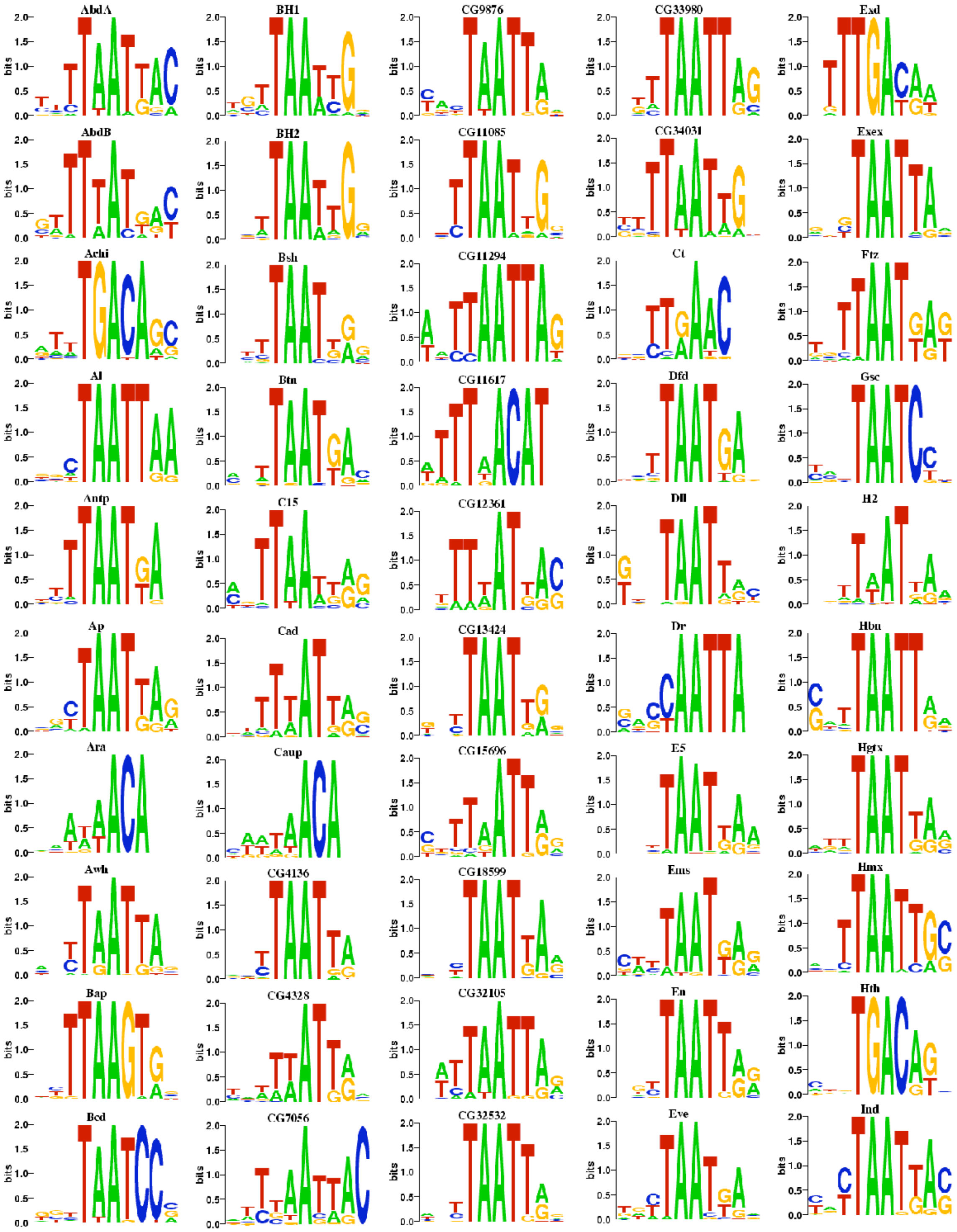

a)
RNA Polymerase
HD
Zif12
2 1
or
21 bp
18bp Randomized region
-35
-10
Weak promoter
HIS3
URA3
Randomized region
Zif12 site
?
b)
ZFHD1 MEME
bits
5′
3′
c)
ZFHD1 Bioprospector - two motifs
d)
10 bp Randomized region
Zif12 site
e)
f)
mutlacUV5(UV2)
omega
KpnI
Zif12
Engrailed
XbaI
f1 origin
Ampicillin
ColE1 origin
pB1H2ω2-12En 4079bp
lacUV5
pB1H2ω5-12En 4079bp
Linker between Zif12 and HD (15bp)
HisThrGlyThrGlyAsp
Zif 1,2 -cacaccggtaccggtgac-HD(En)-TAAtctaga
KpnI
stop XbaI

Figure 4.2. The ZFHD1 framework for determining DNA-binding specificity. A. (left) Cartoon depicting the interaction of the omega-ZFHD1 construct with a reporter with a 18 bp randomized window and (right) one that has a fixed binding site for the Zif12 DNA-binding domain in the context of the ZF10 library. B and C. The logos resulting from the ZFHD1 selection using the 28 bp library as determined by MEME (B. searching for one motif) and by BioProspector (Liu et al., 2001) (C. searching for 2 separate motifs). The discovered motifs in the BioProspector analysis are consistent with the specificity of Zif12 and Oct1 (Pomerantz et al., 1995). D. Sequence chromatogram of the ZF10 library prior to selection with a transcription factor where the 10 bp randomized region and the Zif12 binding sites are indicated. E. The logo resulting from the ZFHD1 selection on the ZF10 library. 5x10^7 bacteria containing the ZF-10 library and the omega-Zif12-Oct1 expression vector were selected on minimal medium lacking histidine and containing 10 mM 3-AT. Approximately 2000 colonies survived the selection, which represented a > 100-fold increase over the number of surviving clones in the omega-Zif12 negative control. MEME analysis of 22 unique sequenced clones recovered a motif consistent with the specificity of Oct1 from 22 of 22 sequences (Pomerantz et al., 1995; Verrijzer et al., 1992). F. The maps of the bait plasmids used to characterize the homeodomains as omega-Zif12 fusions with key features annotated. These plasmids allow a homeodomain to be characterized at two different promoter strengths: either expressed under control of the *lacUV5* or a mutant *lacUV5* (*lacUV5mut*) promoter. Homeodomains were introduced by simply subcloning between the unique KpnI and XbaI sites in each plasmid. The linker between Zif12 and the TF is indicated below.

Figure 4.3


AbdA
BH1
CG9876
CG33980
Exd
AbdB
BH2
CG11085
CG34031
Exex
Achi
Bsh
CG11294
Ct
Ftz
Al
Btn
CG11617
Dfd
Gsc
Antp
C15
CG12361
Dll
H2
Ap
Cad
CG13424
Dr
Hbn
Ara
Caup
CG15696
E5
Hgtx
Awh
CG4136
CG18599
Ems
Hmx
Bap
CG4328
CG32105
En
Hth
Bcd
CG7056
CG32532
Eve
Ind
bits
2.0
1.5
1.0
0.5
0.0

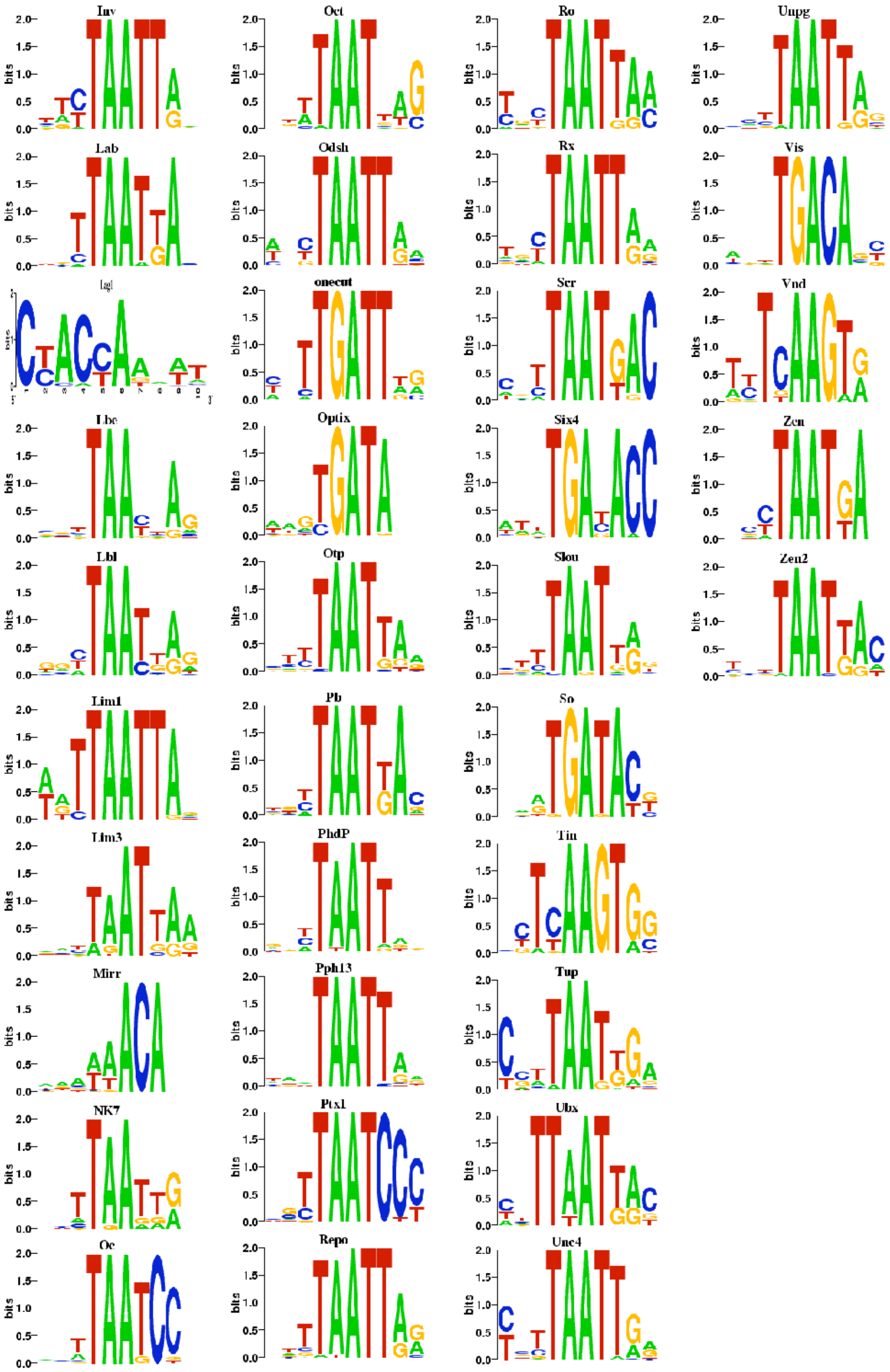

Inv
Lab
Lag1
Lbe
Lbl
Lim1
Lim3
Mirr
NK7
Oc
Oct
Odsh
onecut
Optix
Otp
Pb
PhdP
Pph13
Ptx1
Repo
Ro
Rx
Scr
Six4
Slou
So
Tin
Tup
Ubx
Unc4
Unpg
Vis
Vnd
Zen
Zen2
bits

Figure 4.3. Sequence logos for 84 homeodomains determined in this study. Binding site alignments for each of the 84 homeodomains were extracted from the master alignment of sites. We generated Sequence logos (Schneider and Stephens, 1990) for each of these alignments using WebLogo version 2.8 (Crooks et al., 2004). WebLogo converts alignments to count matrices. Gaps are treated as missing data and ignored, thus not all column count totals are equal. There are no gaps in the core portion of the master alignment, however (columns 4 through 8, see Experimental Procedures).

Figure 4.4

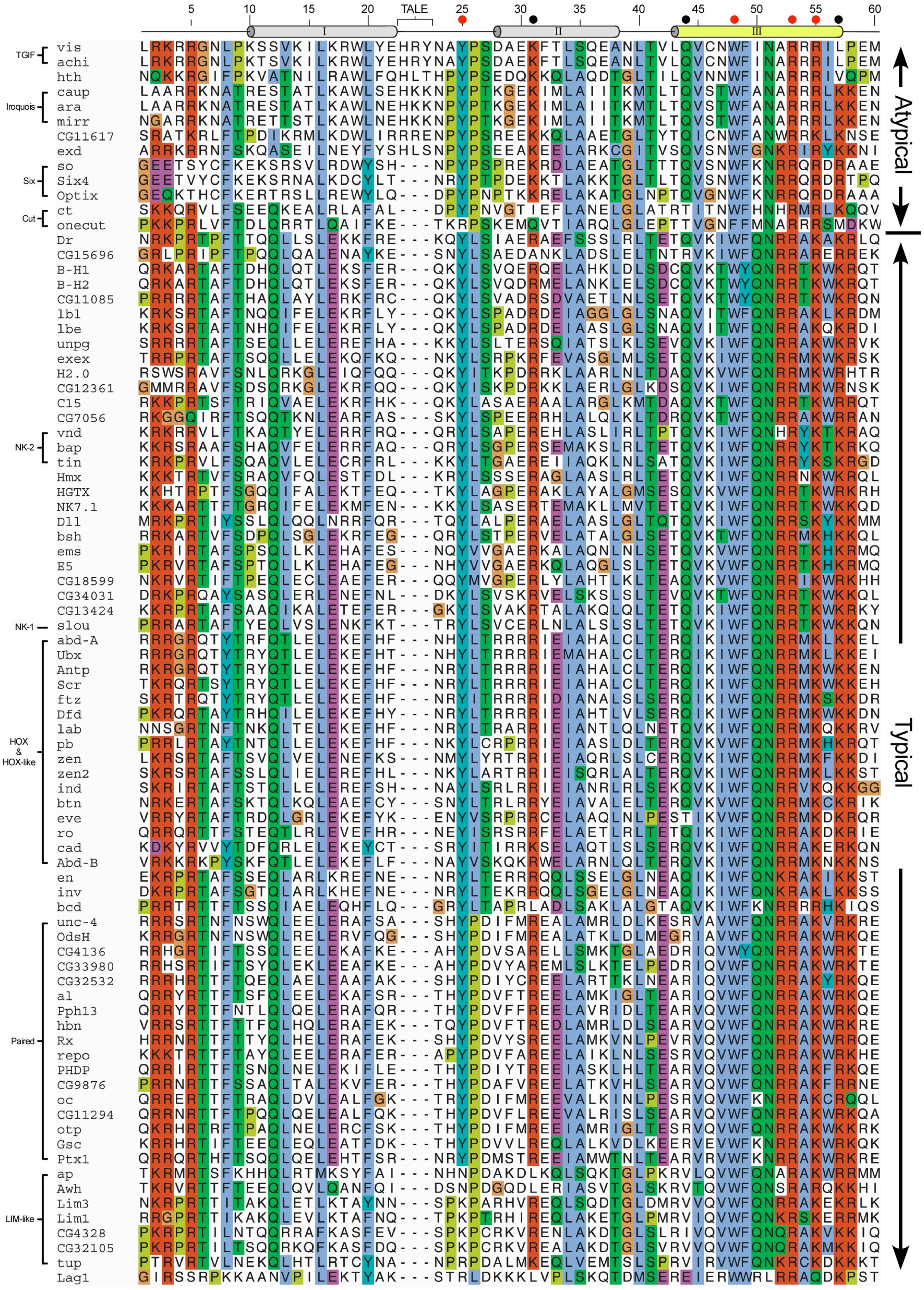

5
10
15
20
TALE
25
30
35
40
45
50
55
60
I
II
III
TGIF
Iroquois
Six
Cut
NK-2
NK-1
HOX & HOX-like
Paired
LIM-like
Atypical
Typical
vis LRKRRGNLPKSSVKILKRWLYEHRYNAYPSDAEKFTLSQEANLTVLQVCNWFINARRRILPEM
achi LRKRRGNLPKTSVKILKRWLYEHRYNAYPSDAEKFTLSQEANLTVLQVCNWFINARRRILPEM
hth NQKKRGIFPKVATNILRAWLFQHLTHPYPSEDQKKQLAQDTGLTILQVNNWFINARRRIVQPM
caup LAARRKNATRESTATLKAWLSEHKKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
ara LAARRKNATRESTATLKAWLNEHKKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
mirr NGARRKNATRETTSTLKAWLNEHKKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
CG11617 SRATKRLFTPDIKRMLKDWLIRRRENPYPSREEKKQLAAETGLTYTQICNWFANWRRKLKNSE
exd ARRKRRNFSKQASEILNEYFYSHLSNPYPSEEAKEELARKCGITVSQVSNWFGNKRIRYKKNI
so GEETSYCFKEKSRSVLRDWYSH---NPYPSPREKRDLAEATGLTTTQVSNWFKNRRQRDRAAE
Six4 GEETVYCFKEKSRNALKDCYLT---NRYPTPDEKKTLAKKTGLTLTQVSNWFKNRRQRDRTPQ
Optix GEQKTHCFKERTRSLLREWYLQ---DPYPNPTKKRELAKATGLNPTQVGNWFKNRRQRDRAAA
ct SKKQRVLFSEEQKEALRLAFAL---DPYPNVGTIEFLANELGLATRTITNWFHNHRMRLKQQV
onecut PKKPRLVFTDLQRRTLQAIFKE---TKRPSKEMQVTIARQLGLEPTTVGNFFMNARRRSMDKW
Dr NRKPRTPFTTQQLLSLEKKFRE---KQYLSIAERAEFSSSLRLTETQVKIWFQNRRAKAKRLQ
CG15696 GRLPRIPFTPQQLQALENAYKE---SNYLSAEDANKLADSLELTNTRVKIWFQNRRARERREK
B-H1 QRKARTAFTDHQLQTLEKSFER---QKYLSVQERQELAHKLDLSDCQVKTWYQNRRTKWKRQT
B-H2 QRKARTAFTDHQLQTLEKSFER---QKYLSVQDRMELANKLELSDCQVKTWYQNRRTKWKRQT
CG11085 PRRRRTAFTHAQLAYLERKFRC---QKYLSVADRSDVAETLNLSETQVKTWYQNRRTKWKRQN
lbl KRKSRTAFTNQQIFELEKRFLY---QKYLSPADRDEIAGGLGLSNAQVITWFQNRRAKLKRDM
lbe KRKSRTAFTNHQIFELEKRFLY---QKYLSPADRDEIAASLGLSNAQVITWFQNRRAKQKRDI
unpg SRRRRTAFTSEQLLELEREFHA---KKYLSLTERSQIATSLKLSEVQVKIWFQNRRAKWKRVK
exex TRRPRTAFTSQQLLELEKQFKQ---NKYLSRPKRFEVASGLMLSETQVKIWFQNRRMKWKRSK
H2.0 RSWSRAVFSNLQRKGLEIQFQQ---QKYITKPDRRKLAARLNLTDAQVKVWFQNRRMKWRHTR
CG12361 GMMRRAVFSDSQRKGLEKRFQQ---QKYISKPDRKKLAERLGLKDSQVKIWFQNRRMKWRNSK
C15 RKKPRTSFTRIQVAELEKRFHK---QKYLASAERAALARGLKMTDAQVKTWFQNRRTKWRRQT
CG7056 RKGGQIRFTSQQTKNLEARFAS---SKYLSPEERRHLALQLKLTDRQVKTWFQNRRAKWRRAN
vnd KRKRRVLFTKAQTYELERRFRQ---QRYLSAPEREHLASLIRLTPTQVKIWFQNHRYKTKRAQ
bap KKRSRAAFSHAQVFELERRFAQ---QRYLSGPERSEMAKSLRLTETQVKIWFQNRRYKTKRKQ
tin KRKPRVLFSQAQVLELECRFRL---KKYLTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
Hmx KKKTRTVFSRAQVFQLESTFDL---KRYLSSSERAGLAASLRLTETQVKIWFQNRRNKWKRQL
HGTX KKHTRPTFSGQQIFALEKTFEQ---TKYLAGPERAKLAYALGMSESQVKVWFQNRRTKWRKRH
NK7.1 KKKARTTFTGRQIFELEKMFEN---KKYLSASERTEMAKLLMVTETQVKIWFQNRRTKWKKQD
Dll MRKPRTIYSSLQLQQLNRRFQR---TQYLALPERAELAASLGLTQTQVKIWFQNRRSKYKKMM
bsh RRKARTVFSDPQLSGLEKRFEG---QRYLSTPERVELATALGLSETQVKTWFQNRRMKHKKQL
ems PKRIRTAFSPSQLLKLEHAFES---NQYVVGAERKALAQNLNLSETQVKVWFQNRRTKHKRMQ
E5 PKRVRTAFSPTQLLKLEHAFEG---NHYVVGAERKQLAQGLSLTETQVKVWFQNRRTKHKRMQ
CG18599 NKRVRTIFTPEQLECLEAEFER---QQYMVGPERLYLAHTLKLTEAQVKVWFQNRRIKWRKHH
CG34031 DRKPRQAYSASQLERLENEFNL---DKYLSVSKRVELSKSLSLTEVQVKTWFQNRRTKWKKQL
CG13424 KKRPRTAFSAAQIKALETEFER---GKYLSVAKRTALAKQLQLTETQIKIWFQNRRTKWKRKY
slou PRRARTAFTYEQLVSLENKFKT---TRYLSVCERLNLALSLSLTETQVKIWFQNRRTKWKKQN
abd-A RRRGRQTYTRFQTLELEKEFHF---NHYLTRRRRIEIAHALCLTERQIKIWFQNRRMKLKKEL
Ubx RRRGRQTYTRYQTLELEKEFHT---NHYLTRRRRIEMAHALCLTERQIKIWFQNRRMKLKKEI
Antp RKRGRQTYTRYQTLELEKEFHF---NRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN
Scr TKRQRTSYTRYQTLELEKEFHF---NRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEH
ftz SKRTRQTYTRYQTLELEKEFHF---NRYITRRRRIDIANALSLSERQIKIWFQNRRMKSKKDR
Dfd PKRQRTAYTRHQILELEKEFHY---NRYLTRRRRIEIAHTLVLSERQIKIWFQNRRMKWKKDN
lab NNSGRTNFTNKQLTELEKEFHF---NRYLTRARRIEIANTLQLNETQVKIWFQNRRMKQKKRV
pb PRRLRTAYTNTQLLELEKEFHF---NKYLCRPRRIEIAASLDLTERQVKVWFQNRRMKHKRQT
zen LKRSRTAFTSVQLVELENEFKS---NMYLYRTRRIEIAQRLSLCERQVKIWFQNRRMKFKKDI
zen2 SKRSRTAFSSLQLIELEREFHL---NKYLARTRRIEISQRLALTERQVKIWFQNRRMKLKKST
ind SKRIRTAFTSTQLLELEREFSH---NAYLSRLRRIEIANRLRLSEKQVKIWFQNRRVKQKKGG
btn NRKERTAFSKTQLKQLEAEFCY---SNYLTRLRRYEIAVALELTERQVKVWFQNRRMKCKRIK
eve VRRYRTAFTRDQLGRLEKEFYK---ENYVSRPRRCELAAQLNLPESTIKVWFQNRRMKDKRQR
ro QRRQRTTFSTEQTLRLEVEFHR---NEYISRSRRFELAETLRLTETQIKIWFQNRRAKDKRIE
cad KDKYRVVYTDFQRLELEKEYCT---SRYITIRRKSELAQTLSLSERQVKIWFQNRRAKERKQN
Abd-B VRKKRKPYSKFQTLELEKEFLF---NAYVSKQKRWELARNLQLTERQVKIWFQNRRMKNKKNS
en EKRPRTAFSSEQLARLKREFNE---NRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKST
inv DKRPRTAFSGTQLARLKHEFNE---NRYLTEKRRQQLSGELGLNEAQIKIWFQNKRAKLKKSS
bcd PRRTRTTFTSSQIAELEQHFLQ---GRYLTAPRLADLSAKLALGTAQVKIWFKNRRRRHKIQS
unc-4 RRRSRTNFNSWQLEELERAFSA---SHYPDIFMREALAMRLDLKESRVAVWFQNRRAKVRKRE
OdsH KRRGRTNFNSWQLRELERVFQG---SHYPDIFMREALATKLDLMEGRIAVWFQNRRAKWRKQE
CG4136 RRHGRTIFTSSQLEELEKAFKE---AHYPDVSARELLSMKTGLAEDRIQVWYQNRRAKWRKTE
CG33980 RRHSRTIFTSYQLEKLEEAFKE---AHYPDVYAREMLSLKTELPEDRIQVWFQNRRAKWRKTE
CG32532 RRRHRTTFTQEQLAELEAAFAK---SHYPDIYCREELARTTKLNEARIQVWFQNRRAKYRKQE
al QRRYRTTFTSFQLEELEKAFSR---THYPDVFTREELAMKIGLTEARIQVWFQNRRAKWRKQE
Pph13 QRRYRTTFNTLQLQELERAFQR---THYPDVFFREELAVRIDLTEARVQVWFQNRRAKWRKQE
hbn VRRSRTTFTTFQLHQLERAFEK---TQYPDVFTREDLAMRLDLSEARVQVWFQNRRAKWRKRE
Rx HRRNRTTFTTYQLHELERAFEK---SHYPDVYSREELAMKVNLPEVRVQVWFQNRRAKWRRQE
repo KKKTRTTFTAYQLEELERAFER---APYPDVFAREELAIKLNLSESRVQVWFQNRRAKWRKHE
PHDP QRRIRTTFTSNQLNELEKIFLE---THYPDIYTREEIASKLHLTEARVQVWFQNRRAKFRKQE
CG9876 PRRNRTTFSSAQLTALEKVFER---THYPDAFVREELATKVHLSEARVQVWFQNRRAKFRRNE
oc QRRERTTFTRAQLDVLEALFGK---TRYPDIFMREEVALKINLPESRVQVWFKNRRAKCRQQL
CG11294 QRRNRTTFTPQQLQELEALFQK---THYPDVFLREEVALRISLSEARVQVWFQNRRAKWRKQA
otp QKRHRTRFTPAQLNELERCFSK---THYPDIFMREEIAMRIGLTESRVQVWFQNRRAKWKKRK
Gsc KRRHRTIFTEEQLEQLEATFDK---THYPDVVLREQLALKVDLKEERVEVWFKNRRAKWRKQK
Ptx1 QRRQRTHFTSQQLQELEHTFSR---NRYPDMSTREEIAMWTNLTEARVRVWFKNRRAKWRKRE
ap TKRMRTSFKHHQLRTMKSYFAI---NHNPDAKDLKQLSQKTGLPKRVLQVWFQNARAKWRRMM
Awh TKRVRTTFTEEQLQVLQANFQI---DSNPDGQDLERIASVTGLSKRVTQVWFQNSRARQKKHI
Lim3 NKRPRTTITAKQLETLKTAYNN---SPKPARHVREQLSQDTGLDMRVVQVWFQNRRAKEKRLK
Lim1 RRGPRTTIKAKQLEVLKTAFNQ---TPKPTRHIREQLAKETGLPMRVIQVWFQNKRSKERRMK
CG4328 PKRPRTILNTQQRRAFKASFEV---SPKPCRKVRENLAKDTGLSLRIVQVWFQNQRAKVKKIQ
CG32105 PKRPRTILTSQQRKQFKASFDQ---SPKPCRKVREALAKDTGLSVRVVQVWFQNQRAKMKKIQ
tup PTRVRTVLNEKQLHTLRTCYNA---NPRPDALMKEQLVEMTSLSPRVIRVWFQNKRCKDKKKT
Lag1 GIRSSRPKKAANVPILEKTYAK---STRLDKKKLVPLSKQTDMSEREIERWWRLRRAQDKPST

Figure 4.4. ClustalW alignment of the homeodomain motifs from the 84 homeodomains examined in this study. The position of each amino acid in the canonical homeodomain numbering scheme is indicated at the top of the alignment, along with cylinders that indicate the position of the three alpha helices in the structure. The homeodomains can be divided into two broad superclasses: the Atypical (top) and Typical (bottom), which have distinguishing sequence features and recognition motifs. One of the most prominent sequence differences is the presence of a three amino acid loop extension (TALE) between helix 1 and 2 within the majority of the atypical homeodomains (Banerjee-Basu, 2001; Mukherjee, 2007). Where clearly defined, the classes that the homeodomains fall within are indicated to the left of their names (Banerjee-Basu, 2001; Harvey, 1996). Colored residues at each position within the alignment indicate a high frequency of occurrence of residues with similar properties, with the exception of Gly and Pro, which are highlighted throughout these sequences. Filled circles above the sequences indicate positions that make phosphate contacts in the structures of both typical and atypical homeodomain-DNA complexes (Wolberger, 1991; Fraenkel, 1998; Joshi, 2007). Red circles indicate positions where an amino acid type capable of making a phosphate contact is conserved in all Asn51 containing homeodomains. At these seven common phosphate-contacting positions 95% (79 of 83) have an appropriate residue for this interaction at 6 of the 7 positions. In the case of position 31, the absence of a Lys or Arg can be compensated by one at position 46, as either position can contact a common phosphate (Grant, 2000). Not surprisingly, two of the outliers with regards to conserved phosphate contacts are Cut and Onecut, which display unique specificities.

One notable exception to this diversity is Asn at position 51 of the recognition helix, which is present in all but one of these homeodomains.

Several observations indicate that the motifs obtained by the B1H method accurately reflect the DNA-binding specificities of homeodomains. Some of the most thoroughly characterized homeodomains within *D. melanogaster* are members of the homeotic (HOX) gene family (Gehring et al., 1994a; Pearson et al., 2005). These factors provide a benchmark for assessing the quality of the B1H-generated specificity data. All of our determined specificities for the HOX factors share a common consensus – T(A/T)AT(T/G)(A/G), which is consistent with the previously described binding site selection data on these factors (Figure 4.5). Previous *in vitro* binding site selections and gel shift assays revealed subtle differences in the specificity of Ubx, Dfd and Abd-B (Ekker et al., 1994; Ekker et al., 1992); these factors exhibit similar preferences in our data, such as the preference of Abd-B for Thy over Ade at binding site position 2 (Ekker et al., 1994). Thus, even subtle differences in homeodomain specificity can be captured by the B1H analysis. Other examples where the B1H dataset is supported by previously identified binding sites in *D. melanogaster* or mammals are described for individual specificity groups below. To further assess the accuracy of our B1H-generated data, competition gel mobility shift assays were performed for 9 factors that display different specificities using binding sites representing 8 different specificity groups; these results are consistent with the B1H specificities (Figure 4.6).

Figure 4.5

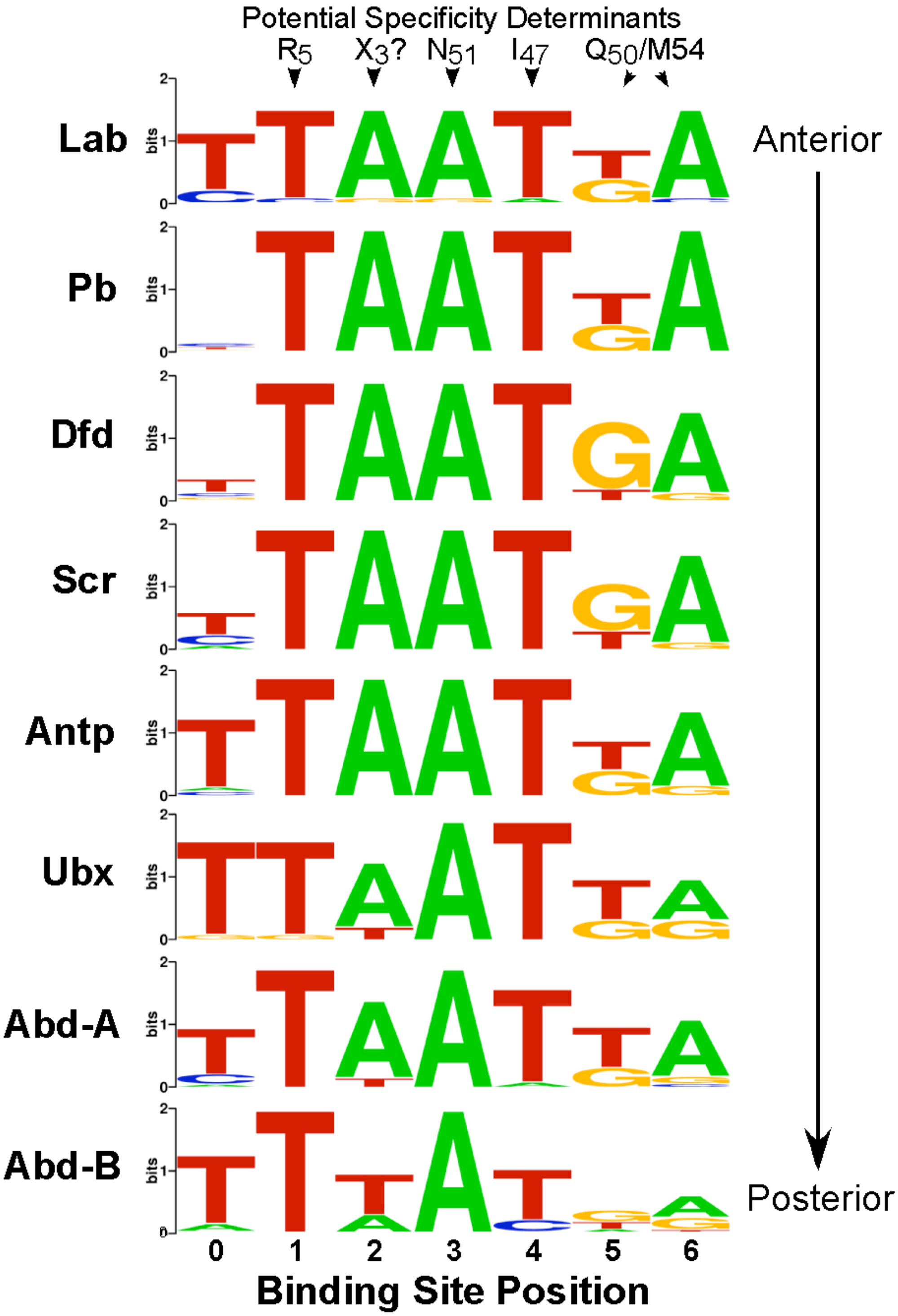
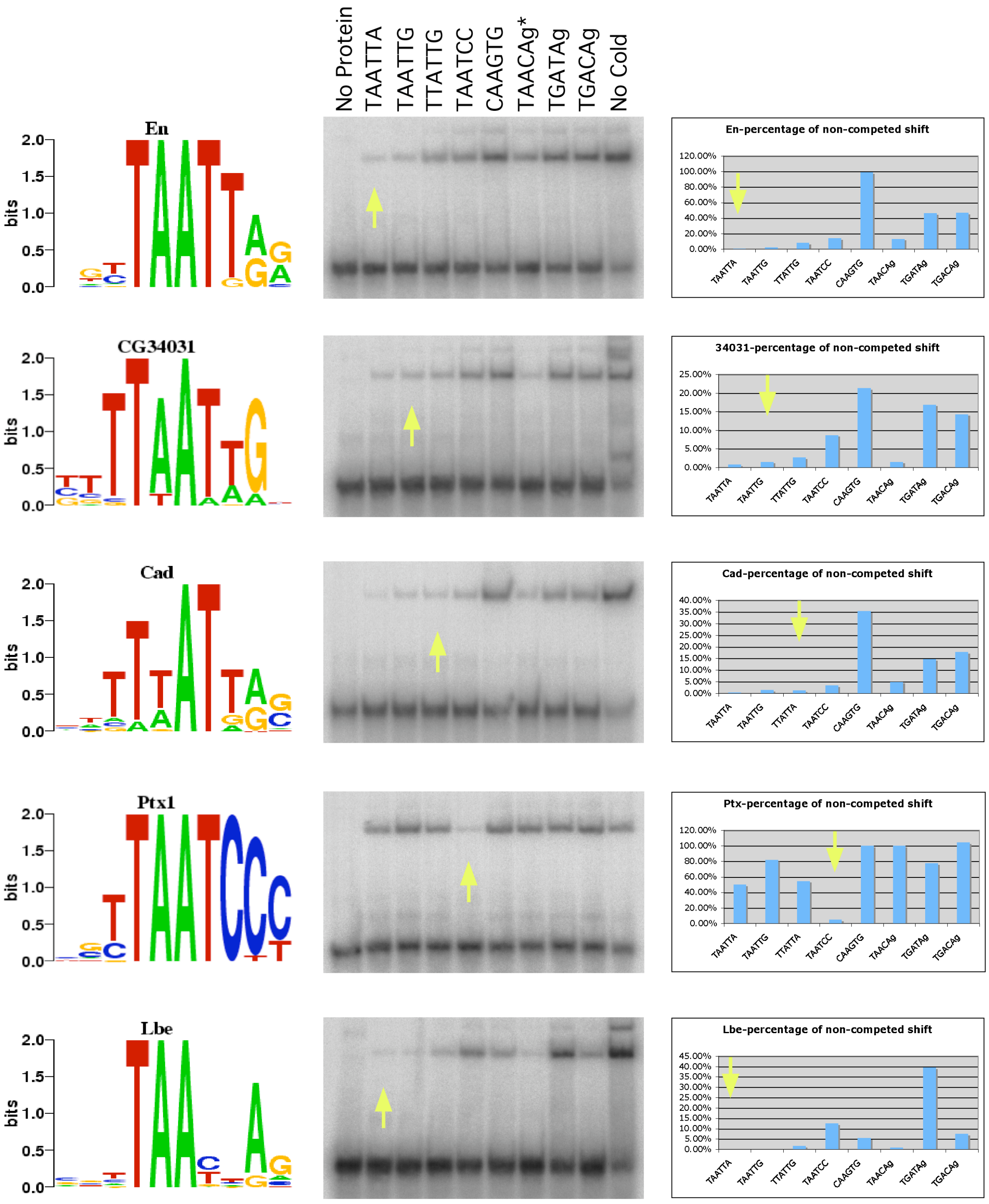

Potential Specificity Determinants
R5
X3?
N51
I47
Q50/M54
Lab
Pb
Dfd
Scr
Antp
Ubx
Abd-A
Abd-B
bits
0
1
2
Anterior
Posterior
0
1
2
3
4
5
6
Binding Site Position

Figure 4.5. Binding site logos for the eight *D. melanogaster* HOX factors displayed in order of anterior to posterior expression (top to bottom). Potential specificity determinants for each position of the binding site are indicated above the logos. There are subtle differences between the specificity of each HOX factor that are highlighted when they are arranged according to their regional expression pattern within the embryo. For example, at binding site position 2, the specificity shifts from a strong preference for Ade in more anterior factors toward an increasing tolerance for Thy in the most posterior factors, with Abd-B displaying a significant preference for Thy over Ade as was previously noted in the study by Beachy and colleagues(Ekker et al., 1994). This change in base preference was ascribed to residues at position 3, 6 and 7 in the N-terminal arm of Abd-B(Ekker et al., 1994). A second trend is the increasing preference for Thy at position 0 in the more posterior factors. This observation is consistent with a greater preference for Thy in Ubx and Abd-B compared to Dfd in *in vitro* binding site selections(Ekker et al., 1994; Ekker et al., 1992). Another notable difference previously observed for Dfd and Ubx was a stronger preference of Dfd for Gua over Thy at binding site position 5, whereas Ubx did not display an obvious preference for Gua over Thy at this position(Ekker et al., 1992).

Figure 4.6


No Protein
TAATTA
TAATTG
TTATTG
TAATCC
CAAGTG
TAACAg*
TGATAg
TGACAg
No Cold
En
CG34031
Cad
Ptx1
Lbe
bits
2.0
1.5
1.0
0.5
0.0
En-percentage of non-competed shift
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
TAATTA
TAATTG
TTATTG
TAATCC
CAAGTG
TAACAg
TGATAg
TGACAg
34031-percentage of non-competed shift
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
Cad-percentage of non-competed shift
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
TTATTA
Ptx-percentage of non-competed shift
Lbe-percentage of non-competed shift
45.00%

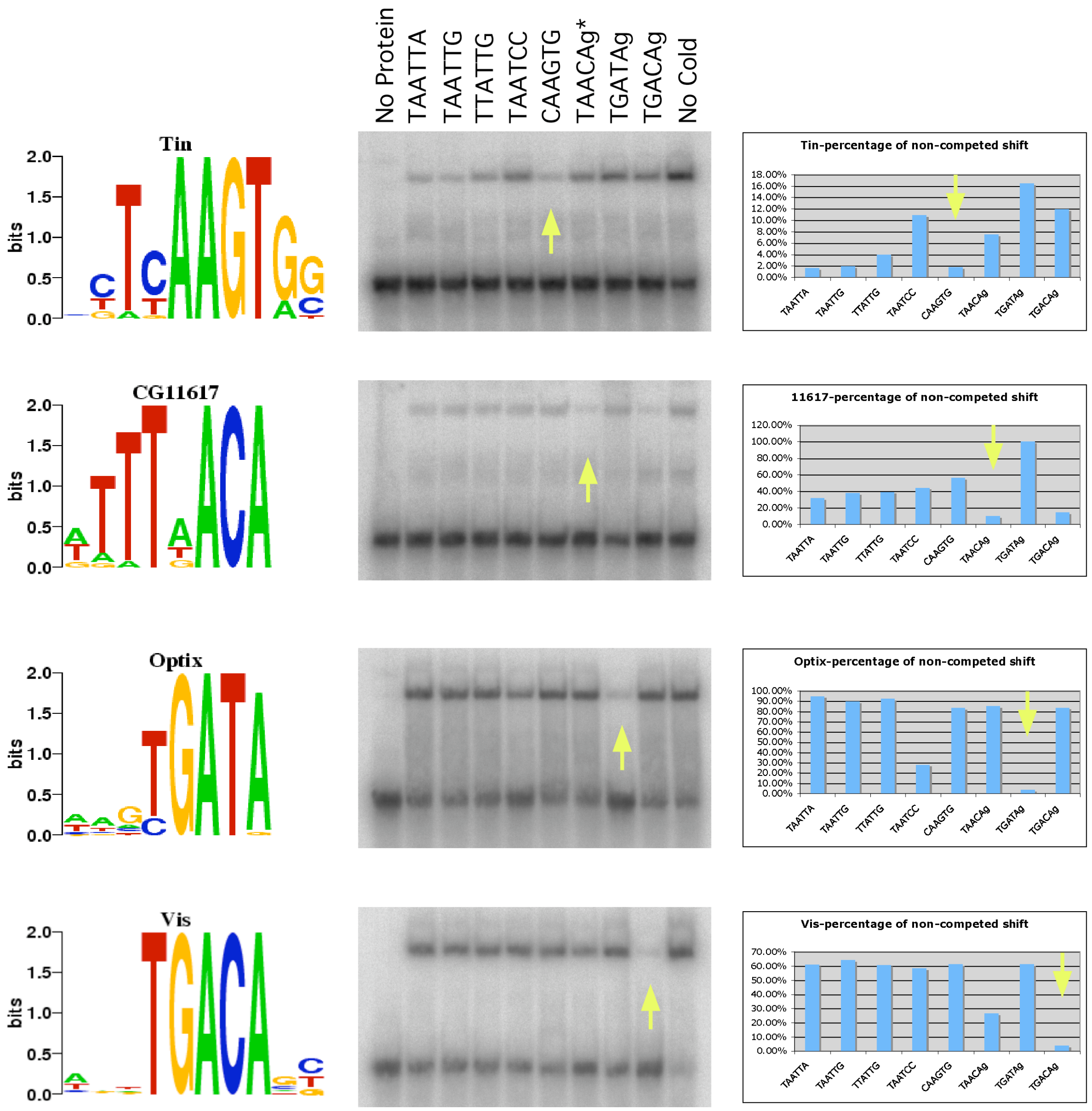

No Protein
TAATTA
TAATTG
TTATTG
TAATCC
CAAGTG
TAACAg*
TGATAg
TGACAg
No Cold
Tin
bits
2.0
1.5
1.0
0.5
0.0
Tin-percentage of non-competed shift
18.00%
16.00%
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
TAATTA
TAATTG
TTATTG
TAATCC
CAAGTG
TAACAg
TGATAg
TGACAg
CG11617
11617-percentage of non-competed shift
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Optix
Optix-percentage of non-competed shift
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Vis
Vis-percentage of non-competed shift
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%

Figure 4.6. Competitive gel shift assays for nine different fly homeodomains representing nine different specificity groups. The sequence logo for the factor used in each assay is shown on the left of each row. Radiolabeled DNA containing the consensus binding site for each factor was challenged with excess cold competitor DNA containing sequences from the various specificity groups. Each lane represents a different competition, where the cold competitor and protein are at constant concentrations, with the exceptions of the far left (no protein) and far right (no competitor) lanes. The sequence of the homeodomain binding site used in each competition is indicated at the top of the page, where they are at identical positions in each assay. A yellow arrow indicates the competitor that is identical in sequence to the radiolabeled probe. The quantification of the change in percent shift as a function of the concentration of cold competitor is shown on the right where these values are normalized to the percent shift in the no competitor lane. A yellow arrow indicates the competitor that is identical in sequence to the radiolabeled probe. The neighboring 5' constant region of each binding site is CAG with the exception of the indicated site (*), which contains an additional T (CAG**T**). The reverse complement of this element with a portion of the binding site creates an element "TAACTG" which may compete somewhat effectively with some of the typical homeodomains.

Figure 4.7

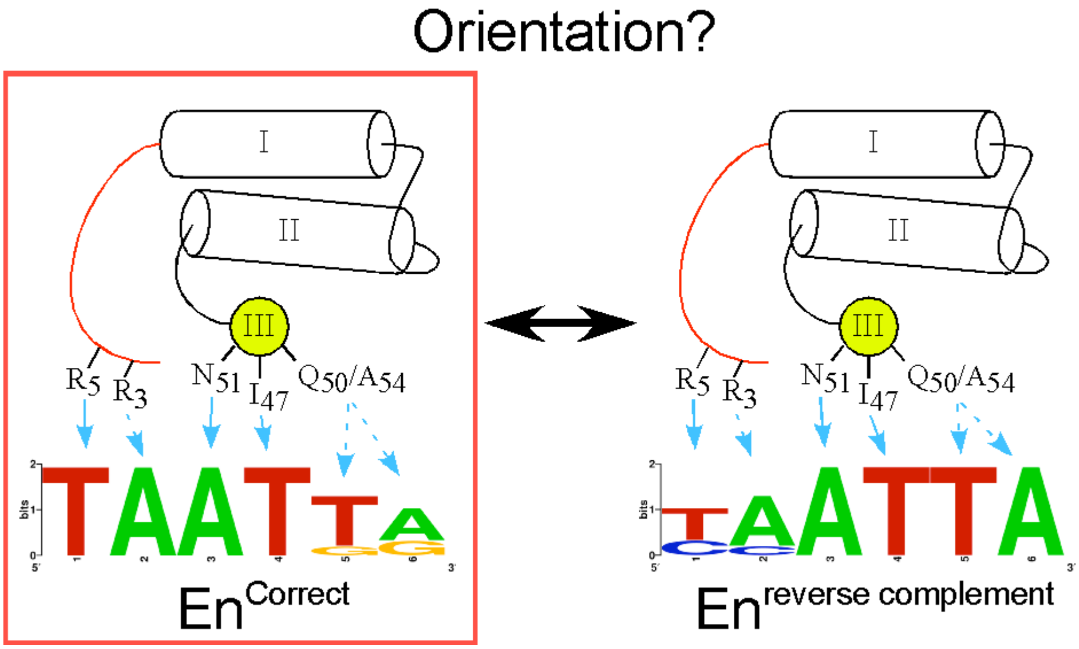

Orientation?
I
II
III
R5
R3
N51
I47
Q50/A54
bits
2
1
0
5′
3′
1
2
3
4
5
6
TAATTA
En Correct
I
II
III
R5
R3
N51
I47
Q50/A54
bits
TAATTA
En reverse complement

Figure 4.7. Determining the orientation of the homeodomain on a selected site. To correlate differences in homeodomain specificity with differences in their amino acids, it is also critical to determine the orientation of the homeodomain with respect to the motif. For example, if the consensus sequence for a factor defined from a selection is TCATTA, the N-terminal arm might either specify the TC at the 5' end of this sequence or the TA at the 5' end of the complimentary sequence TAATGA. In this study, base positions predicted to interact with the N-terminal arm are at the 5' positions of the motif (Figure 1). Previous structural and mutagenesis studies of well-characterized homeodomain proteins such as En, Ubx, and Bcd have defined the orientation of the homeodomain on its binding site (Ades and Sauer, 1994; Ekker et al., 1992; Fraenkel et al., 1998; Hanes and Brent, 1989; Percival-Smith et al., 1990; Treisman et al., 1989). For factors, such as En, that can bind a pseudo-C2 symmetric binding sequence (*e.g.* TAATTA), it is not immediately obvious from the binding site logo alone how the sequence motif should be orientated. Fortunately, the short tether between the omega-Zif12 module used for the B1H selection and the homeodomain biases the location of binding sites recovered from the randomized library. Homeodomain binding sites that neighbor the fixed Zif12 binding site are preferentially recovered on the DNA strand complementary to the Zif12 binding site. In contrast, more distant homeodomain sites are generally recovered on the same strand as the Zif12 binding site. This positional bias is most easily observed in the selected binding sites of a factor such as Bcd, which binds a clearly asymmetric binding sequence (TAATCC, Supplementary Table 2). This site bias can be used to define the

Figure 4.7 cont.

orientation of the homeodomain on the consensus sequence, which facilitates the comparison of recognition sequences for each homeodomain in a common register.

Table 4.1

Bicoid selected sequences

>bcd1
TGTTAATCCG
>bcd2
ATGGATTAGA
>bcd3
CGTTAATCTC
>bcd4
GGTTTAATCC
>bcd5
TCTATAATCC
>bcd6
TCTTAATCCC
>bcd7
GCTTAATCCG
>bcd8
GGGTTAATCC
>bcd9
GAGATAATCC
>bcd10
AGCTTATCC
>bcd11
CGGGTAATCC
>bcd13
TGTTAATCC
>bcd14
TGGGATTATA
>bcd15
GCGTAATCCA
>bcd16
GGCTTAAGCC
>bcd17
GGTTATCCG
>bcd18
TGTTAATCCC
>bcd20
CGCTTAATCC
>bcd21
TTACTAATCC
>bcd22
GTCCTAATCC
>bcd23
GGTTAATCCG
>bcd24
ATGGATTAGA

Deformed selected sequences

>dfd1
CTTCATTAAG
>dfd2
GGTCATTAAT
>dfd3
TATCATTAAA
>dfd4
GGTCATTAAT
>dfd5
GTCATTAACA
>dfd6
CCTAATTAAG
>dfd7
GATAATTAAT
>dfd8
CCTAATTAAG
>dfd9
CCCCATTAAT
>dfd10
TTTTTAATGA
>dfd11
AGCTATTAAA
>dfd12
GCACTAATGA
>dfd13
TCGTAATGA
>dfd14
TGCTTAATGG
>dfd15
ATCGTAATTA
>dfd16
CTCATTACT
>dfd17
CTTCATTAAG
>dfd18
AGTCATTAGG
>dfd19
TACCTAATGA
>dfd20
TGGATAATGA
>dfd21
CGACTAATGA
>dfd22
TATCATTAAC
>dfd23
CCGTTAATGA
>dfd24
CAATTAATGA

Engrailed selected sequences

>En1A1
CGCAATTAGA
>En1A2
GACTTAATGA
>En1A3
CTACTAATTG
>En1A4
CCAATTATTT
>En1A5
CCAATTACC
>En1A6
CATAATTAAA
>En1A7
CACAATTAAC
>En1A8
AGTAATTACC
>En1A9
CTAATTAGAG
>En1A10
TTAATTAGAC
>En1A11
TTAATTAGGT
>En1A12
CCAATTATGT
>En1B2
TTAATTAGTA
>En1B3
TTTTTAATTA
>En1B4
ATAATTAGTG
>En1B5
TCCAATTAAG
>En1B6
CCAATTAGAT
>En1B7
TAATTAACA
>En1B8
GGTAATTAAA
>En1B9
GGTAATTAAC
>En1B10
TCAATTAAGG
>En1B11
GCTAATTAAT
>En1*A1
GTGTTAATGA
>En1*A4
GCATAATTA
>En1*A5
ACAATTATAA
>En1*A6
ATAATTAAAA
>En1*A8
ATCATTAACC
>En1*A9
GATAATTATC
>En1*A10
TATCAACCCC
>En1*A11
CCTCATTAAA
>En1*B2
ATTAATTATC
>En1*B3
GTAATTAGG
>En1*B4
AGCAATTAAG
>En1*B5
TGTAATTAGA
>En1*B7
TTTAATTAAG
>En1*B8
TAGTTAATTA
>En1*B9
ATCAATTAAG
>En1*B10
CTCATTAAGA
>En1*B11
CTCATTAGTG

Table 4.1. Selected sequences for Bicoid, Deformed and En. All sequences are listed in the same orientation relative to the promoter of the reporter genes (toward the 3' side). Yellow highlight indicates the position of a binding site in the positive strand, magenta indicates the position of a binding site in the negative strand, and green indicates a palidromic site (orientation unknown). "Yellow" sites trend toward the 3' end of the sequence whereas "magenta" sites trend toward the 5' end of the sequence.

Global alignment and clustering of homeodomain binding sites

Remarkable diversity exists in the B1H-determined DNA-binding specificities for the entire set of homeodomains (Figure 4.3). The almost absolute conservation of Asn51, which specifies Ade at binding site position 3 (Fraenkel et al., 1998; Grant et al., 2000; Wolberger et al., 1991), in combination with our ability to infer the orientation of each homeodomain on its binding site (Figure 4.7 and Table 4.1) provides a basis for aligning all of these recognition sequences. Using this master alignment (Appendix Table A.3), hierarchical clustering of the *D .melanogaster* homeodomains was performed based on the similarity of their DNA-binding specificities (Figure 4.8). Based on this analysis the majority of these factors can be organized into eleven different specificity groups (Figure 4.8). We determined the average specificity of each group for the purposes of comparison (Group recognition motif, Figure 4.8B). In this calculation, we utilized the core 6 base pair element recognized by these factors, although some individual factors recognize larger motifs. Slightly more than half (43) of the homeodomains fall into the Antp or En specificity groups. There are also a number of specificity groups, such as the Abd-B and NK-1 group, which differ in sequence preference from the Antp or En groups at only one or two positions. Other groups, such as the TGIF-Exd group, differ at four or five positions relative to the Antp or En groups. Outside of these specificity groupings are five factors, such as CG11617, that exhibit unique specificities. This diversity of specificities reveals the adaptability of the homeodomain architecture for the recognition of a variety of DNA sequences.

Figure 4.8

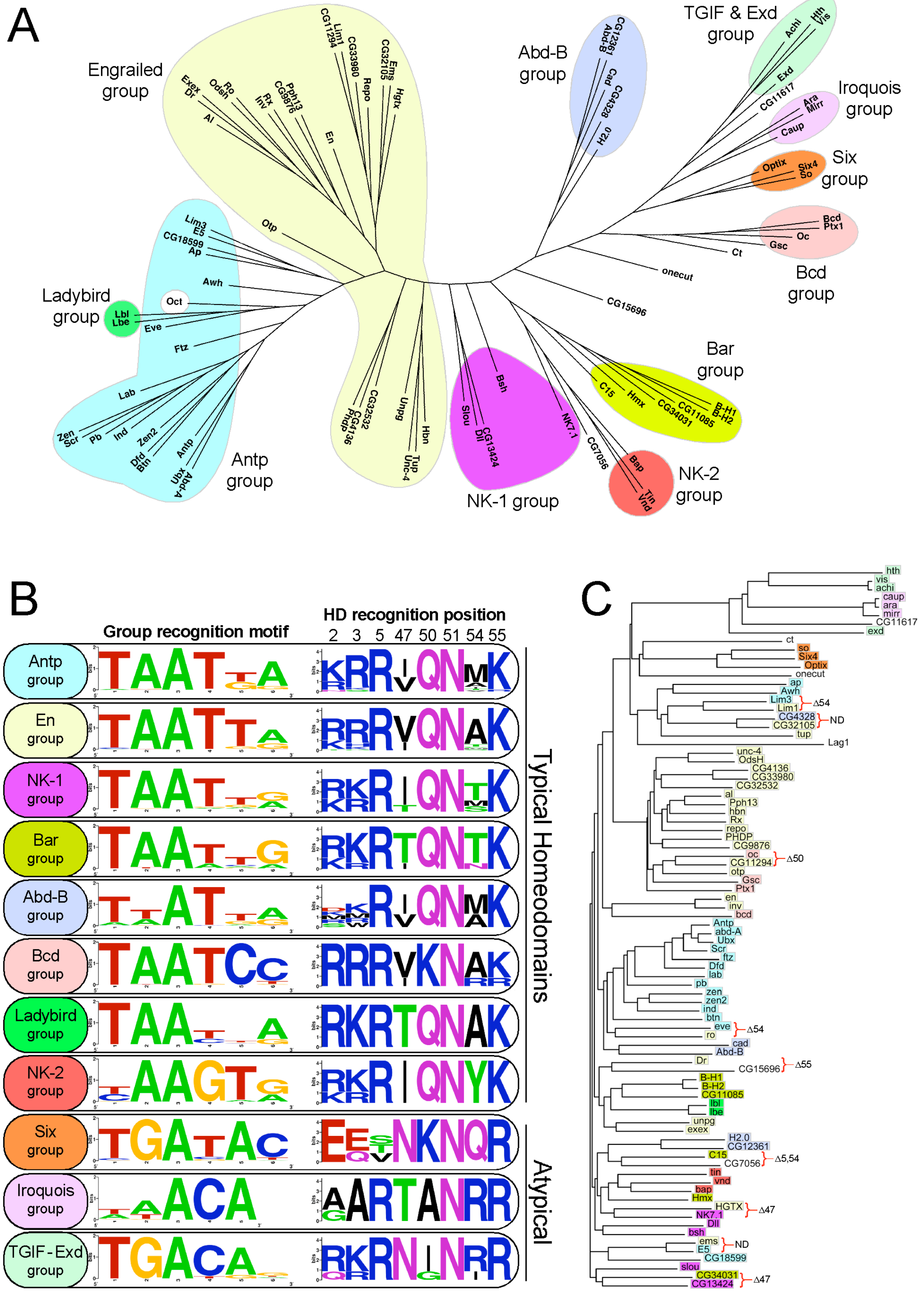

A
Engrailed group
Abd-B group
TGIF & Exd group
Iroquois group
Six group
Bcd group
Bar group
NK-2 group
NK-1 group
Antp group
Ladybird group
Hth
Vis
Achi
Exd
CG11617
Ara
Mirr
Caup
Optix
Six4
So
Bcd
Ptx1
Oc
Gsc
Ct
onecut
CG15696
B-H1
B-H2
CG11085
CG34031
Hmx
C15
Bap
Tin
Vnd
CG7056
NK7.1
Bsh
CG13424
Dll
Slou
CG12361
Abd-B
Cad
CG4328
H2.0
CG11294
Lim1
CG33980
Ems
CG32105
Repo
Hgtx
Exex
Dr
Ro
OdsH
Rx
Inv
Pph13
CG9876
Al
En
Otp
Unpg
Hbn
Tup
Unc-4
CG32532
CG4136
PhdP
Lim3
E5
CG18599
Ap
Awh
Oct
Lbl
Lbe
Eve
Ftz
Lab
Zen
Scr
Pb
Ind
Zen2
Dfd
Btn
Antp
Ubx
Abd-A
B
Group recognition motif
HD recognition position
2 3 5 47 50 51 54 55
Antp group
En group
NK-1 group
Bar group
Abd-B group
Bcd group
Ladybird group
NK-2 group
Six group
Iroquois group
TGIF-Exd group
TAATTA
TAATTA
TAATTG
TAATTG
TTATTA
TAATCC
TAATTA
TAAGTG
TGATAC
AAACA
TGACA
KRRIQNMK
RRRVQNAK
RKRIQNTK
RKRTQNTK
RRIQNMK
RRRVKNAK
RKRTQNAK
RKRIQNYK
EESNKNQR
AARTANRR
RKRNINRR
Typical Homeodomains
Atypical
C
hth
vis
achi
caup
ara
mirr
CG11617
exd
ct
so
Six4
Optix
onecut
ap
Awh
Lim3
Lim1
Δ54
CG4328
CG32105
ND
tup
Lag1
unc-4
OdsH
CG4136
CG33980
CG32532
al
Pph13
hbn
Rx
repo
PHDP
CG9876
oc
CG11294
Δ50
otp
Gsc
Ptx1
en
inv
bcd
Antp
abd-A
Ubx
Scr
ftz
Dfd
lab
pb
zen
zen2
ind
btn
eve
ro
Δ54
cad
Abd-B
Dr
CG15696
Δ55
B-H1
B-H2
CG11085
lbl
lbe
unpg
exex
H2.0
CG12361
C15
CG7056
Δ5,54
tin
vnd
bap
Hmx
HGTX
Δ47
NK7.1
Dll
bsh
ems
E5
ND
CG18599
slou
CG34031
CG13424
Δ47

Figure 4.8. Specificity-based clustering of the 84 Drosophila homeodomains. (A) Based on the sequence similarity between the clustered recognition motifs, these factors can be organized into eleven different specificity groups. Oct1 (Oct) was also included this analysis, but was not included in a specificity group. (B) The typical and atypical homeodomains are distributed into different groups, with the typical groups listed at the top. The average specificity of each group is indicated under the Group recognition motif, and to the right is the Sequence logo of the key homeodomain recognition positions for each group. (C) The specificity groups (colored rectangles) are mapped onto the homeodomain amino acid sequence similarity tree. In instances where highly similar neighbors have been assigned into different specificity groups (10 pairs indicated by red brackets) any difference in residue type at a key recognition position (5, 47, 50, 54 or 55) is noted (ND = No difference).

Homeodomains that share strong sequence similarities are not always found in the same specificity group. We mapped the specificity group assignments onto an amino acid neighbor-joining tree to examine the correlation between overall sequence similarity and DNA-binding specificity within the homeodomain family (Figure 4.8C). As expected, when there is a strong similarity in amino acid sequence for two factors they usually share a common specificity. However, in 10 instances, two factors share strong sequence similarity, but fall into different specificity groups. In eight of these comparisons, this difference can be explained by the presence of a different residue at one or more of the key DNA-recognition positions (5, 47, 50, 51, 54 and 55, see below). Pairs of factors with high overall sequence similarity, but different specificities due to changes in their specificity determinants, may represent recently diverged gene duplications where one factor has acquired new target genes and a different *in vivo* function.

Distinguishing features of homeodomain specificity groups

Clear correlations exist between the specificity of each group and the amino acid distributions that are present at the key DNA recognition positions (Figure 4.8B). Below, we compare and contrast these amino acid distributions to highlight distinguishing features of the different homeodomain specificity groups. In some cases, the contribution of specific residues toward specificity for one or more group members has been demonstrated in previous mutagenesis or structural studies. This dataset allows the categorization of these determinants in the context of the full range of *Drosophila* homeodomain specificities.

Typical superclass

Antp and En groups: The largest and most extensively-studied groups of homeodomains provide our benchmark for describing how differences in amino acid sequence correlate with DNA-binding specificity. The Antp and En groups share very similar recognition motifs and corresponding similarities in their amino acid distributions at the key recognition positions. One difference in specificity is at binding site position 5: the En class prefers Thy, whereas the Antp class tolerates either Gua or Thy. There is a corresponding difference between these groups in the most common residue at position 54: Ala for the En group and Met for the Antp group. In the Antp-DNA structure, the side chain of Met54 is close (4.25Å) to binding site position 5, which could directly influence the preferred base (Fraenkel and Pabo, 1998). This correlation is not absolute, as a small number of factors with Ala at position 54 are found in the Antp class.

Bcd group: Typical homeodomains containing Lys50 (Bcd, Oc, Gsc and Ptx1) prefer Cyt at binding site positions 5 and 6. The importance of Lys in defining this preference is well-established (Hanes and Brent, 1989; Percival-Smith et al., 1990; Treisman et al., 1989) and is a consequence of a direct contact between Lys50 and the complementary guanines at binding site positions 5 and 6 (Tucker-Kellogg et al., 1997). These factors

Figure 4.9

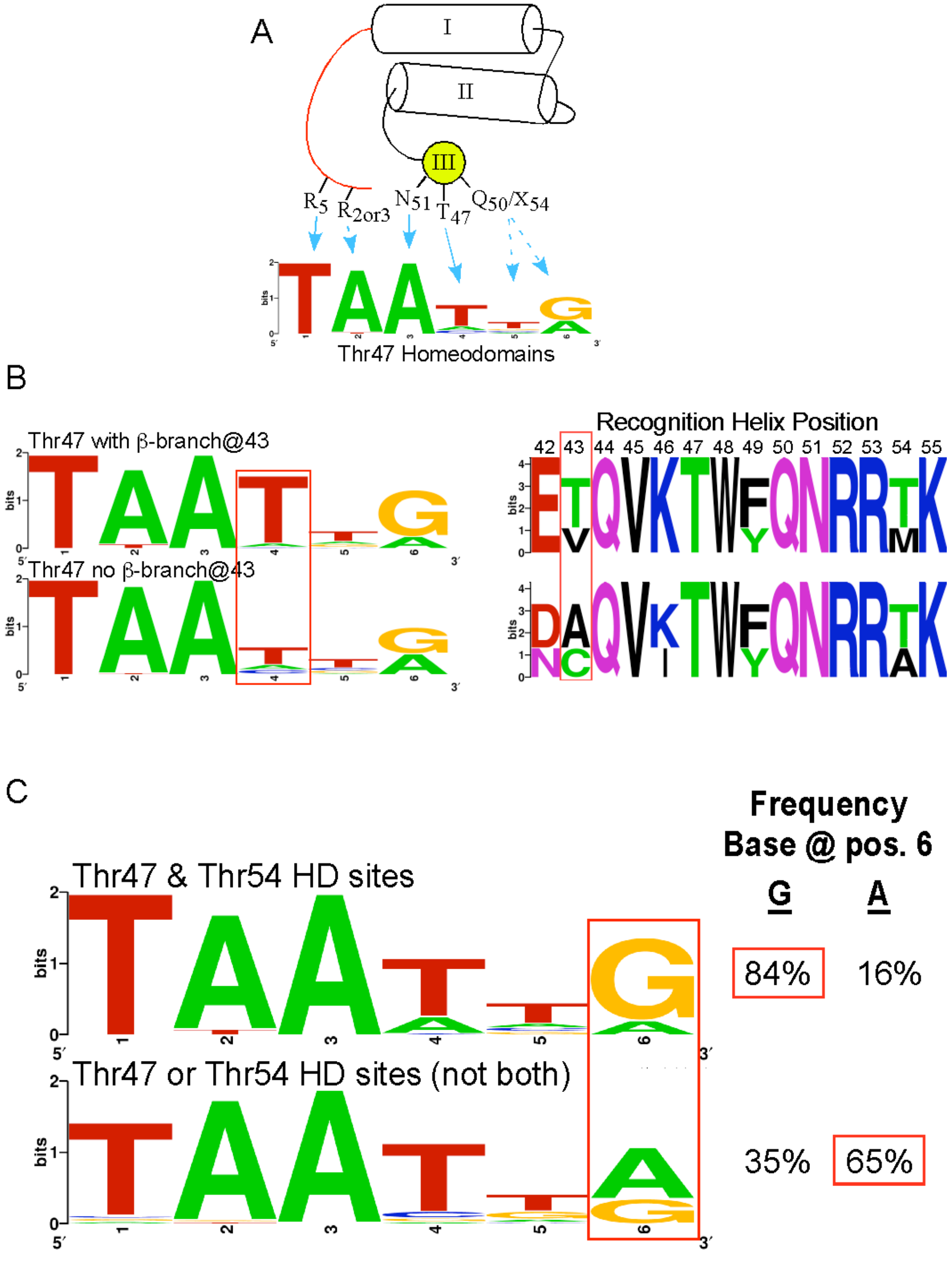

A
I
II
III
R5
R2or3
N51
T47
Q50/X54
Thr47 Homeodomains
B
Thr47 with β-branch@43
Thr47 no β-branch@43
Recognition Helix Position
42 43 44 45 46 47 48 49 50 51 52 53 54 55
C
Thr47 & Thr54 HD sites
Thr47 or Thr54 HD sites (not both)
Frequency
Base @ pos. 6
G
A
84%
16%
35%
65%
bits
5′
3′

Figure 4.9. Secondary effects influence recognition for Thr47 containing homeodomains. A) Schematic of the potential interactions between the Thr47 containing homeodomains and their binding sites. When the eight typical homeodomains that contain Thr at position 47 are considered together, they display a clear reduction in specificity at binding site positions 4 and 5 of their as assessed by the information content at these positions. B) This set of Thr47-containing factors can be divided into two sub-sets with regards to the degree of preference for Thy at position 4 of their binding site. Factors that contain a β-branched amino acid at position 43 (Bsh, CG11085 & CG34031) display a strong preference for Thy at position 4, while the remaining five factors (Lbe, Lbl, B-H1, B-H2 & C15) display only a weak preference The difference between these two motifs is statistically significant (p-value = 2.48e-4). Because amino acids at positions 43 and 47 are on the same face of the recognition helix and are in van der Waals contact in some crystal structures (Piper et al., 1999), it is plausible that the type of residue at position 43 could bias the preferred conformation of Thr47 and thereby influence its binding site preference. The Sequence logos of the recognition helices for each subgroup (not corrected for small sample size) are shown to the right of their motifs. C) The combination of Thr47 and Thr54 leads to a preference for Gua over Ade at binding site position 6 for the factors within the Bar group. Five of these factors (BH-1, BH-2, CG11085, CG34031 & C15) contain Thr at positions 47 and 54 of the recognition helix and a display a strong preference for Gua at binding site position 6, a preference that is not observed in any of the other homeodomain groups. Ten other homeodomains (Bsh, Lbl, Lbe, NK7.1, CG7056, Slou, CG13424, Hgtx, E5 & Ems) that contain Thr at

Figure 4.9 cont.

either but not both positions, display a preference for an alternate base, typically A, at position 6. This difference is statistically significant (p-value = 7.05e-18). Mechanistically, how the presence of Thr at positions 47 and 54 would create a preference for Gua at position 6 of the binding site is not obvious as these residues are separated by 2 helical turns and appear more appropriately positioned to influence specificity at binding site positions 4 and 5. This combination could indirectly affect specificity at binding site position 6 through an influence on the sequence preference of Gln50.

prefer the classical TAAT motif at binding site positions 1 through 4 and correspondingly have canonical (En/Antp-like) recognition residues at most other recognition positions.

NK-1, Bar and Ladybird groups: Many of these homeodomains are members of the NK or DL homeodomain classes (Banerjee-Basu and Baxevanis, 2001), which generally have Thr at position 47 or 54, and have specificities similar to the Antp and En groups. Those homeodomains with Thr47 generally display reduced specificity at binding site positions 4 and 5 (Figure 4.9A) with the exception of homeodomains with a beta-branched amino acid at position 43. These homeodomains display a stronger preference for Thy at position 4 (Figure 4.9B). Within the Bar group, factors that contain both Thr47 and Thr54 display a unique preference for Gua at binding site position 6 (Figure 4.9C).

NK-2 group: The three members of this group display a unique preference for Gua at binding site position 4, due to an interaction between Tyr54 and the complementary Cyt in this base pair (Gruschus et al., 1997). Their specificities differ only at binding site position 1, which correlates with amino acid differences at positions 6 and 7 of the N-terminal arm (Damante et al., 1996) (Figure 4.10).

Abd-B group: These factors display a slight preference for Thy over Ade at binding site position 2. A diverse set of amino acids is present at positions 2 and 3 of the N-terminal arm when compared with other groups of typical homeodomains, which frequently have Arg or Lys at both positions. Although the preference for Thy at binding site position 2

Figure 4.10

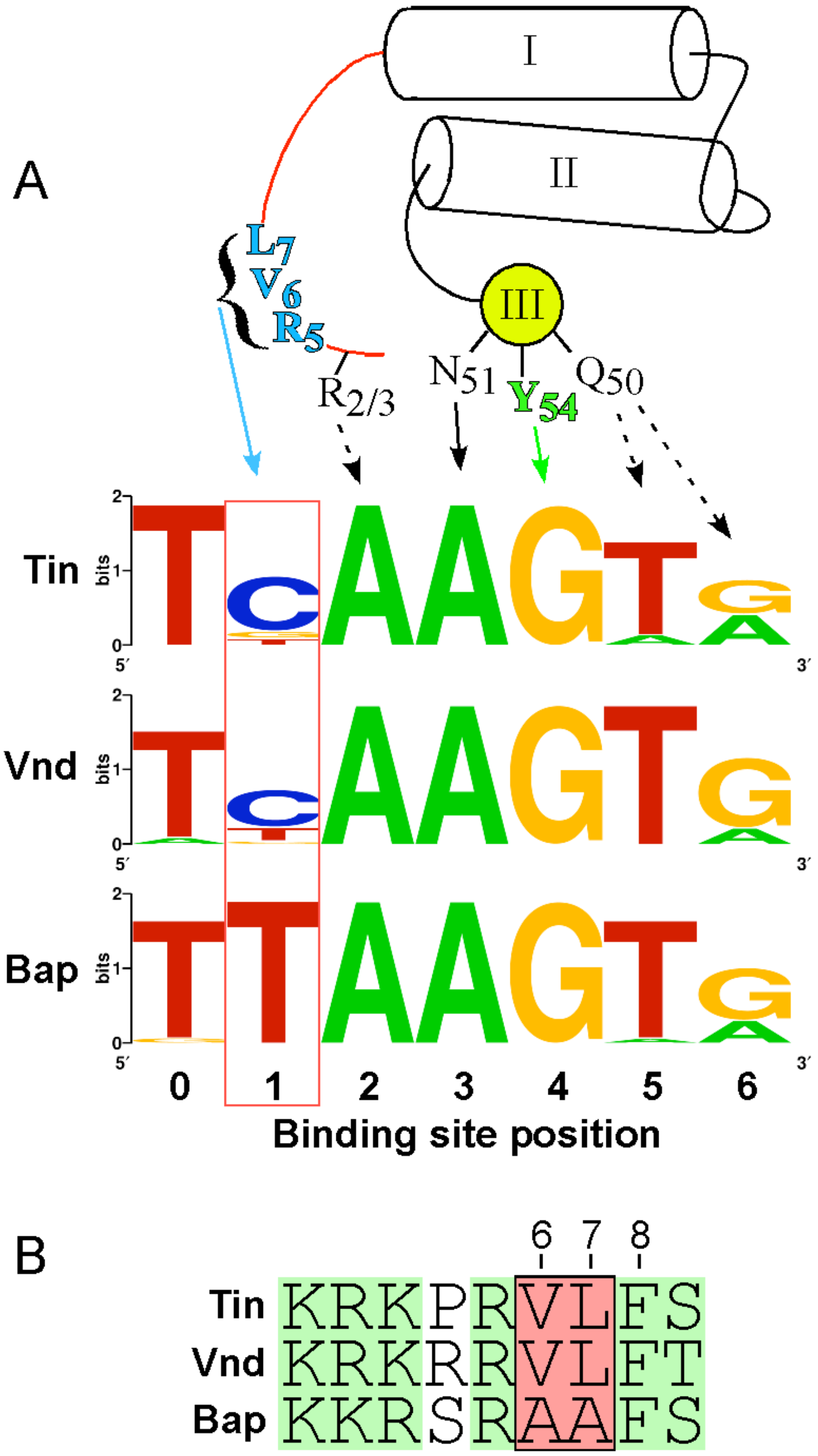
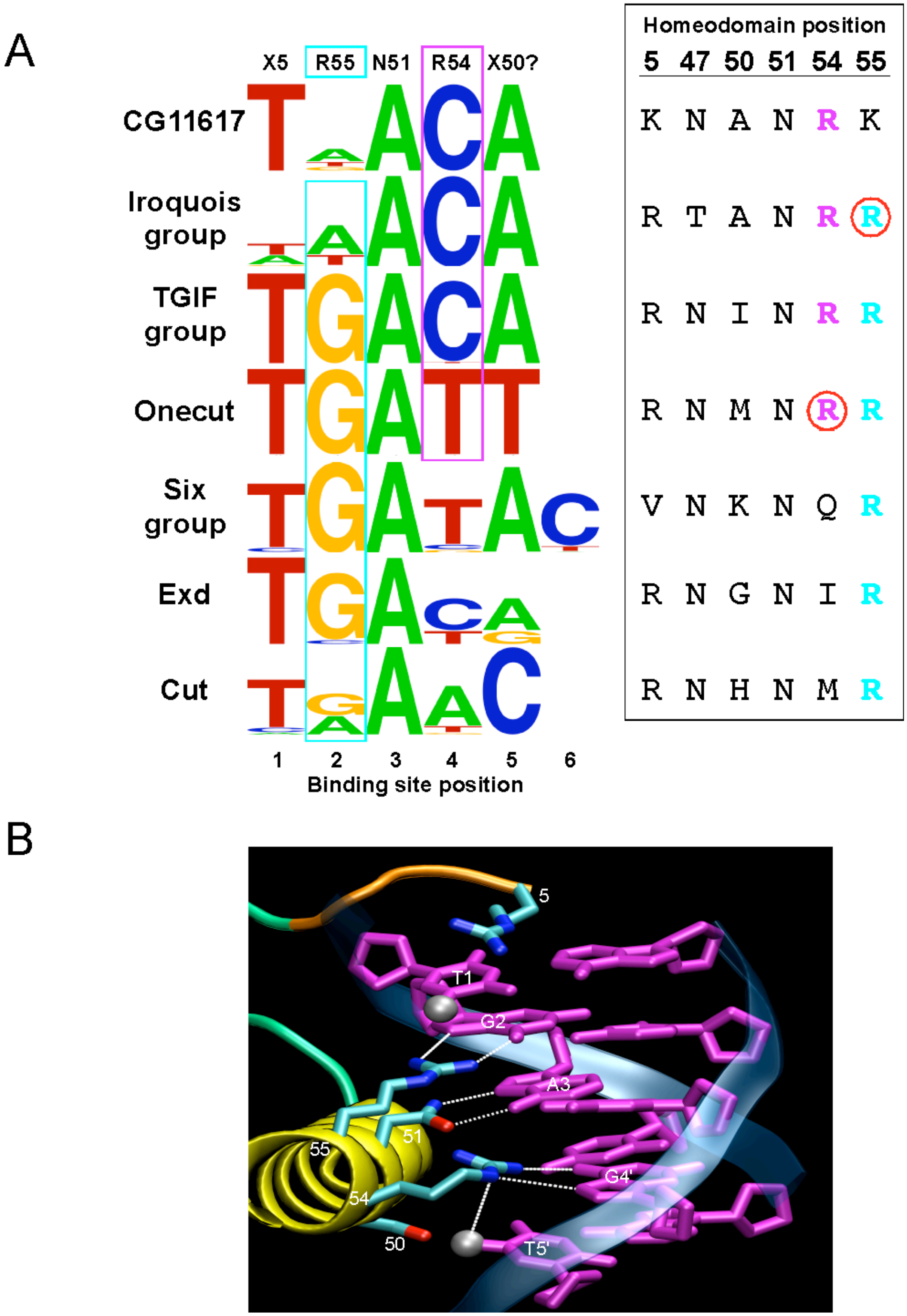

A
I
II
III
L7
V6
R5
R2/3
N51
Y54
Q50
Tin
Vnd
Bap
bits
2
1
0
5′
3′
0
1
2
3
4
5
6
Binding site position
B
6
7
8
Tin KRKPRVLFS
Vnd KRKRRVLFT
Bap KKRSRAAFS

Figure 4.10. Residues at positions 6 and 7 of the N-terminal arm can influence the specificity at position 1 of the binding site. A) The three NK-2 class family members that share Tyr at position 54 display very similar overall specificities with the exception of binding site position. Bap (NK3) prefers a **<u>T</u>**AAGTG sequence, while Tin (NK4) and Vnd (NK2) prefer **<u>C</u>**AAGTG. A vertebrate homolog of Bap, Nkx3.2 displays an identical consensus sequence TAAGTG (Kim et al., 2003). The modest preference for Cyt at binding site position 1 has been observed for homologs of Tin and Vnd in other species and has been particularly well characterized in bovine TTF-1, which recognizes CAAGTG(Damante et al., 1996). Damante and colleagues narrowed the residues responsible for the Cyt preference at position 1 to the residues at positions 6 through 8 (VLF) in the N-terminal arm(Damante et al., 1996). They demonstrated that when these residues are substituted for the QTY in the N-terminal arm of Antp, its specificity was altered from a strong preference for Thy at binding site position 1 to a tolerance for Cyt or Thy. B) Consistent with this analysis, both Tin and Vnd have VLF motif at these positions in the N-terminal arm, while Bap contains an AAF motif. Our results allow us to further narrow the key residues responsible for the alteration in the specificity observed in the NK-2 family at binding site position 1 to residues at positions 6 and 7 in the N-terminal arm, since F is shared between all three NK-2-type factors. Only one other *D. melanogaster* homeodomain, Cut, has the VL motif at positions 6 and 7 of the N-terminal arm and it tolerates either Thy or Cyt at position 1 in the binding site, although there is a modest preference for Thy (Figure 4.4). This Thy preference in Cut may be due to the presence of a contact to binding site position 2 by Arg55, which can

Figure 4.10 cont.

influence the preference for Thy at position 1 in the binding site as described in the text (Figure 4.11B). A structural explanation for this influence is unavailable, but it is possible that the packing of the VL motif against the homeodomain fold positions a backbone carbonyl to favorably interact with the N2 position of the complementary Gua in the minor groove.

Figure 4.11



Homeodomain position	5	47	50	51	54	55
CG11617	K	N	A	N	R	K
Iroquois group	R	T	A	N	R	R
TGIF group	R	N	I	N	R	R
Onecut	R	N	M	N	R	R
Six group	V	N	K	N	Q	R
Exd	R	N	G	N	I	R
Cut	R	N	H	N	M	R

Figure 4.11. Correlations are observed within the atypical homeodomains between the presence of Arg at positions 54 and 55 and the preferred specificity at binding site positions 2 and 4. A) (Left) Sequence logos for the different types of atypical homeodomains (either groups or outliers). (Right) The amino acid sequences at the key DNA contact positions for each factor or group. When Arg is present at position 54 (magenta) there is usually a preference for Cyt at binding site position 4. When Arg is present at position 55 (cyan) there is usually a preference for Gua at binding site position 2. Notable exceptions are indicated by red circles. B) Structural model of DNA recognition for atypical family members containing Arg at positions 54 and 55 constructed from a superposition of the contacts observed in the MATα2-DNA (Wolberger et al., 1991) and Exd-Ubx-DNA structures (Passner et al., 1999). These arginines potentially specify not only the contacted Gua, but also the 5’ Thy due to the favorable van der Waals interaction (~4 Å) with the T-methyl group (silver sphere).

in Abd-B has been mapped to amino acid positions 3, 6 and 7 of the N-terminal arm (Ekker et al., 1994), the variability within this group at these positions prevents a straightforward correlation of binding site preference and amino sequence.

Atypical homeodomains

The atypical superclass is distinguished from the typical superclass by different amino acid preferences at a number of positions within the homeodomain fold and by the presence of a three amino acid loop extension (TALE) between helix 1 and 2 (Figure 4.4). The *D. melanogaster* atypical homeodomains comprise three different specificity groups with additional outliers (e.g. Cut, Onecut and CG11617, Figure 4.11A).

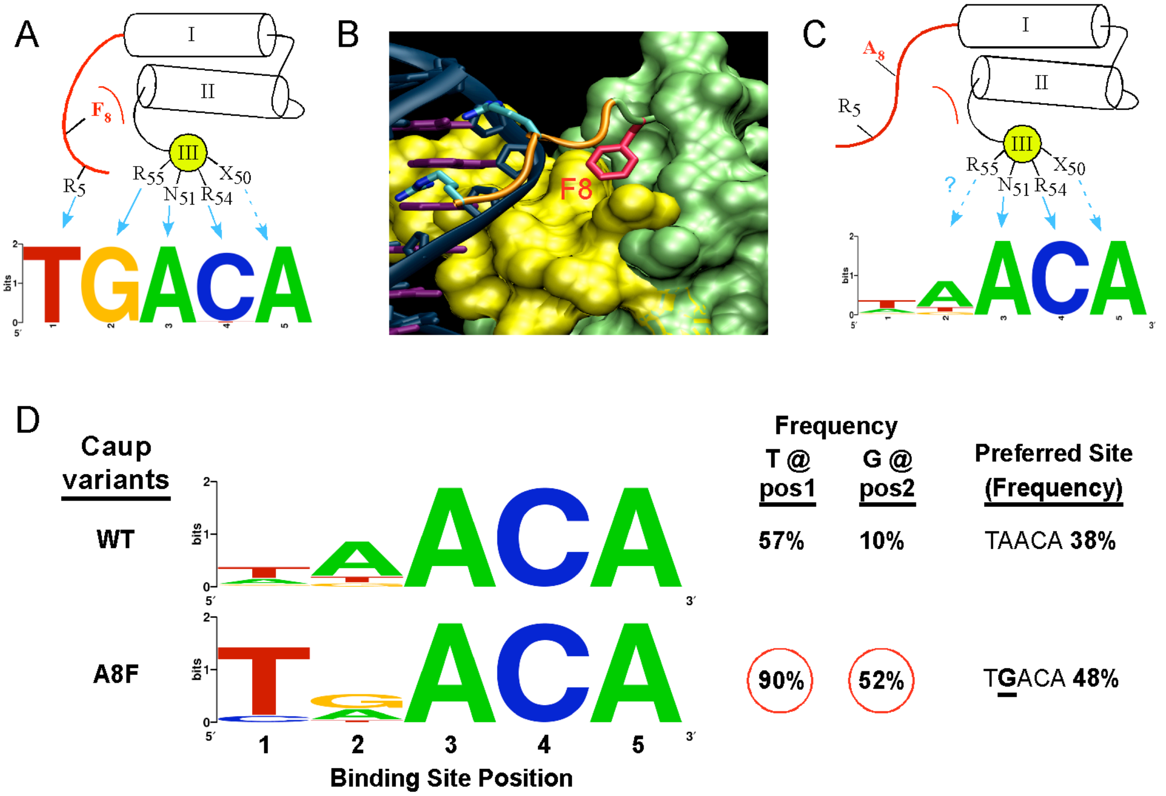
Overall, Recognition motifs for the atypical groups are distinguished by a tendency toward Gua at binding site position 2, and Cyt and Ade at positions 4 and 5 (Figures 4.8B and 4.11A). In CG11617 and the Iroquois and TGIF groups, the preference for Cyt and Ade at positions 4 and 5 correlates with the presence of Arg54, which is consistent with the contact between Arg54 and the complementary Gua and Thy at positions 4 and 5 described in the MATα2 co-complex crystal structures (Wolberger et al., 1991) (Figure 4.11B). The single exception to this correlation, Onecut, contains a unique specificity determinant (Met50), which may be responsible for its distinct recognition preference. Likewise, with the exception of the Iroquois group, homeodomains that contain Arg55 prefer Gua at position 2, suggesting that the contact observed between this amino acid and base in the Exd and Pbx co-complex crystal structures (Passner et al., 1999; Piper et al., 1999) is another general recognition feature (Figure 4.11B). Consistent with this model, CG11617, which has Lys55, has reduced specificity at position 2 (Figure 4.6).

TGIF-Exd group: Our data is consistent with the described specificities for individual members of the TGIF - Exd group (Bertolino et al., 1995; Chang et al., 1996).

Six group: All members of this group (So, Six4 and Optix) display a specificity that overlaps with the recognition motif TGATAC and share identical residues at the key DNA-recognition positions (47, 50, 51, 54 and 55). Our recognition motif is consistent with binding sites extracted from footprinting data for So ((T/C)GATAC) (Hazbun et al., 1997), but is inconsistent with the specificity (TAAT) reported for murine Six3, an Optix homolog (Zhu et al., 2002). Both Six3 and Optix contain Arg55 and so would be expected to specificity G instead of A at position 2. This discrepancy is examined in more detail below.

Iroquois group: Our motif (ACA) represents the specificity of a monomer, whereas the previous described specificity of a full-length protein (Mirr) revealed that it binds as a homodimer to the palindrome ACANNTGT (Bilioni et al., 2005). All three members of the Iroquois group (Caup, Mirr and Ara) exhibit a weak preference for Thy at position 1 despite the presence of Arg5 in the N-terminal arm, which can specify this position in other homeodomains (Passner et al., 1999; Piper et al., 1999), and exhibit a weak preference for Ade or Thy instead of Gua at position 2 despite the presence of Arg55.

Figure 4.12



Caup variants	Frequency T @ pos1	Frequency G @ pos2	Preferred Site (Frequency)
WT	57%	10%	TAACA **38%**
A8F	90%	52%	TGACA **48%**

Figure 4.12. The role of position 8 in organizing the N-terminal arm. A) Most homeodomains have a large hydrophobic residue at position 8 that docks into a pocket formed by the three-helix bundle of the homeodomain fold. This interaction anchors the N-terminal arm over the minor groove for DNA-recognition. B) Surface rendering of the homeodomain (residues 9-60, recognition helix shown in yellow; Msx-1 structure (Hovde et al., 2001)). Phe8 (red) sits in a structural pocket and directs the N-terminal arm (orange) over the minor groove of the DNA. C) Iroquois family members have Ala at position 8 and as a consequence the N-terminal arm is free to sample other conformations that reduce the specificity of the factor. D) Reintroduction of the Phe at position 8 in Caup (A8F) dramatically alters the specificity of the protein at positions 1 and 2 of the binding site.

One striking difference between the Iroquois group and all other homeodomains (typicals and atypicals) is the presence of Ala instead of a large hydrophobic residue at position 8 (Figure 4.4). The large hydrophobic residue binds in a cleft against the homeodomain helices and appears to position the N-terminal arm over the 5' end of the binding site (Figure 4.12). This residue can influence the efficiency of DNA-binding, as mutation of Phe8 to Ala in Bcd attenuates its DNA-binding affinity (Subramaniam et al., 2001). To test if the limited specificity of the Iroquois family at the 5' end of their site is due to the absence of this interaction, a Ala8Phe mutation was introduced into Caup (Figure 4.12D). Remarkably, this single mutation increases Thy specificity at position 1 and Gua specificity at position 2. The relatively weak specificity for Gua at position 2 suggests that another feature within the homeodomain architecture may directly or indirectly mask the ability of Arg55 to specify this base; one potential confounding feature is Glu59, which could create a cationic trap in conjunction with the phosphodiester backbone (Figure 4.13).

Overall, a comparison of the typical and atypical specificity groups suggests two overlapping, but distinct sets of protein-DNA interactions (Figure 4.8B and 4.11B). In both, Arg5 and Asn51 generally specify Thy and Ade at binding site positions 1 and 3. These contacts in conjunction with a common set of phosphate contacts that are shared by 95% of the homeodomains in this set imply that the docking geometry of all of these homeodomains with the DNA will be similar (Figure 4.4). The specificity differences arise from different combinations of residues on this common recognition platform,

Figure 4.13

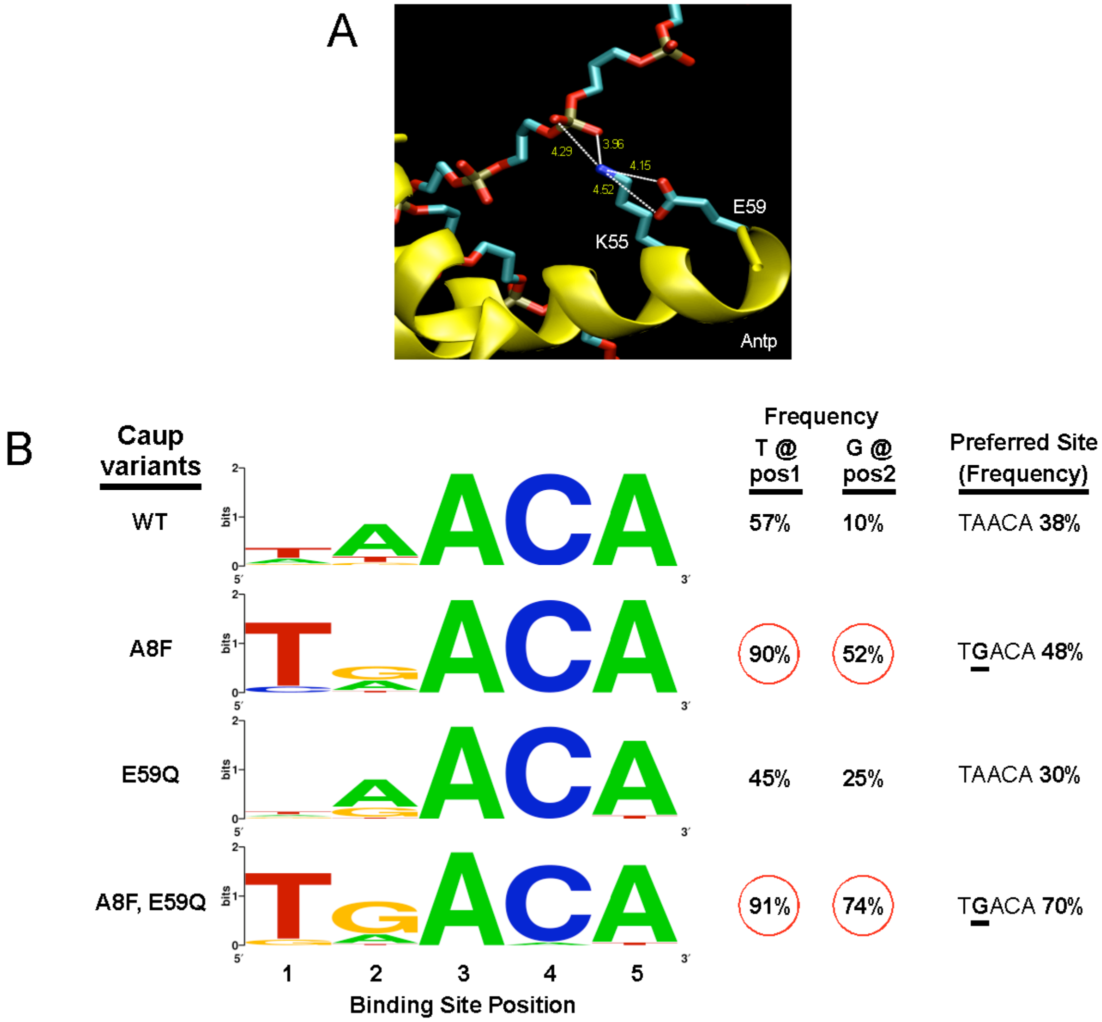

A
4.29
3.96
4.15
4.52
E59
K55
Antp
B
Caup variants
Frequency
T @ pos1
G @ pos2
Preferred Site (Frequency)
WT
57%
10%
TAACA 38%
A8F
90%
52%
TGACA 48%
E59Q
45%
25%
TAACA 30%
A8F, E59Q
91%
74%
TGACA 70%
bits
2
1
0
5′
3′
1
2
3
4
5
Binding Site Position

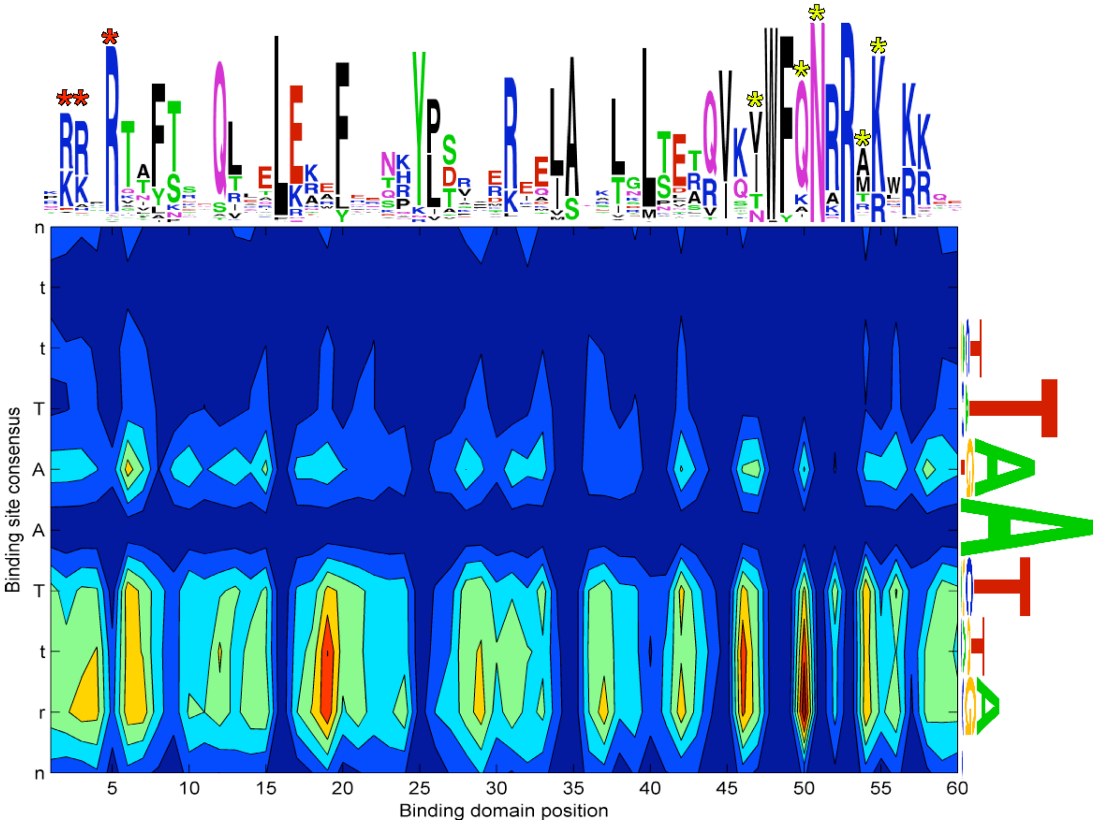
Figure 4.13. Glu59 may influence the specificity imparted by Arg55 in Caup. The presence of glutamate at position 59 of the homeodomain may also be influencing the degree of preference for Gua at position 2 of the binding site by capturing Arg55. A) A "cationic trap" is present in the Antp structure where Lys55 is captured against the phosphate backbone by Glu59 (Fraenkel and Pabo, 1998). Dotted lines indicate the distances between the lysine amino group and the negatively charged oxygens on the phosphodiester backbone and the acidic side-chain. The Iroquois family has a conserved glutamate at this position that could potentially capture Arg55 against the phosphodiester backbone. B) Mutation of Glu59 to Gln in the presence or absence of the A8F mutation results in an increase in the frequency of recovery of Gua at position 2 in the binding site. However, this change in specificity is modest when compared to the effect of the A8F mutation on specificity.

which can be analyzed as a set to elucidate how each position in the binding site is influenced by the homeodomain sequence.

Common specificity determinants for homeodomain proteins

We have undertaken both computational and qualitative approaches to decipher the common DNA-binding determinants for homeodomains. Our dataset provides an excellent resource for the computational identification of correlations between amino acid sequence and base preferences within the binding site, as it contains significantly more breadth (1860 different binding sites for 84 different proteins) than any previous catalog of homeodomain DNA-binding specificities. Mutual information (MI) analysis was used to identify potential specificity determinants by evaluating positions within our set of homeodomains that co-vary with changes at specific positions in the binding site (Gutell et al., 1992; Mahony et al., 2007). Protein and DNA positions that have high MI scores (i.e. their variation appears correlated) are candidates to interact either directly or indirectly. MI analysis identified a number of protein/DNA positions that appear to be correlated, some of which correspond to expected interactions based on previously identified DNA recognition positions (Appendix Table A.4). However, simple MI analysis is complicated by the limited variability at some individual positions (Figure 4.14). In order to normalize the values over both columns and rows of the MI matrix, we transformed the MI matrix into a joint rank product matrix, which provides a more uniform landscape (Figure 4.15). Prominent features of this plot correspond to expected homeodomain-DNA interactions. For example, strong MI is observed between

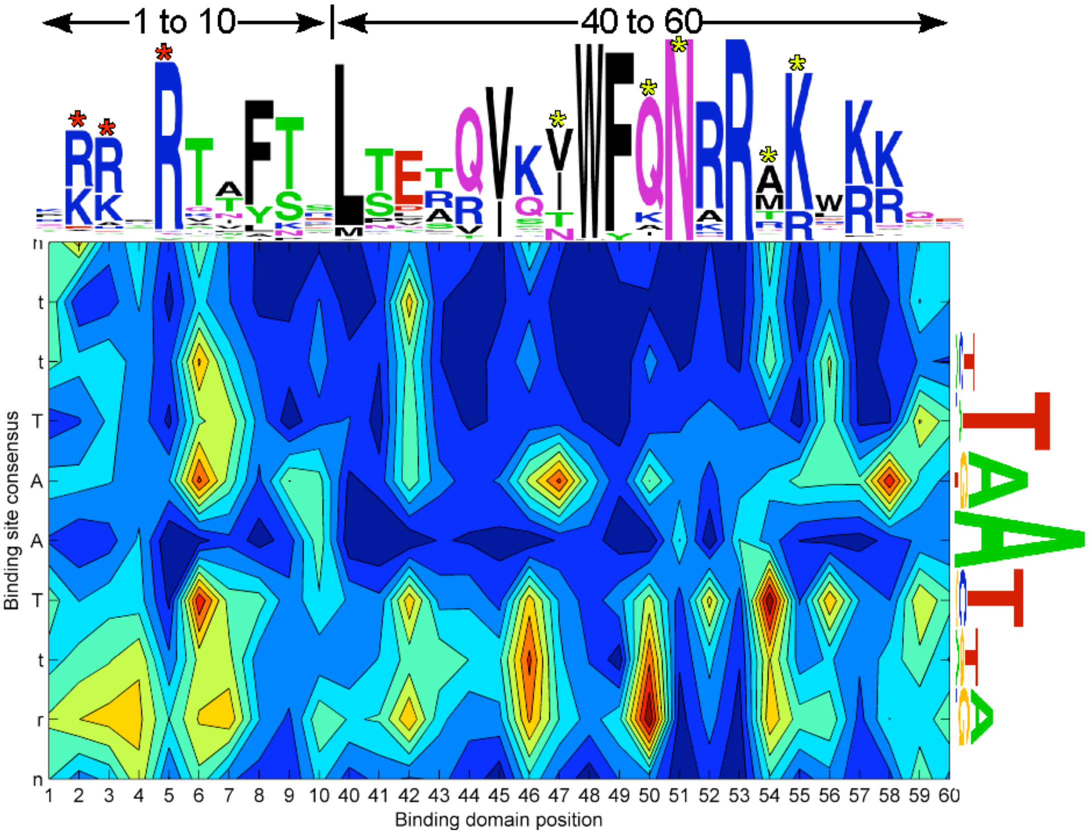
Figure 4.14



Binding site consensus
n
t
t
T
A
A
T
t
r
n
5
10
15
20
25
30
35
40
45
50
55
60
Binding domain position

Figure 4.14. Heat plot of Mutual Information (MI) analysis between the amino acid sequences of the homeodomain and the master alignment of their selected binding sites. A two-dimensional heat plot is shown where background levels of covariance are indicated in dark blue and positions with strong covariance are indicated in red. The x-axis indicates the position within the homeodomain (1 through 60), where the TALE insertions in the atypical family have been removed to achieve a common sequence framework. The amino acid diversity at each position is indicated by a Sequence logo above the plot. Key DNA-recognition positions are indicated by asterisks above the amino acid sequence. The y-axis indicates the position within the binding site (5' top, 3' bottom). The overall motif of all of the binding sites is represented as a Sequence logo to the right of the plot. Overall the degree of covariation is higher at the 3' end of the binding site than for the 5' end of the site. Positions with little diversity in sequence composition have lower MI values, which is highlighted by the trough in the plot at the position corresponding to the Ade at position 3, which is highly conserved among the population of binding sites. Strong peaks are observed within the plot for expected base-amino acid contacts, such as between position 50 of the homeodomain and positions 5 and 6 of the binding site.

Figure 4.15



1 to 10
40 to 60
Binding site consensus
n
t
t
T
A
A
T
t
r
n
1 2 3 4 5 6 7 8 9 10 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
Binding domain position

Figure 4.15. Heat plot of the joint-ranked Mutual Information (jrMI) analysis. The background levels of covariance are indicated in dark blue and positions with strong covariation are indicated in red. The x-axis indicates the position within the homeodomain alignment. Only positions 1 through 10 and 40 through 60 are shown to highlight the regions that are most likely involved in recognition. The amino acid diversity at each position is indicated by the Sequence logo above the plot. Key DNA-recognition positions are indicated by asterisks. The y-axis denotes the position within the binding site (5’ top, 3’ bottom, with positions 1-4 capitalized) where the overall motif of the aligned binding sites is represented by a Sequence logo.

recognition helix positions 50 and 54 and binding site positions 6 and 4, respectively. Both of these peaks correctly predict positions of interaction between the homeodomain and DNA. Strong peaks are also observed between positions 6 and 47 of the homeodomain and binding site position 2. Position 6 on the N-terminal arm is expected to affect the specificity of binding site position 2 (Ekker et al., 1994). However, the correlation at position 47 is likely due to evolutionary linkage; the type of residue at position 47 is correlated to the superclass of the homeodomain (atypical or typical) where each superclass typically prefers different bases at binding site position 2. Although this evolutionary history complicates the MI analysis, a number of known homeodomain-DNA interactions are recovered and other homeodomain positions are identified that may be new hallmarks for predicting specificity based on amino acid sequence.

While the MI analysis can predict positions involved in determining DNA-binding specificity, it does not identify which amino acids lead to different binding site preferences. This information can be extracted from our dataset by examining the correlations between amino acid sequence and recognition preference in the context of the large body of reported structural and biochemical studies on homeodomain recognition. The keystone for this analysis is the Asn51-mediated recognition of Ade at binding site position 3. In the presence of an alternate amino acid-base combination at this position, other inferences about the key specificity determinants may not be valid. Below, the residues that frequently contribute to specificity are summarized for each position in the binding site (Figure 4.16).

Figure 4.16

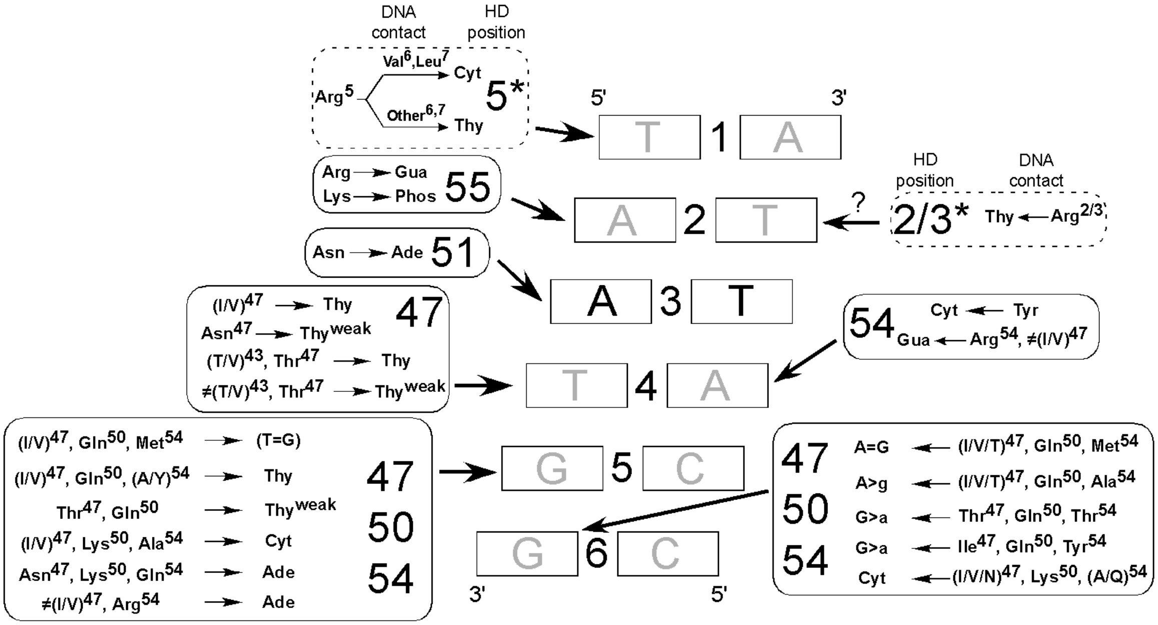

DNA contact
HD position
Arg^{5}
Val^{6},Leu^{7} → Cyt
Other^{6,7} → Thy
5*
5'
3'
T 1 A
Arg → Gua
Lys → Phos
55
A 2 T
?
HD position
DNA contact
2/3*
Thy ← Arg$^{2/3}$
Asn → Ade
51
A 3 T
(I/V)^{47} → Thy
Asn^{47} → Thy^{weak}
(T/V)^{43}, Thr^{47} → Thy
≠(T/V)^{43}, Thr^{47} → Thy^{weak}
47
T 4 A
54
Cyt ← Tyr
Gua ← Arg^{54}, ≠(I/V)^{47}
(I/V)^{47}, Gln^{50}, Met^{54} → (T=G)
(I/V)^{47}, Gln^{50}, (A/Y)^{54} → Thy
Thr^{47}, Gln^{50} → Thy^{weak}
(I/V)^{47}, Lys^{50}, Ala^{54} → Cyt
Asn^{47}, Lys^{50}, Gln^{54} → Ade
≠(I/V)^{47}, Arg^{54} → Ade
47
50
54
G 5 C
G 6 C
3'
5'
47
50
54
A=G ← (I/V/T)^{47}, Gln^{50}, Met^{54}
A>g ← (I/V/T)^{47}, Gln^{50}, Ala^{54}
G>a ← Thr^{47}, Gln^{50}, Thr^{54}
G>a ← Ile^{47}, Gln^{50}, Tyr^{54}
Cyt ← (I/V/N)^{47}, Lys^{50}, (A/Q)^{54}

Figure 4.16. Catalog of common specificity determinants for Asn51-containing homeodomains. The specificity of a homeodomain for its binding site can be inferred from the amino acid sequence that is present at positions 2, 3, 5, 43, 47, 50, 51, 54 and 55. Amino acid positions that are most likely to influence the sequence preference of the homeodomain at a particular position in the binding site are indicated in boxes (solid line – major groove, dotted line – minor groove) surrounding the core 6 bp binding element. An arrow points from the box of potential interactions to the base within each base pair that it describes. For simplicity some interactions, such as Lys50 with binding site positions 5 and 6, are described as influencing specificity on the primary strand of the DNA when in reality direct contacts are made to the complementary strand. The DNA base specified by a certain type of amino acid at each position is indicated in each box. Where multiple amino acid positions can influence specificity at a particular binding site position, the position of each amino acid is indicated as a superscript. The prediction of specificity of a homeodomain based on the amino acids at these recognition positions will not be perfect, as many factors can influence the sequence preference of these determinants.

BS Position 1: 89% of the aligned recognition sequences have Thy at this position. Consistent with this preference, the majority of homeodomains (94%) have Arg5 in the N-terminal arm, which specifies Thy (Ades and Sauer, 1995). In the Six family, Thy is preferred in the absence of Arg5, which is potentially due to the favorable stacking of the T-methyl group over Arg55 as it interacts with Gua at position 2 of the binding site (Figure 4.11B). In the NK-2 class, Val6 and Leu7 promote a preference for Cyt over Thy (Damante et al., 1996).

BS Position 2: Preferences for Ade, Gua or Thy are observed among the different homeodomains. 83% of the aligned sequences have Ade at this position. Consistent with this preference, most typical homeodomains contain Arg2 or Arg3 within the N-terminal arm, which structural and biochemical studies have implicated in biasing specify toward Ade (Ades and Sauer, 1995; Hovde et al., 2001). However, both Lab and HGTX, which lack either arginine, strongly prefer Ade, which is consistent with reports of an additional unidentified specificity determinant (Ades and Sauer, 1995). Members of the Abd-B group display a preference for Thy over Ade, but the responsible specificity determinants are unclear. In atypical homeodomains Arg55 specifies Gua; however, several typical homeodomains contain Arg55, yet display a preference for Ade (e.g. Awh and Bcd), indicating that additional residues contribute to specification of Gua.

BS Position 3: Asn51 specifies Ade at this position.

BS Position 4: Any base can be specified at this position. Thy is the most common base (80%) and is strongly correlated with the presence of Ile or Val at position 47. The three exceptions that contain Val at position 47 - Bap, Tin and Vnd – are NK-2 class homeodomains, which prefer Gua due to a direct contact by Tyr54 to the complementary strand Cyt (Gruschus et al., 1997). Cyt specificity is strongly correlated with the presence of Arg54, which interacts with the complementary Gua. Bcd is one notable exception; it contains both Ile47 and Arg54, yet displays strong preference for Thy, which suggests that the base preference of Ile47 may supercede Arg54 (explored below). Decreased specificity at position 4 is correlated with either Thr47 or Asn47, although the presence of a β-branched amino acid at position 43 modifies the degree of preference (Figure 4.9B). Finally, Cut displays a unique preference for Ade at position 4; His50, which only occurs in this homeodomain, may be responsible.

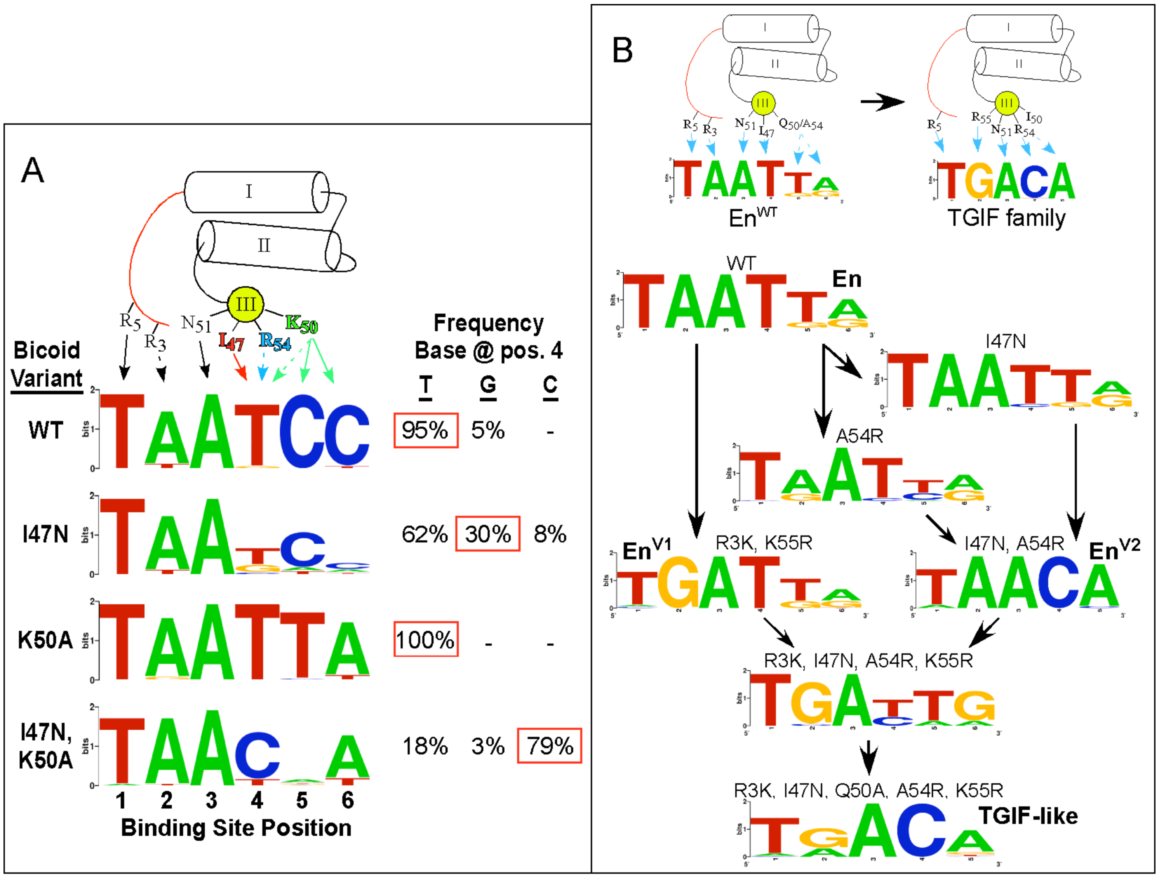
BS Position 5: For many specificity groups correlations exist between the residues at positions 47, 50 and 54 and certain base preferences. Specificity groups that have Gln50 and either Ile47 or Val47 display preferences that correlate with the residue at position 54. When Met54 is present (Antp group) a tolerance is observed for either Gua or Thy, whereas when Ala54 (En group) or Tyr54 (NK-2 group) are present, a preference for Thy is observed. When Lys50 occurs with Ile47 or Val47 and Ala54 (Bcd group) a strong preference for Cyt is observed (Ades and Sauer, 1994; Hanes and Brent, 1989; Percival-Smith et al., 1990; Treisman et al., 1989). However, Lys50 in the presence of Asn47 and Gln54 (Six family) specifies Ade instead of Cyt. An influence of position 54 on the specificity mediated by Lys50 has also been observed in other contexts (Pellizzari et al., 1997). Finally, when Arg54 is present in the absence of Ile47 or Val47 (Iroquois and TGIF groups) a preference for Ade is observed, which is likely due to preferential stacking of the complementary T-methyl group over the guanidinium group of Arg54 as it interacts with Gua at position 4 (Figure 4.11B). Typical homeodomains containing Thr47 display relaxed specificity at position 5.

BS Position 6: Like position 5, the determinants of specificity at this position appear influenced by the residues at positions 47, 50 and 54. Ade is modestly preferred over Gua when Gln50 and Met54 (Antp group) or Ala54 (En group) are present. Conversely, Gua is preferred over Ade when both Thr47 and Thr54 are present with Gln50 (Bar and NK-1 groups, or less exclusively when Tyr54 is present with Ile47 and Gln50 (NK-2 group). Finally, the presence of Lys50 provides a preference for Cyt at position 6 (Bcd and Six groups). The majority of atypical homeodomains with the exception of the Six group display no strong preference at this position.

Competing contact residues in Bcd

One complicating feature of these specificity determinants, as observed for other types of DNA-binding domains, is the absence of a one-to-one correlation between the presence of an individual residue at a certain position and the specificity at a particular binding site position; residues at different positions on the homeodomain can potentially

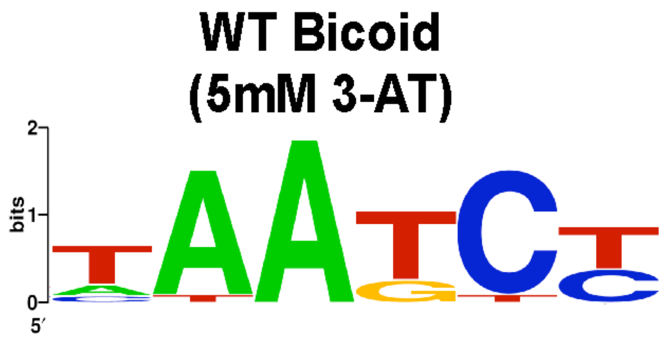
Figure 4.17



Bicoid Variant	Frequency Base @ pos. 4 T	G	C
WT	95%	5%	-
I47N	62%	30%	8%
K50A	100%	-	-
I47N, K50A	18%	3%	79%

Figure 4.17. Exploring DNA-binding specificity through mutagenesis. A) Mutational analysis of the specificity determinants for binding site position 4 in Bcd. Amino acids at three different positions (Ile47, Lys50 and Arg54) can potentially influence the base preference at position 4. The wild-type protein prefers Thy. Three different mutants (I47N, K50A and I47N with K50A) were characterized to determine the alteration in base preference at this position. The frequency that each base was recovered at position 4 is indicated to the right of the Sequence logo for each factor. B) Conversion of Engrailed (En) into a homeodomain with TGIF-like specificity. (Top) Schematic representation of the critical base contacts responsible for specificity in En and TGIF family members. (Bottom) Flow diagram of the mutations required to complete the specificity conversion. Two intermediate specificity conversions (En^{V1} and En^{V2}) were obtained first, and these mutations were combined along with Q50A to produce TGIF-like specificity.

Figure 4.18



Frequency Base @ pos. 4

T	G	C
64%	32%	-

Figure 4.18. Binding site selection data for Bicoid collected from a selection performed at 5mM 3-AT instead of the typical stringency of 10 mM. The sequence preferences are preserved in this binding site selection except that Cyt and Thy are equally represented at position 6. There is increased tolerance for other bases at many positions, such as position 4 where Thy and Gua are both acceptable, but Cyt is still excluded from the binding sites at this position.

interact with a common base pair to coordinately influence the base preference. For example, residues at positions 47 and 54 can both directly influence the base preference at binding site position 4 (Fraenkel et al., 1998; Gruschus et al., 1997; Wolberger et al., 1991). When multiple determinants with different preferences for a single binding site position are present, it is possible that these potential interactions compete. Consequently, the overall specificity of a homeodomain may be a composite of more than one binding site motif.

We have used Bcd to explore whether competing interactions can contribute to binding site specificity. Bcd contains Ile47 and Arg54, which appear to respectively specify Thy and Cyt at binding site position 4 in other homeodomains. At this position Bcd displays a strong preference for Thy, a weak preference for Gua and no evidence of tolerance for Cyt (Figure 4.17A). This hierarchy of preferences is also observed in a binding site selection performed at lower stringency (Figure 4.18). A weak tolerance for Gua is consistent with previous SELEX-based characterization of Bcd (Wilson et al., 1996) and is likely a consequence of Lys50. In the structure of the Engrailed Q50K mutant with the binding site TAATCC, this Lys has two different conformations: one conformation interacts with the O6 positions of Gua5' and Gua6' (' indicates the complementary strand) while the other conformation interacts with the O4 of Thy4 and the O6 of Gua5' (Tucker-Kellogg et al., 1997). The latter interaction would likely be more favorable with the O6 positions of Gua4 and Gua5' in the context of the sequence TAAGCC (e.g. see (Kim and Berg, 1996)). In fact, Bcd has been previously shown to bind specifically to the sequence TAA**<u>G</u>**CT (Dave et al., 2000), although the authors ascribed the preference for Gua at position 4 to Arg54, which appears unlikely given our mutational results (described below) and the fact that the Oc homeodomain protein, which also contains Lys50 but lacks Arg54, displays a weak preference for Gua at position 4 based on our data and previous SELEX data (Wilson et al., 1996).

The absence of Cyt in the recognition motif suggests that Ile47 or Lys50 may prevent Arg54 from contributing to the base preference at position 4. To help define the hierarchy of interactions that influence specificity, single and double mutations were examined at positions 47 and 50. When Ile47 is mutated to Asn, a residue commonly found in atypical homeodomains that contain Arg54, the specificity of Bcd was modestly altered (Figure 7A). In this mutant, the relative preference for Thy and Gua is preserved, but a tolerance for Cyt is also apparent, possibly due to the increased influence of Arg54. Moreover, every occurrence of Cyt is in the context of the sequence TAA**<u>CAC</u>** (3 of 34), which is consistent with Arg54 specifying Cyt and Ade at positions 4 and 5, respectively (Figure 4.11B), and Lys50 specifying Cyt at position 6 as seen in the Six group. When Lys50 is mutated to Ala, a complete shift to an En-like specificity (TAATTA) is observed with no evidence of other specificity determinants influencing position 4. In the double mutant Ile47Asn and Lys50Ala, a preference for Cyt at position 4 - the base specified by Arg54 in most atypical homeodomains - is revealed. Thus, three different potential specificities are embedded within the Bcd scaffold as a consequence of the residues at positions 47, 50 and 54. Lys50 and Arg54 are less influential likely because they have greater flexibility, which allows them to obtain other favorable interactions: Lys50 with bases at positions 5 and 6, and Arg54 with the phosphodiester backbone.

Engineering the DNA-binding specificity of En

To assess the utility of our catalog of specificity determinants, we attempted to shift the specificity of En, a typical homeodomain (TAATTA), to that of a TGIF-type atypical homeodomain (TGACA). These homeodomains differ in sequence preference at four of the six recognition positions (Figure 4.17B), and share ~28% sequence identity overall. While small numbers of substitutions have been previously introduced into homeodomains to alter their specificity at one or two binding site positions, attempts to produce dramatic changes in specificity, such as grafting the specificity of a typical homeodomain onto an atypical homeodomain, have failed (Mathias et al., 2001).

En was chosen as our scaffold for engineering because of its detailed biochemical and structural characterization (Ades and Sauer, 1994, 1995; Fraenkel et al., 1998; Grant et al., 2000; Tucker-Kellogg et al., 1997). Two partial specificity conversions were carried out in parallel (Figure 4.17B): One conversion (variant En^{V1}) to alter specificity at position 2 (T**G**ATTA), and another conversion (variant En^{V2}) to alter specificity at positions 4-6 (TAA**CA**). Two mutations (R3K and K55R), which both influence the base preference at position 2 (Figure 4.16), were sufficient to generate a strong preference for Gua (Figure 4.17B). Likewise, two mutations (I47N and A54R), which influence specificity at the 3' end of the recognition sequence (Figure 4.16 and 4.17A), were sufficient to promote the En^{V2} specificity. Consistent with the Bcd mutagenesis data, individual substitutions did not achieve the desired alteration in specificity. It is noteworthy that the specificity of the En^{V2} mutant (TAA**<u>CA</u>**) was obtained in the presence of Gln50, which is not found in atypical homeodomains but is almost universally conserved among typical homeodomains. In this particular context, the Arg54 specificity determinant is dominant, which is consistent with previous studies that found no evidence of a prominent role in specificity for Gln50 in En (Ades and Sauer, 1994; Grant et al., 2000).

The two pairs of mutations (R3K, I47N, A54R and K55R) present in En^{V1} and En^{V2} were combined in anticipation that they would produce the desired composite specificity (TGACA). However, this mutant displayed a 3' specificity that is intermediate between an En-like and TGIF-like specificity (TGA(T/C)(T/A)(G/A); Figure 7B). Even though Gln50 appears passive in the En^{V2} mutant, its potential influence on specificity in this context was examined by mutation; including Q50A provides an almost complete conversion to the desired TGACA specificity. When this motif is included with the other fly homeodomains to analyze the overall similarity of all of the recognition motifs, it clusters with the TGIF-Exd group (Figure 4.19). The less pronounced preference for Gua at binding site position 2 relative to a TGIF homeodomain suggests that additional residues in atypical homeodomains can contribute to specificity at this position. This nearly complete transformation of binding specificity demonstrates that En is a robust scaffold for engineering novel DNA-binding specificities while highlighting the complex interplay between residues at the recognition surface that can influence specificity of the entire motif.

Figure 4.19

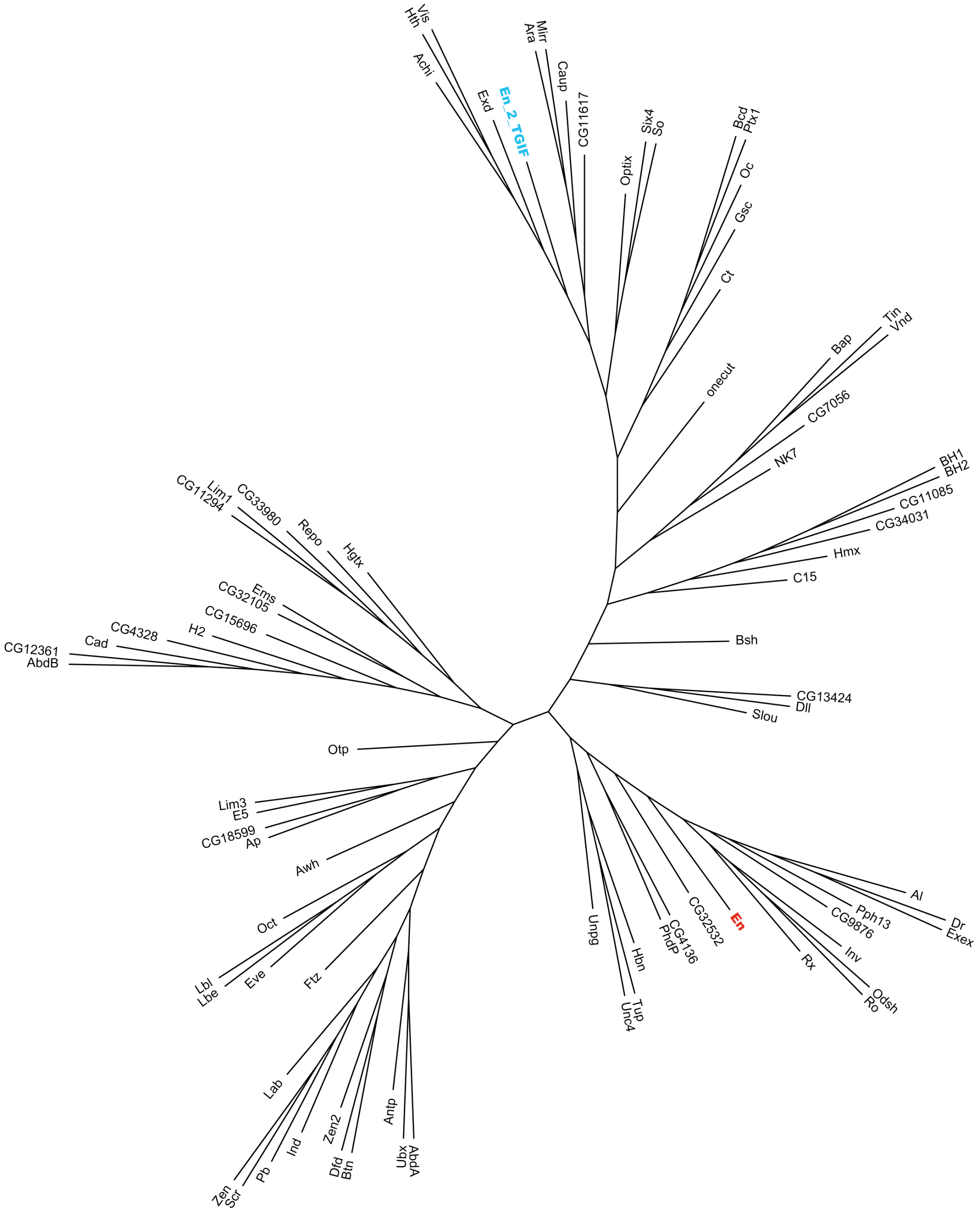
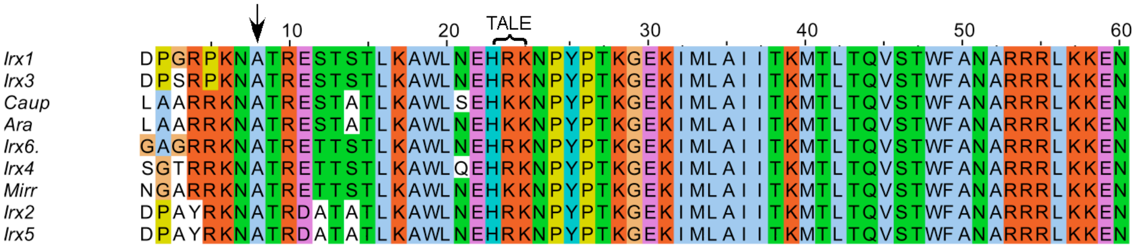

Vis
Hth
Achi
Exd
En_2_TGIF
Mirr
Ara
Caup
CG11617
Optix
Six4
So
Bcd
Ptx1
Oc
Gsc
Ct
onecut
Bap
Tin
Vnd
CG7056
NK7
BH1
BH2
CG11085
CG34031
Hmx
C15
Bsh
CG13424
Dll
Slou
Al
Dr
Exex
Pph13
CG9876
Inv
Odsh
Ro
Rx
En
CG32532
CG4136
PhdP
Hbn
Tup
Unc4
Unpg
Antp
Ubx
AbdA
Zen2
Dfd
Btn
Ind
Pb
Lab
Zen
Scr
Ftz
Eve
Lbl
Lbe
Oct
Awh
CG18599
Ap
Lim3
E5
Otp
CG12361
AbdB
Cad
CG4328
H2
CG15696
CG32105
Ems
CG11294
Lim1
CG33980
Repo
Hgtx

Figure 4.19. Clustergram of engineered En homeodomain that has TGIF-like specificity (**En_2_TGIF**; R3K, I47N, Q50A, A54R, K55R) with the entire set of fly homeodomains. The specificity of this factor clusters with the other members of the TGIF-Exd specificity group (Vis, Achi, Hth & Exd), which is significantly removed from its original position in the clustergram (**En**) with the En-group.

Figure 4.20



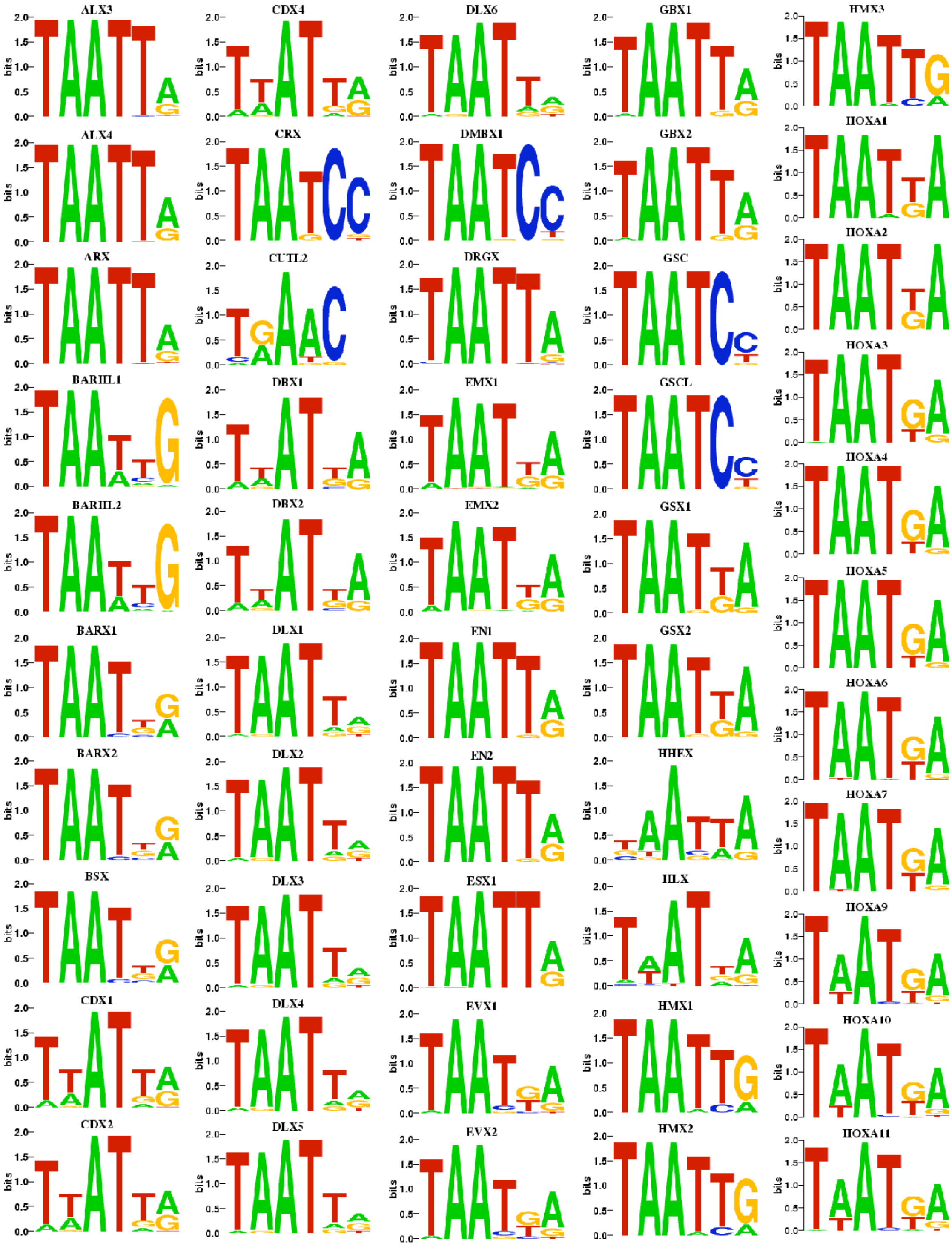
```
Irx1   DPGRPKNATRESTSTLKAWLNEHRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
Irx3   DPSRPKNATRESTSTLKAWLNEHRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
Caup   LAARRKNATRESTATLKAWLSEHKKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
Ara    LAARRKNATRESTATLKAWLNEHKKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
Irx6.  GAGRRKNATRETTSTLKAWLNEHRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
Irx4   SGTRRKNATRETTSTLKAWLQEHRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
Mirr   NGARRKNATRETTSTLKAWLNEHKKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
Irx2   DPAYRKNATRDATATLKAWLNEHRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
Irx5   DPAYRKNATRDATATLKAWLNEHRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
```

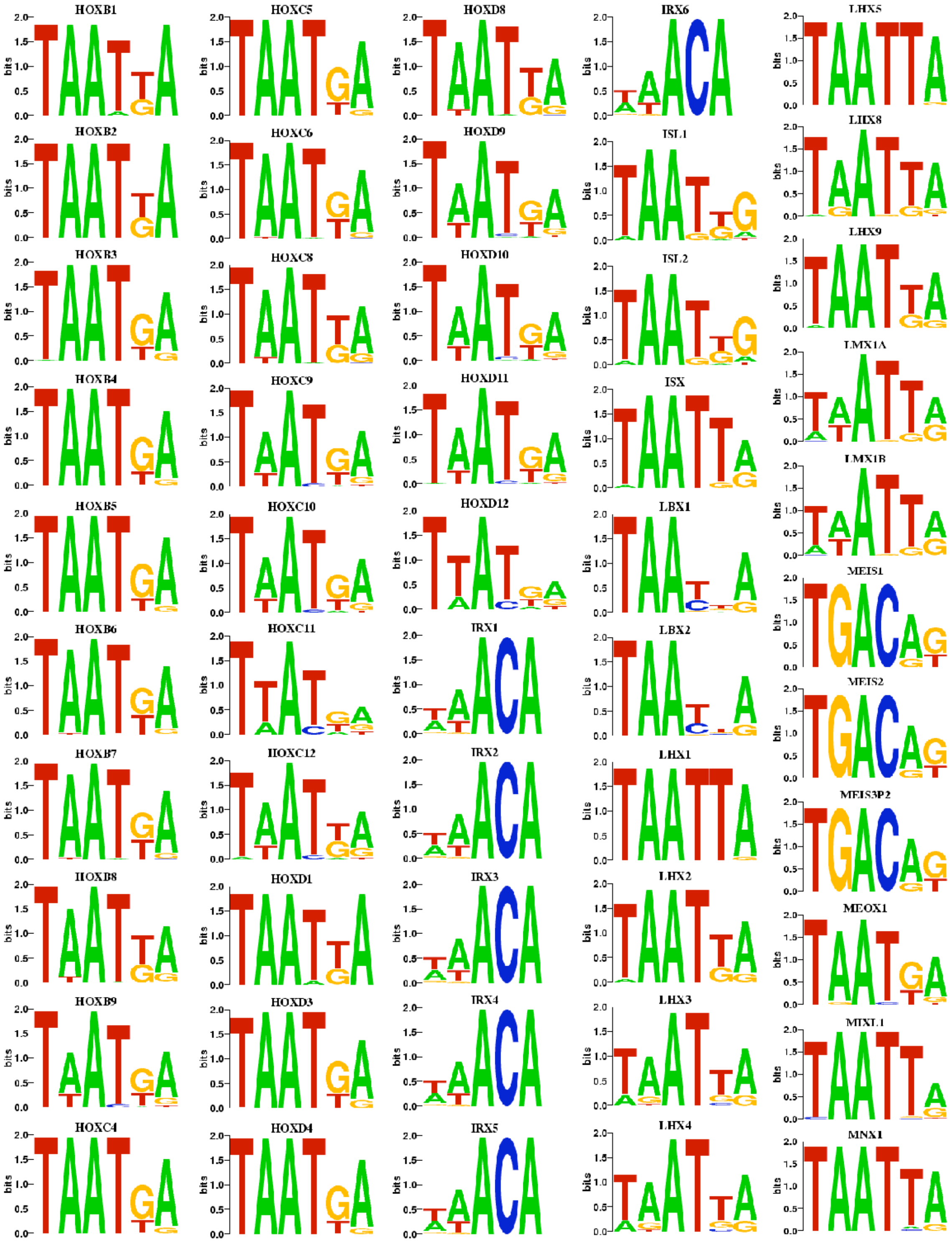
Figure 4.20. ClustalW alignment of the *D. melanogaster* and *H. sapiens* Iroquois family members. The numbering scheme has been adjusted to ignore the presence of the TALE insertion in this atypical family. Position 8 is indicated by an arrow, where all nine proteins have an alanine at this position. It is noteworthy that the majority of sequence divergence that has occurred within this family during evolution is present in the N-terminal arm, which in the context of alanine at position 8 may no longer play an important role in sequence-specific binding by this factor.

Predicting the specificity of the human homeodomain set

Based on our complete description of the specificities of the fly homeodomains and our analysis of homeodomain recognition via mutagenesis, we predicted the specificity of the majority of independent homeodomains that are present in the human genome. This effort is facilitated by the large degree of evolutionary conservation within the family (e.g. Figure 20). However, there are numerous instances in our analysis where pairs of homeodomains with the highest overall sequence similarity have different specificities, which are likely due to differences at their key recognition positions (Figure 4.8C). Therefore, three criteria were employed in making predictions for the independent human homeodomains: 1) the presence of Asn51, 2) the overall sequence similarity of each human homeodomain to each fly homeodomain, and for each of these comparisons 3) the number of identical residues at five key recognition positions (5, 47, 50, 54 and 55). Based on these criteria the recognition motif for 153 of 193 human homeodomains (79%) was constructed from the selected binding sites of up to three fly factors that share the highest overall sequence homology and the most similar DNA-recognition residues (Figure 4.21). A cross-validation test was performed on the fly homeodomain set to assess the accuracy of these predictions (Table 4.2). Based on this evaluation, the human predictions were binned into four confidence levels (Appendix Table A.5): 1 represents the highest confidence level and 4 the lowest confidence level. 113 (74%) of the predictions fall in the top two confidence levels and only 8 (5%) of our predictions fall in the lowest confidence level. The quality of these predictions was confirmed by determining the specificity of six human homeodomains representing different specificity groups using the B1H system (Figure 4.22) and by comparing the predicted and defined specificities of the non-fly homeodomains in TRANSFAC (Matys et al., 2003)(Table 4.3). The predicted and defined specificities for the six human homeodomains were highly correlated (all P-values < $2\text{x}10^{-6}$). To make this predictive method accessible to the scientific community, we have created an interactive web-page where a user may enter the sequence of a homeodomain to be evaluated using our predictive criteria. A recognition motif will be constructed for the input homeodomain sequence if fly homeodomains are present in our dataset that meet the user-defined criteria (Figure 4.23). Our specificity predictions for the human homeodomain set, their corresponding PWMs, and the interactive prediction tool are available at http://ural.wustl.edu/flyhd.

Figure 4.21 Sequence logos of the 153 predicted human homeodomains.




HOXB1
HOXC5
HOXD8
IRX6
LHX5
HOXB2
HOXC6
HOXD9
ISL1
LHX8
HOXB3
HOXC8
HOXD10
ISL2
LHX9
HOXB4
HOXC9
HOXD11
ISX
LMX1A
HOXB5
HOXC10
HOXD12
LBX1
LMX1B
HOXB6
HOXC11
IRX1
LBX2
MEIS1
HOXB7
HOXC12
IRX2
LHX1
MEIS2
HOXB8
HOXD1
IRX3
LHX2
MEIS3P2
HOXB9
HOXD3
IRX4
LHX3
MEOX1
HOXC4
HOXD4
IRX5
LHX4
MIXL1
MNX1
bits
2.0
1.5
1.0
0.5
0.0

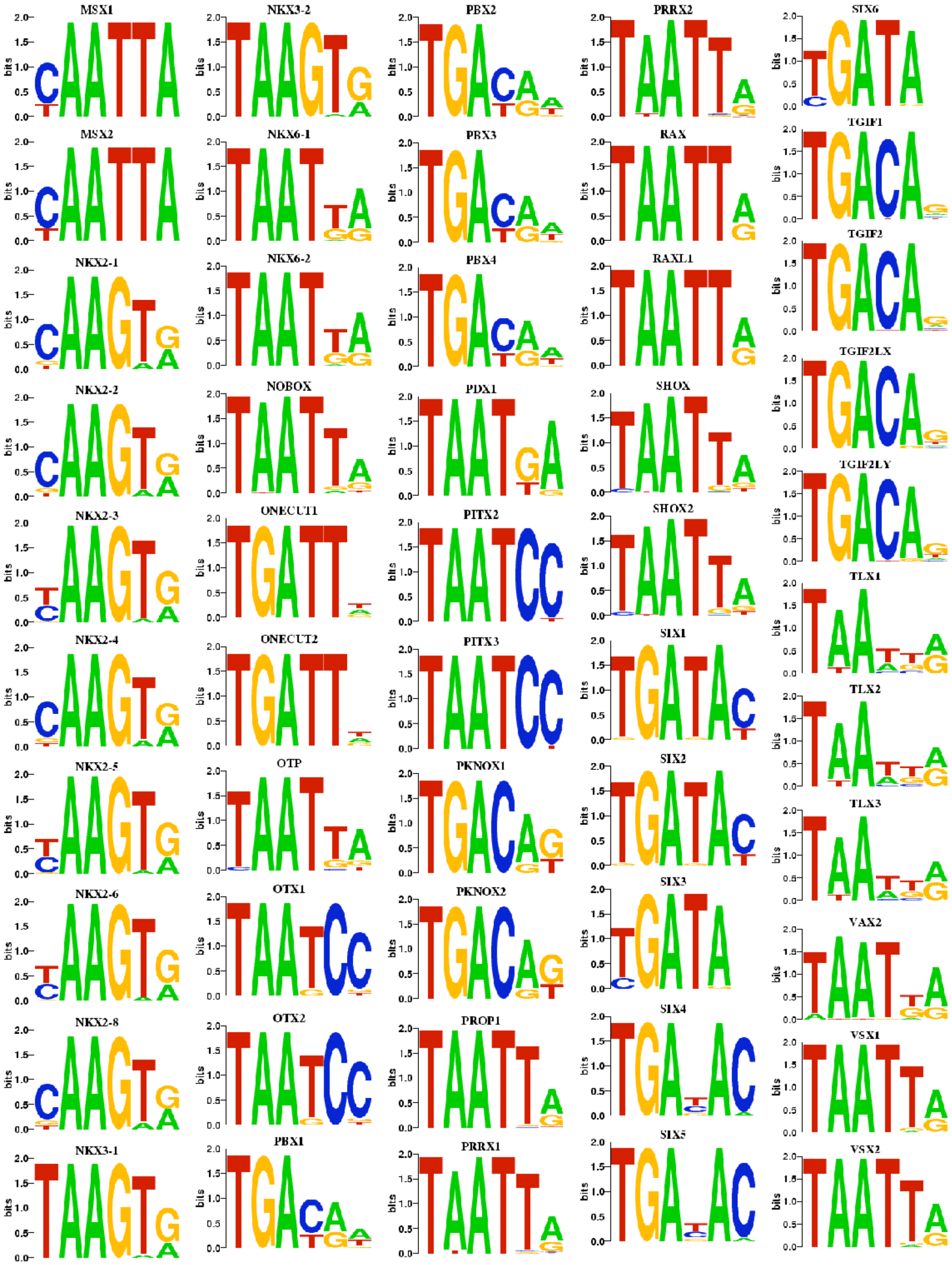

MSX1
MSX2
NKX2-1
NKX2-2
NKX2-3
NKX2-4
NKX2-5
NKX2-6
NKX2-8
NKX3-1
NKX3-2
NKX6-1
NKX6-2
NOBOX
ONECUT1
ONECUT2
OTP
OTX1
OTX2
PBX1
PBX2
PBX3
PBX4
PDX1
PITX2
PITX3
PKNOX1
PKNOX2
PROP1
PRRX1
PRRX2
RAX
RAXL1
SHOX
SHOX2
SIX1
SIX2
SIX3
SIX4
SIX5
SIX6
TGIF1
TGIF2
TGIF2LX
TGIF2LY
TLX1
TLX2
TLX3
VAX2
VSX1
VSX2
bits
0.0
0.5
1.0
1.5
2.0

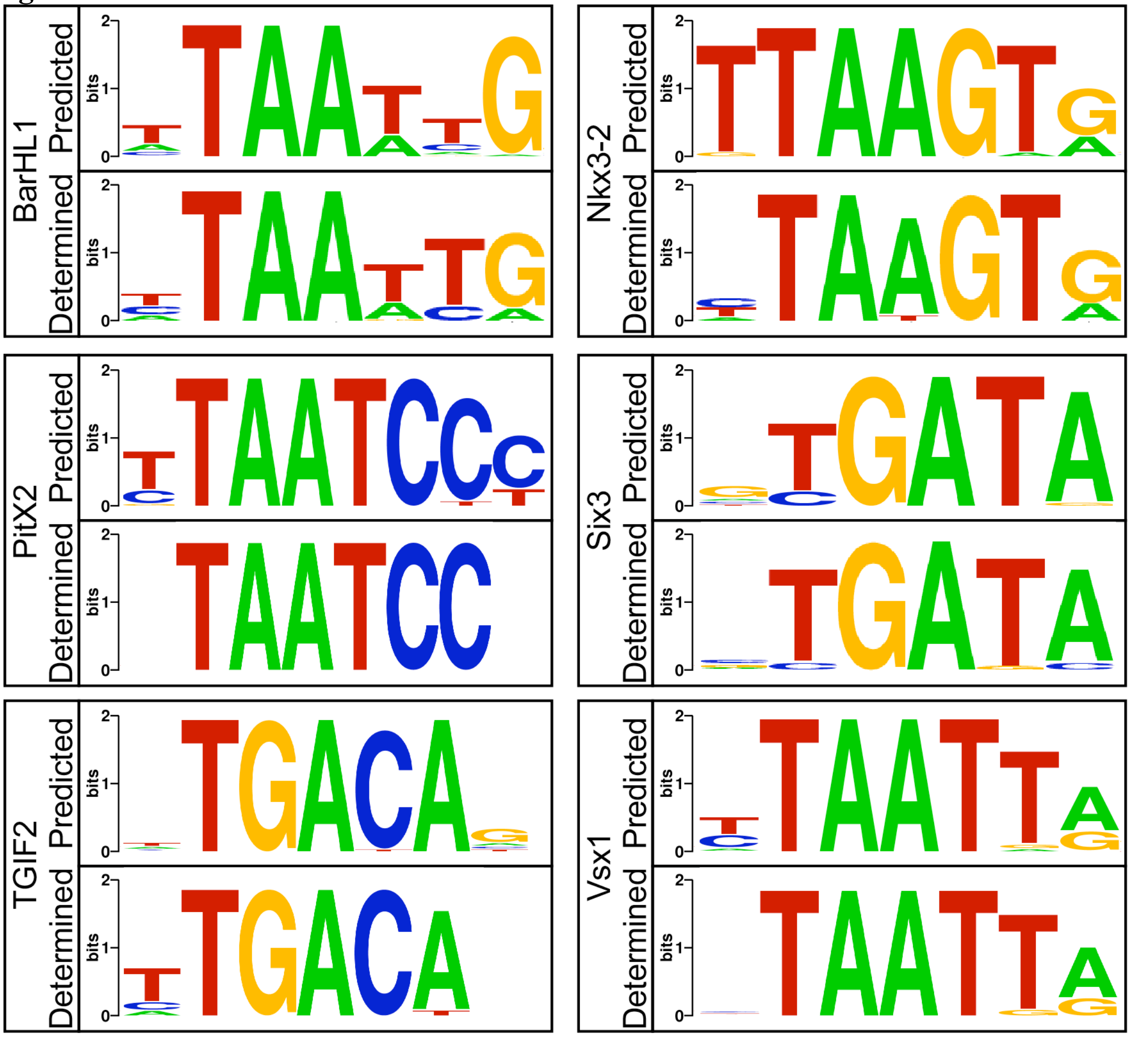
Table 4.2

Crossvalidation analysis of fly homeodomains.

Query	ALLR	Distance	p-value	e-value	Pred. Cons.	Actual Cons.	Combined Cons.
Repo	9.9809	0.0512	7.90E-009	7.90E-009	TAATTa	TAATTA	TAATTa
Rx	9.9038	0.1058	9.29E-009	9.29E-009	TAATTa	TAATTR	TAATTa
Hbn	9.8023	-0.2713	1.15E-008	1.15E-008	TAATTa	TAATTR	TAATTa
Al	9.7524	0.4103	1.28E-008	1.28E-008	TAATTR	TAATTA	TAATTa
Ptx1	9.6421	0.6916	1.61E-008	1.61E-008	TAATCC	TAATCC	TAATCC
Oc	9.6284	0.3647	1.65E-008	1.65E-008	TAATCC	TAATCC	TAATCC
Pph13	9.5617	0.1223	1.90E-008	1.90E-008	TAATTA	TAATTa	TAATTA
Gsc	9.5553	0.6155	1.93E-008	1.93E-008	TAATCC	TAATCc	TAATCC
Dfd	9.4429	0.2754	2.44E-008	2.44E-008	TAATgA	TAATGA	TAATGA
CG32532	9.4385	0.074	2.46E-008	2.46E-008	TAATTa	TAATTR	TAATTa
CG11294	9.3819	0.3812	2.77E-008	2.77E-008	TAATTA	TAATTA	TAATTA
Odsh	9.3406	0.2868	3.02E-008	3.02E-008	TAATTR	TAATTR	TAATTR
Scr	9.2996	0.3014	3.29E-008	3.29E-008	TAATKA	TAATGA	TAATKA
Ind	9.2047	0.2277	4.01E-008	4.01E-008	TAATKA	TAATKA	TAATKA
Zen2	9.1465	0.6703	4.53E-008	4.53E-008	TAATGA	TAATKA	TAATKA
Lab	9.1042	0.7574	4.95E-008	4.95E-008	TAATGA	TAATKA	TAATgA
Inv	9.0763	0.3488	5.25E-008	5.25E-008	TAATTR	TAATTa	TAATTR
En	9.0763	0.3488	5.25E-008	5.25E-008	TAATTa	TAATTR	TAATTR
Unc4	9.0702	0.8337	5.32E-008	5.32E-008	TAATTa	TAATTg	TAATTR
CG9876	9.067	0.3904	5.35E-008	5.35E-008	TAATTa	TAATTa	TAATTa
Otp	9.0266	0.456	5.83E-008	5.83E-008	TAATTA	TAATTA	TAATTA
Ftz	8.9761	0.0222	6.48E-008	6.48E-008	TAATKA	TAATKA	TAATKA
Zen	8.9689	0.699	6.57E-008	6.57E-008	TAATKA	TAATgA	TAATKA
Antp	8.8538	0.4076	8.37E-008	8.37E-008	TAATKA	TAATKA	TAATKA
PhdP	8.8293	0.1977	8.81E-008	8.81E-008	TAATTA	TAATTn	TAATTa
AbdA	8.7462	0.1147	1.05E-007	1.05E-007	TAATKA	TAATtA	TAATKA
Dr	8.6782	2.0874	1.21E-007	1.21E-007	TAATTA	TAATTG	TAATTR
Btn	8.6048	0.3618	1.41E-007	1.41E-007	TAATKA	TAATgA	TAATKA
Vis	8.5718	1.0373	1.51E-007	1.51E-007	TGACAg	TGACAn	TGACA
Achi	8.5718	1.0373	1.51E-007	1.51E-007	TGACAn	TGACAg	TGACA
Unpg	8.5501	1.4249	1.58E-007	1.58E-007	TAATTA	TAATTa	TAATTA
Ro	8.39	1.8609	2.21E-007	2.21E-007	YAATTA	TAATTA	tAATTA
Eve	8.3891	0.4253	2.21E-007	2.21E-007	TAATKA	TAATKA	TAATKA
CG33980	8.3842	0.9261	2.24E-007	2.24E-007	TAATTa	TAATTA	TAATTA
CG4136	8.3842	0.9261	2.24E-007	2.24E-007	TAATTA	TAATTa	TAATTA
Ubx	8.3467	0.5666	2.42E-007	2.42E-007	TAATKA	TAATta	TAATKA
Hth	8.3339	1.5492	2.48E-007	2.48E-007	TGACAg	TGACAg	TGACA
CG18599	8.325	1.2562	2.53E-007	2.53E-007	TAATKA	TAATtA	TAATKA
Pb	8.2429	1.6834	3.01E-007	3.01E-007	TAATKA	TAATKA	TAATKA
Awh	8.2022	1.0905	3.27E-007	3.27E-007	TAATtA	TAATTA	TAATTA
Ap	8.2022	1.0905	3.27E-007	3.27E-007	TAATTA	TAATtA	TAATTA
So	8.1879	2.4025	3.37E-007	3.37E-007	TGATAC	TGATAC	TGATAC
Tin	8.0367	1.177	4.63E-007	4.63E-007	YAAGTR	cAAGTG	YAAGTg

Bap	7.8641	2.6081	6.64E-007	6.64E-007	CAAGTg	TAAGTg	YAAGTg
Dll	7.8095	1.8343	7.45E-007	7.45E-007	YAATTA	YAATTA	YAATTA
BH2	7.7818	1.3614	7.89E-007	7.89E-007	TAAWtG	TAATtG	TAAttG
BH1	7.7818	1.3614	7.89E-007	7.89E-007	TAATtG	TAAWtG	TAAttG
Optix	7.7705	2.0452	8.08E-007	8.08E-007	TGATAC	TGATAn	TGATAC
CG34031	7.7609	1.7932	8.24E-007	8.24E-007	TAATtG	TAATtG	TAATtG
CG13424	7.7235	0.9239	8.92E-007	8.92E-007	TAATtR	TAATtR	TAATtR
Vnd	7.6846	1.9895	9.67E-007	9.67E-007	tAAGTG	CAAGTR	YAAGTg
Slou	7.4995	1.1247	1.43E-006	1.43E-006	TAATtR	TAATtR	TAATtR
Six4	7.4962	3.3833	1.44E-006	1.44E-006	TGATAC	TGAnAC	TGATAC
Ems	7.4919	0.9041	1.45E-006	1.45E-006	TAATKA	TAATKa	TAATKA
E5	7.4919	0.9041	1.45E-006	1.45E-006	TAATKa	TAATKA	TAATKA
CG11085	7.4246	1.0618	1.67E-006	1.67E-006	TAAttG	TAATtG	TAAttG
Hgtx	7.4223	1.6526	1.68E-006	1.68E-006	TAATtR	TAATtA	TAATtR
NK7	7.376	1.4885	1.85E-006	1.85E-006	TAATtR	TAATtR	TAATtR
Bsh	7.3414	1.302	1.98E-006	1.98E-006	TAATtR	TAATKR	TAATtR
Exex	7.2781	4.8019	2.27E-006	2.27E-006	TAATKA	TAATTA	TAAT
Lim1	7.1013	3.0796	3.28E-006	3.28E-006	TAATtA	TAATTA	TAATTA
Mirr	6.8517	0.3406	5.53E-006	5.53E-006	taACAn	WaACAn	WaACA
Ara	6.7625	0.493	6.67E-006	6.67E-006	WAACAn	WaACAn	WaACA
Caup	6.6398	0.3492	8.62E-006	8.62E-006	WaACAn	tAACAn	WaACA
Lbe	6.5811	1.563	9.75E-006	9.75E-006	TAATtA	TAAYnA	TAAtnA
Lbl	6.5811	1.563	9.75E-006	9.75E-006	TAAYnA	TAATtA	TAAtnA
Lim3	6.5461	2.1844	1.05E-005	1.05E-005	TWATTR	TAATtA	TaATTa
C15	6.4381	2.6673	1.32E-005	1.32E-005	TAATtG	TAAttR	TAAttG
CG32105	6.4272	2.9105	1.35E-005	1.35E-005	TWATTR	TAATTA	TWATTR
CG4328	6.4272	2.9105	1.35E-005	1.35E-005	TAATTA	TWATTR	TWATTR
Hmx	6.1889	4.7795	2.22E-005	2.22E-005	TAAKTR	TAATTG	TAAtTg
H2	5.7098	1.9904	6.04E-005	6.04E-005	TWATKA	TWATnA	TWATnA
AbdB	5.647	4.8563	6.89E-005	6.89E-005	TAATKA	TTATga	TAATKA
CG12361	5.2354	5.7449	0.000163	0.000163	TAATTA	TMATWA	TAATtA
Cad	5.0652	6.9451	0.0002328	0.0002328	TAATgA	TtATtR	TAATKA

Figure 4.22



Similarity of predicted and determined recognition motifs for six human homeodomains

Factor Comparison	ALLR	Distance	E value	P value	predicted consensus	actual consensus	combined consensus
BarHL1	7.8848	1.5896	8.48E-007	8.48E-007	TAAttG	nTAAtTGn	TAAtTG
Nkx3-2	9.1699	0.6191	5.04E-008	5.04E-008	TAAGTg	YTAAGTG	TAAGTG
PitX2	10.1543	0.114	5.50E-009	5.50E-009	TAATCC	TAATCC	TAATCC
Six3	7.6043	0.962	1.27E-006	1.27E-006	TGATA	nnSTGATA	TGATA
TGIF2	8.5538	1.3843	1.83E-007	1.83E-007	TGACAg	ntTGACA	TGACA
Vsx1	9.5817	0.2637	2.13E-008	2.13E-008	TAATTa	tTAATTA	TAATTa

Figure 4.22. Comparison of the predicted and actual recognition motifs for 6 human homeodomains. The specificities of the human factors were predicted as described in the text. Specificities for these factors were determined using the B1H system and the ZF10 library. Comparisons between the predicted and determined motifs were determined as described in the Supplementary Methods with the local alignments and similarity calculated by ALLR and the E value and P value calculated by Matalign. Of particular note, the specificity of Six3 is consistent with other Six family members; it does not specify TAAT as previously described {Zhu, 2002}.

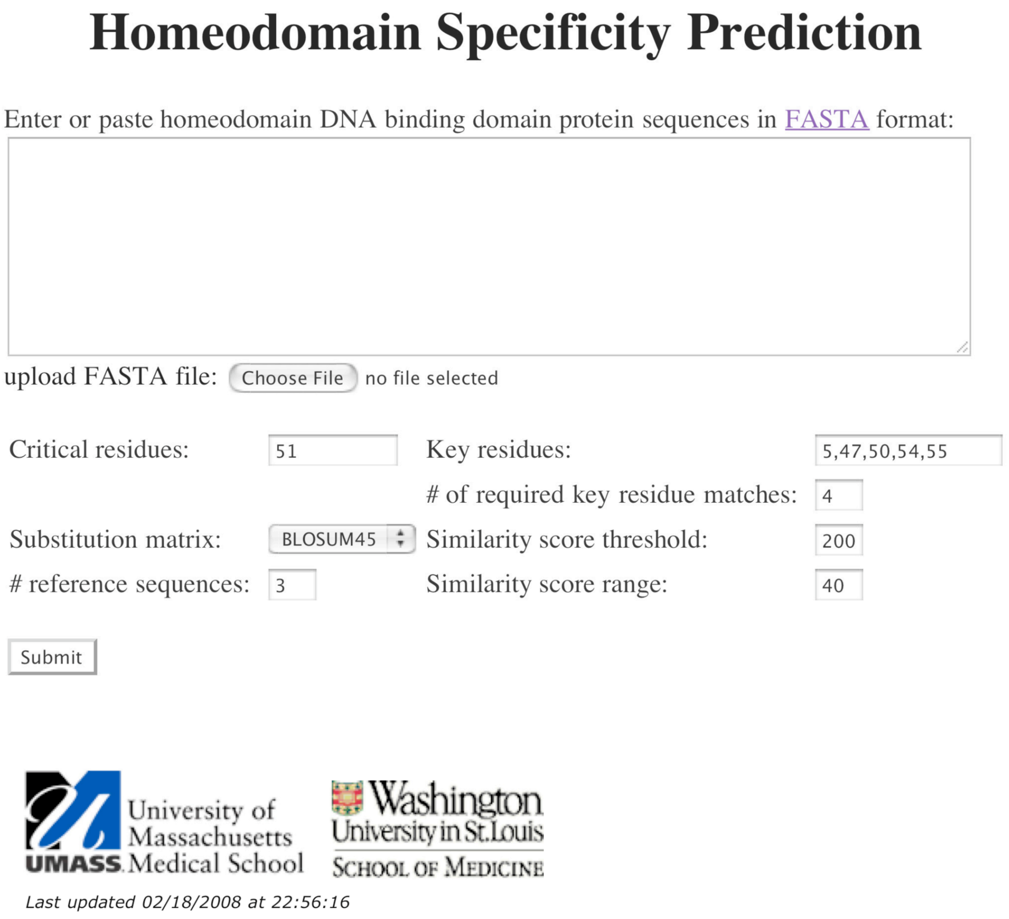
Table 4.3 Prediction of TRANSFAC homeodomains.

Predicted	Actual	ALLR	Dist	E value	P value	Predicted Cons.	Actual Cons.	Combined Cons.
T02970 mouse Chx10	M00437.cons.L8.c2.topMx1	9.2938	4.1572	4.44E-008	4.44E-008	TAATTa	GCTAATTA	TAATTA
T04139 human Chx10	M00437.cons.L8.c2.topMx1	9.2938	4.1572	4.44E-008	4.44E-008	TAATTa	GCTAATTA	TAATTA
T04142 chick Chx10	M00437.cons.L8.c2.topMx1	9.2938	4.1572	4.44E-008	4.44E-008	TAATTa	GCTAATTA	TAATTA
T08863 mouse S8	M00099.cons.L9.c2.topMx1	8.8487	1.2748	1.27E-007	1.27E-007	TAATTa	TAATTRRnt	TAATTR
T03978 human Cart-1	M00416.cons.L14.c2.topMx1	7.836	7.8714	1.64E-006	1.64E-006	TAATTA	SnTAATtRnATTAn	TAATTA
T03981 clawed frog Cart-1	M00416.cons.L14.c2.topMx1	7.836	7.8714	1.64E-006	1.64E-006	TAATTA	SnTAATtRnATTAn	TAATTA
T03980 rat Cart-1	M00416.cons.L14.c2.topMx1	7.836	7.8714	1.64E-006	1.64E-006	TAATTA	SnTAATtRnATTAn	TAATTA
T08295 mouse Nkx2-2	M00485.cons.L9.c2.topMx1	7.2991	5.2662	3.25E-006	3.25E-006	CAAGTR	TAAGTRnTT	YAAGTR
T04337 human Nkx2-2	M00485.cons.L9.c2.topMx1	7.2991	5.2662	3.25E-006	3.25E-006	CAAGTR	TAAGTRnTT	YAAGTR
T04272 chick Nkx2-2	M00485.cons.L9.c2.topMx1	7.2991	5.2662	3.25E-006	3.25E-006	CAAGTR	TAAGTRnTT	YAAGTR
T04265 golden Syrian hamster Nkx2-2	M00485.cons.L9.c2.topMx1	7.2991	5.2662	3.25E-006	3.25E-006	CAAGTR	TAAGTRnTT	YAAGTR
T03489 cattle Crx	M00623.cons.L15.c2.topMx1	6.4759	5.537	3.04E-005	3.04E-005	TAATCC	YnnnTAAtCnnMnnn	TAATC
T03458 rat Crx	M00623.cons.L15.c2.topMx1	6.4759	5.537	3.04E-005	3.04E-005	TAATCC	YnnnTAAtCnnMnnn	TAATC
T03461 mouse Crx	M00623.cons.L15.c2.topMx1	6.4759	5.537	3.04E-005	3.04E-005	TAATCC	YnnnTAAtCnnMnnn	TAATC
T02792 human Crx	M00623.cons.L15.c2.topMx1	6.4759	5.537	3.04E-005	3.04E-005	TAATCC	YnnnTAAtCnnMnnn	TAATC
T00857 human Nkx2-1	M00432.cons.L8.c2.topMx1	6.264	4.6036	2.52E-005	2.52E-005	CAAGTR	ASTCAAGT	CAAGT
T00856 rat Nkx2-1	M00432.cons.L8.c2.topMx1	6.264	4.6036	2.52E-005	2.52E-005	CAAGTR	ASTCAAGT	CAAGT
T02098 dog Nkx2-1	M00432.cons.L8.c2.topMx1	6.264	4.6036	2.52E-005	2.52E-005	CAAGTR	ASTCAAGT	CAAGT
T00859 mouse Nkx2-1	M00432.cons.L8.c2.topMx1	6.264	4.6036	2.52E-005	2.52E-005	CAAGTR	ASTCAAGT	CAAGT
T00856 rat Nkx2-	M00794.cons.L9.c2.to	5.911	5.2132	5.95E-005	5.94E-005	CAAGTR	cTcAAGnGY	cAAGtg

1	pMx1							
T04271 chick Nkx2-1	M00794.cons.L9.c2.topMx1	5.911	5.2132	5.95E-005	5.94E-005	CAAGTR	cTcAAGnGY	cAAGtg
T00859 mouse Nkx2-1	M00794.cons.L9.c2.topMx1	5.911	5.2132	5.95E-005	5.94E-005	CAAGTR	cTcAAGnGY	cAAGtg
T00857 human Nkx2-1	M00794.cons.L9.c2.topMx1	5.911	5.2132	5.95E-005	5.94E-005	CAAGTR	cTcAAGnGY	cAAGtg
T02098 dog Nkx2-1	M00794.cons.L9.c2.topMx1	5.911	5.2132	5.95E-005	5.94E-005	CAAGTR	cTcAAGnGY	cAAGtg
T05157 human RX	M00623.cons.L15.c2.topMx1	5.8299	7.2643	0.0001174	0.0001174	TAATTR	YnnnTAAtCnnMnnn	TAAT
T00863 fruit fly Ubx	M00018.cons.L4.c2.topMx1	5.7005	3.4202	4.11E-005	4.11E-005	TAATKA	TAAT	TAAT
T03848 red flour beetle Ubx	M00018.cons.L4.c2.topMx1	5.7005	3.4202	4.11E-005	4.11E-005	TAATKA	TAAT	TAAT
T01481 human Pbx1a	M00124.cons.L9.c2.topMx1	5.1275	11.4862	0.0003066	0.0003065	TGAcaa	TTGATTGAT	TGA
T01481 human Pbx1a	M00096.cons.L9.c2.topMx1	5.0303	11.8509	0.0003757	0.0003757	TGAcaa	AAGCTTGAT	TGA
T01481 human Pbx1a	M00998.cons.L8.c2.topMx1	4.0228	8.6473	0.002752	0.002749	TGAcaa	tGATTGAT	TGA
T01992 fruit fly Abd-A	M01083.cons.L10.c2.topMx1	2.4675	12.5216	0.08927	0.0854	TAATKA	AARTaAWWWW	TAAT
T04367 human NCX	M00484.cons.L10.c2.topMx1	1.9307	7.2197	0.2746	0.2401	TAAttR	nngtAAntng	TAA
T04368 mouse Ncx	M00484.cons.L10.c2.topMx1	1.9307	7.2197	0.2746	0.2401	TAAttR	nngtAAntng	TAA

Figure 4.23

A.



Homeodomain Specificity Prediction

Enter or paste homeodomain DNA binding domain protein sequences in FASTA format:

upload FASTA file: Choose File no file selected

Critical residues:	51	Key residues:	5,47,50,54,55
		# of required key residue matches:	4
Substitution matrix:	BLOSUM45	Similarity score threshold:	200
# reference sequences:	3	Similarity score range:	40

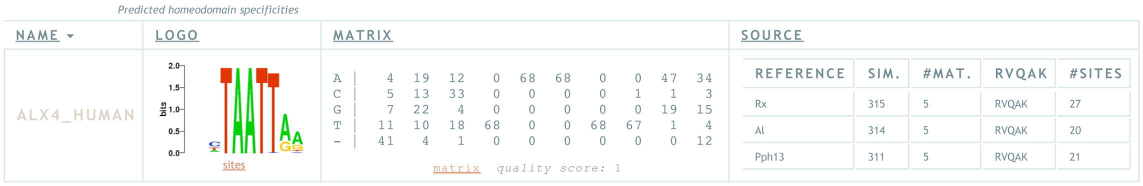
Submit

University of Massachusetts UMASS Medical School

Washington University in St.Louis School of Medicine

Last updated 02/18/2008 at 22:56:16

B.



Predicted homeodomain specificities

NAME ▾ | LOGO | MATRIX | SOURCE

ALX4_HUMAN

```
A |   4  19  12   0  68  68   0   0  47  34
C |   5  13  33   0   0   0   0   1   1   3
G |   7  22   4   0   0   0   0   0  19  15
T |  11  10  18  68   0   0  68  67   1   4
- |  41   4   1   0   0   0   0   0   0  12
```

matrix quality score: 1

REFERENCE	SIM.	#MAT.	RVQAK	#SITES
Rx	315	5	RVQAK	27
Al	314	5	RVQAK	20
Pph13	311	5	RVQAK	21

Figure 4.23. A. Splash page of interactive web-based homeodomain specificity prediction tool at http://ural.wustl.edu/flyhd/. Given a set of homeodomain protein sequences entered as a FASTA file, the program returns a set of predicted sites, a count matrix and a sequence logo for each query protein, as well as the set of reference proteins used to make each prediction. If the homology constraints are not met by any of the reference proteins, no prediction is made. Users can change the default parameters to define alternate sets of key and critical residues and different homology score constraints. **B.** Example of the information returned from the website upon entering the sequence of the Alx4 homeodomain. The sites used to construct the LOGO can be accessed by clicking the “sites” link. The quality score (1 to 4) under the MATRIX indicates the confidence level of the prediction where 1 is the highest confidence. Under SOURCE, the REFERENCE indicates the fly homeodomains used to construct the predicted LOGO where SIM.= similarity score, #MAT= number of matches at the key recognition positions, RVQAK are the amino acids at the key recognition positions in the query, with the corresponding residues of the utilized fly factors listed below, #SITES = number of unique sequences used from each fly factor.

Discussion

The DNA-binding specificity of 84 fly homeodomains, 6 human homeodomains and 16 proteins with mutations in specificity determinants were characterized using the omega-based B1H system. This dataset dramatically increases the number of characterized fly homeodomains; for example, only 18 of the homeodomains in this study have any binding site information present in the FlyREG database (Bergman et al., 2005) and the specificity of two of these factors is described by only a single binding site. B1H binding site data is now available for all the independent homeodomains in *D. melanogaster*. In addition to the ability to efficiently analyze a complete set of DNA-binding domains, the B1H system offers two potential advantages for the analysis of transcription factor specificity. First, selected binding sites are assayed for the ability to activate a biological response in the context of competition from a pool of potential sites in the *E. coli* genome. For some factors, this assay may provide a more relevant measure of specificity than the off-rate measurements obtained in *in vitro* assays using oligonucleotides (Berger et al., 2006). More importantly, the ability to determine the orientation of the homeodomain on each selected binding site allows even partially symmetric sites to be properly aligned when constructing recognition motifs (Figure 4.7). Correct alignment of selected sites is not only important for ranking predicted recognition sequences in genomic DNA sequences, but it is also required to construct an accurate catalog of specificity determinants.

In principal, technical factors such as the use of an N-terminal fusion partner or the number of selected sequences analyzed could limit the accuracy or precision of our dataset. However, several observations argue against these concerns. First, the ability to cluster factors based on their observed binding specificities and to identify correlations in DNA contact residues demonstrates a striking internal consistency within the dataset. The significance of these correlations is confirmed by the ability to subsequently use this data to rationally alter DNA-binding specificity in mutagenesis experiments. Second, validation of the specificities by gel shift for a subset of factors as well as comparisons with external data sources - the previously determined *in vitro* and *in vivo* specificity data for fly homeodomains - provides clear evidence of the accuracy of our data. Combined with the advantages discussed above, this validation of the homeodomain dataset confirms the utility of the B1H system.

This study provides the first global analysis of homeodomain specificities in an organism. We find that the homeodomain family displays a great range of specificities in which a wide variety of bases can be preferred at most positions within the core 6 bp binding site. Overall, the majority of homeodomains (93%) in our dataset can be clustered into 11 different specificity groups. In addition to the 11 specificity clusters, there are 6 individual homeodomains with unique specificities. In general, these orphan factors display different combinations of amino acids at the key DNA-recognition positions than are found in the specificity groups. Thus, while the 43 homeodomains in the Antp and En groups all have related binding specificities, we find evidence for at least

17 different-DNA binding specificities within the entire *D. melanogaster* homeodomain set.

An important conclusion from our analysis is that the overall sequence similarity between two homeodomains is a useful, but sometimes misleading indicator of the degree of similarity in their DNA-binding specificities. Once factors are clustered into specificity groups, it is possible to compare binding specificity with their degree of sequence homology (Figure 4.8C). A substantial correlation between sequence similarity and preferred recognition motif is observed, which is expected since proteins with greater overall homology are also more likely to share similar residues at their DNA-recognition positions. However, we observed multiple examples where pairs of closely related homeodomains cluster into different specificity groups. In both naturally-occurring and engineered homeodomains, single amino acid changes at putative DNA recognition positions are sufficient to significantly alter specificity. Alteration of only 5 recognition positions in Engrailed is sufficient to change its specificity to that of a TGIF family member, although these homeodomains only share 17 identical residues. These observations illustrate the importance of defining the amino acid positions that contribute to variations in binding site specificity in order to make accurate specificity predictions.

Among the specificity groups, there are clear correlations between amino acid sequence and specificity. Using this data we have produced a catalog of common specificity determinants based on our computational and mutagenic analysis of specificity combined with previous biochemical and structural data. The origin of specificity is clearer for residues that are found in the recognition helix. Binding site positions 4, 5 and 6 are specified primarily by the combination of residues at positions 47, 50 and 54. However, the influence of each of these residues does not map to a single position within the binding site and multiple residues can simultaneously affect specificity at a single binding site position. The mutational analysis of Bcd highlights this complexity, as residues at all three of these positions compete to define the nucleotide preference at position 4. Consequently, our catalog of determinants indicates that the presence of specific residues can bias specificity towards a particular base or bases within the binding site, but that the actual specificity is dependent on a combination of primary and secondary effects at the protein-DNA interface.

In addition to defining specificity determinants, this dataset provides an important resource for the prediction and interpretation of homeodomain binding sites in regulatory targets within the *D. melanogaster* genome. The reductionist approach of analyzing the specificity of isolated homeodomains has resulted in a much more detailed understanding of monomeric DNA-recognition. The specificity of individual homeodomains has proven instrumental in the identification of functional regulatory sites utilized by these factors *in vivo* (a subset of examples in *D. melanogaster* are listed in Table 4.4) and in the computational identification of target genes with evolutionarily conserved binding sites (Berman et al., 2004; Kheradpour et al., 2007; Schroeder et al., 2004; Sinha et al., 2003). Comparisons with chromatin immunoprecipitation (ChIP) data confirm that Bicoid

Table 4.4 Examples of functional (*in vivo*) homeodomain binding sites consistent with their monomeric specificity

HD	**Target gene**	**References**	**Notes**	**enhancer element name***
Cad	*ftz*	Dearolf, C.R., Topol, J., and Parker, C.S. (1989). *Nature* 341, 340-343.	Activation of the ftz zebra stripe element is facilitated by cad binding sites where direct mutation of these sites abrogates activity.	ftz zebra stripe
Tin	*mef2*	Gajewski, K., Kim, Y., Lee, Y.M., Olson, E.N., and Schulz, R.A. (1997). *EMBO J* 16, 515-522.	A pair of tinman sites is required for enhancer function in myocyte precursors as demonstrated by loss of *lacZ* expression when the sites are mutated in report assay.	Mef2_IIA237
Tin	*beta3Tub60D*	Kremser, T., Gajewski, K., Schulz, R.A., and Renkawitz-Pohl, R. (1999). *Dev biol* 216, 327-339	Identified 3 tinman sites, 2 are required for reporter transcription in dorsal vessel cells. Mutation of tin sites in enhancer disrupts expression.	betaTub60D_ b3-lac333
Bap	*beta3Tub60D*	Zaffran, S., and Frasch, M. (2002). *Mech. Dev.* 114, 85-93.	Identifies single pair of overlapping bap sites that are responsible for tissue specific gene expression	betaTub60D_ beta3-14/vm1
Tin	*Sur*	Akasaka, T., Klinedinst, S., Ocorr, K., Bustamante, E.L., Kim, S.K., and Bodmer, R. (2006). *Proc Natl Acad Sci U S A* 103, 11999-12004. & Hendren, J.D., Shah, A.P., Arguelles, A.M., and Cripps, R.M. (2007). *Mech. Dev.* 124, 416-426.	In this gene a tin-responsive enhancer was discovered using bioinformatics approaches looking for tin binding sites based on the consensus sequence.	En3
Tin	*svp*	Ryan, K.M., Hendren, J.D., Helander, L.A., and Cripps, R.M. (2007). *Dev biol* 302, 694-702.	In this gene a tin-responsive enhancer was discovered using bioinformatics approaches looking for tin binding sites based on the consensus sequence where the identified sites were conserved over multiple genomes.	SCE
Tin	*eve*	Knirr, S., and Frasch, M. (2001). *Dev biol* 238, 13-26.	Mutation of 2 tin sites identified based on its consensus recognition element inactivates enhancer	EME b3'
Bcd & Cad	*kni*	Rivera-Pomar, R., Lu, X., Perrimon, N., Taubert, H., and Jackle, H. (1995). *Nature* 376, 253-256.	Demonstrates that sets of binding sites for cad and bcd are sufficient for anterior patterned expression in absence of other factor binding sites. Composition of sites not critical, just number and quality.	kni_64
Ap	*Ser*	Yan, S.J., Gu, Y., Li, W.X., and Fleming, R.J. (2004). *Development* 131, 285-298.	14 Ap sites identified by DNaseI footprinting when mutated abrogate activitiy.	Ser_minimal_ wing_enhancer
Bcd	*eve*	Arnosti, D.N., Barolo, S., Levine, M., and Small, S. (1996). *Development* 122, 205-214.	Mutating individual bcd sites reduces activity. Adding novel sites to new positions restores activity demonstrating that the bcd site position is not critical.	eve_stripe2
Ubx	*sal*	Galant, R., Walsh, C.M., and Carroll, S.B. (2002). *Development* 129, 3115-3126.	Ubx represses *sal* expression in the haltere. The development of the haltere is not dependent on Exd or Hth, so the interaction of Ubx with the sal 328 element is thought to be independent of these TFs (although potentially dependent on other unknown TFs). Mutation of individual Ubx sites results in a loss of repression of the reporter gene in the haltere	sal 328
Dfd	*rpr*	Lohmann, I., McGinnis, N., Bodmer, M., and McGinnis, W. (2002). *Cell* 110, 457-466.	Dfd regulates *rpr* in maxillary segment boundary. Loss of 4 Dfd sites severely decreases reporter expression. Gel shift analysis suggests that Exd does not bind cooperatively with Dfd on this element.	rpr_4S3

HD= homeodomain; *Enhancer names were taken directly from the literature or extracted from REDfly monomer binding sites are enriched at sites that are occupied *in vivo* (Li et al., 2008) and that the combination of ChIP data and analysis of conserved transcription factor binding sites generally provides significant improvement in the prediction of functional targets over either method alone (Kheradpour et al., 2007). The complete analysis of *D. melanogaster* specificities also highlights the importance of identifying factors with overlapping specificities to properly interpret conservation and ChIP data. Conserved binding sites might reflect recognition sequences for a number of potential factors with overlapping specificities and factors with overlapping specificities and expression patterns will sometimes compete to occupy binding sites identified in ChIP experiments.

Many homeodomains are known to recognize DNA not only as monomers, but also as homodimers, heterodimers or higher order complexes; in several examples, the preferred recognition sequence of monomers in these complexes may even be modified (Pearson et al., 2005; Ryoo and Mann, 1999; Wilson and Desplan, 1999). Both structural data and our analysis suggest that a likely site for modified specificities is in the flexible N-terminal arm (Figures 4.1, 4.8 and 4.12). The recently described structures of Scr-Exd heterodimers bound to DNA reveal how complex formation with Exd can alter the interaction with DNA of residues within and beyond the N-terminal arm in Scr (Joshi et al., 2007). Thus, while the primary sequence determinants within the N-terminal arm, such as Arg3 and Arg5, are important in defining sequence preferences, the influence of secondary effects, whether intramolecular (e.g. Ala8 in Caup; Figure 4.12) or intermolecular (complex formation with other DNA-binding domains; Scr-Exd), can also influence recognition. It is currently unclear how frequently monomeric specificities are modified by protein-protein interactions, but our systematic characterization of monomeric specificities provides a foundation to begin to fully explore this question.

The analysis of homeodomain specificities in *D. melanogaster* also provides the basis to predict most homeodomains specificities in other organisms. Using a comparative approach that is based both on the overall sequence similarity and on the identity of the key recognition residues, we predicted the DNA-binding specificities of 79% of the independent homeodomains in the human genome with moderate to high confidence (Figure 4.21). The quality of these predictions was validated by characterizing a small number of human homeodomains in the B1H system and, where available, by comparison to previously published data. Our robust prediction scheme can be applied to homeodomains from any species, thereby providing an important community resource to help identify functional binding sites in regulatory regions of target genes. We have developed a web-based tool that will evaluate the specificity of any input homeodomain sequence based on its amino acid sequence using our database and evaluation criteria. Because our understanding of homeodomain recognition is still imperfect, future improvements to our algorithm should lead to more comprehensive specificity predictions based on the composition of residues at the recognition interface, such as the incorporation of a probabilistic recognition code to approximate the specificities of factors that do not have good homologs in our database (Benos et al., 2002; Liu and Stormo, 2005).

Continued B1H analysis of homeodomain specificity should lead to more detailed understanding of recognition by this family. Our current experiments have led to a catalogue of specificity determinants that can be used to rationally engineer the DNA-binding specificity of homeodomains with reasonable success; this is most clearly demonstrated by the conversion of Engrailed into a TGIF-like factor. Intermediate specificity alterations were also obtained in these experiments (Figure 4.17B and Figure 4.24) including homeodomains with the novel specificities TAACA and TGATTA, which are not observed for any of the natural homeodomains characterized in this study. The throughput of the B1H system will facilitate the synthesis of a more comprehensive recognition model as homeodomains from other species with different combinations of recognition residues are characterized and as additional mutagenesis experiments are performed to more thoroughly interrogate specificity. Using the B1H system, it should also be possible to perform selections on pools of mutagenized homeodomains to select proteins with a desired DNA-binding specificity, providing a relatively unbiased assessment of the range of residues that are compatible with recognition of a given motif. Obviously, this type of family-wide specificity analysis can also be applied to other classes of DNA-binding domains to characterize their specificities and decipher their specificity determinants with the ultimate goal of producing a complete map of the specificities of all of the transcription factors in a genome.

Figure 4.24

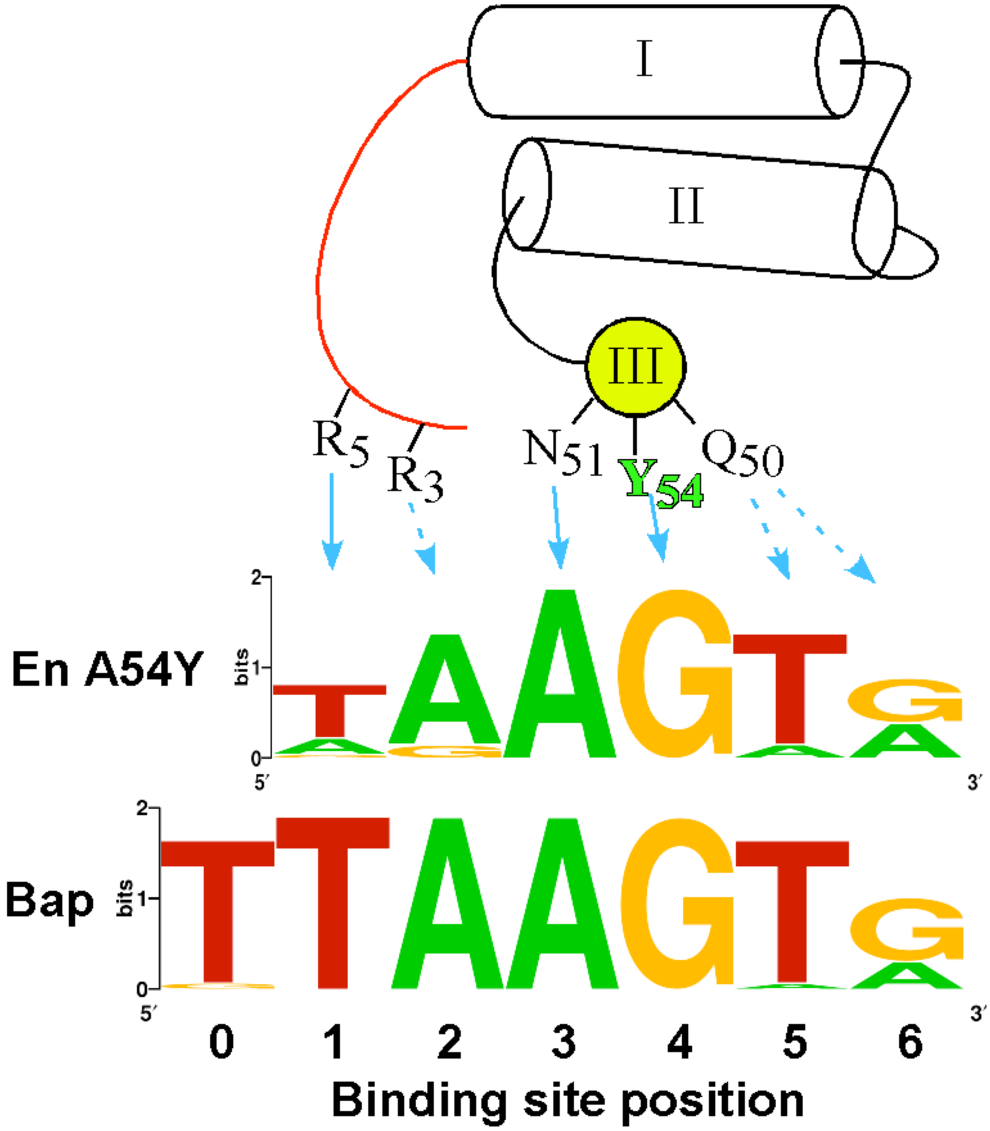

I
II
III
R5
R3
N51
Y54
Q50
En A54Y
Bap
bits
2
1
0
5′
3′
0
1
2
3
4
5
6
Binding site position

Figure 4.24. Conversion of Engrailed into a factor with an NK-2 type specificity. A single mutation was introduced into Engrailed at position 54 (A54Y) to attempt to convert this homeodomain into a specificity that is similar to the NK-2 class. This mutation has been introduced previously into Antp and Gsc with a resulting conversion of sequence preference at position 4 of the binding site from Thy to Gua (Damante et al., 1996; Pellizzari et al., 1997). We observe a similar phenomenon, where this single mutation is sufficient to convert the specificity of Engrailed at position 4 from Thy to Gua. For reference the specificity of Bap, a NK-2 family member, is shown below the specificity of the A54Y mutant.

This selection system also provides a promising platform for the creation of artificial DNA-binding proteins with unique specificities. Previous studies have focused on the utility of zinc fingers as a flexible system to create artificial DNA-binding domains for genetic engineering of cell lines or animals (Beumer et al., 2006; Urnov et al., 2005). Our zinc finger-homeodomain (ZFHD) selection framework, which is derived from ZFHD1 (Pomerantz et al., 1995), provides a system in which the specificity of both components can now be engineered to create hybrid transcription factors that can be used for targeted gene regulation or modification. Our analysis of *D. melanogaster* homeodomains creates a catalog of naturally-occurring components that can be incorporated into chimeric transcription factors and provides a blueprint for engineering homeodomains with novel specificities, expanding the list of modular DNA-binding components that can be used to manipulate natural systems.

Experimental Procedures

Homeodomain binding site selections: A detailed description of the general B1H selection protocol is described in Chapter 1*,* modifications are detailed below. The sequences of the homeodomains used in the B1H selection and the raw selected binding sites are found in Appendix Table A.2.

Identification and boundary definition of the independent homeodomains

These 84 homeodomains represent all of the fly homeodomains in the SMART database that are not associated with another major type of DNA-binding domain: 18 unique homeodomains are associated with PAX, POU or ZnF-C2H2 domains or an additional homeodomain based on the SMART annotation (Letunic et al., 2006; Schultz et al., 1998). There are homeodomains in our set that are associated with another DNA-binding domain, such as those in the Cut family, which were retained because of interesting sequence composition. The sequences of the homeodomains used in the B1H selection and the raw selected binding sites are found in the Appendix, Table A.2. The region of the homeodomain that was fused to omega was defined by the length of the core homeodomain identified by SMART: 60 to 63 amino acids depending on the presence of a TALE insertion. 10 additional amino acids in the protein sequence were included beyond the C-terminus of the homeodomain (if present). The amino acids were removed from the terminus if a hydrophobic residue occurred (YFIVLWPM), since terminal hydrophobic residues can induce protein degredation in E coli (Parsell et al., 1990).

Omega-Zif12 fusions for the selection of homeodomain specificity

Homeodomains were expressed as omega fusions in combination with fingers one and two of Zif268 (Zif12) under control of the *lacUV5mut* promoter plasmid (pB1H2ω2-12HD; Figure 4.2). Each homeodomain (with two additional N-terminal residues) was cloned between the KpnI site and the Xba1 site downstream with a stop codon introduced just prior to the Xba1 site. Two additional amino acids were added after the Kpn1 site prior to the start of the homeodomain. The first amino acid was always glycine. In the majority of the homeodomains, the second amino acid was the -1 amino acid of the specific homeodomain being assayed, however for a subset we used the -1 residue of Oct1, which is arginine, for purely historical reasons. The KpnI site and the inserted residues created a 5 amino acid linker between the 2nd His of Zif268 finger 2 and the beginning of the HD (Zif12-TGTGN-HD; Figure 4.2). All expression constructs were sequence verified.

Alternate selection conditions

The vast majority of the selections were successful in the initial attempt, however a handful did not yield an obvious enrichment in the number of selected clones as defined by a low fold increase (or no increase) in the number of surviving clones on selective media relative to the background when normalized to the number of cells plated. In most cases this was resolved by expressing the omega-Zif12-homeodomain at higher levels using a stronger promoter (*lacUV5*). In the case of one homeodomain, Eve, where our initial selections failed, we found that the removal of a small string of hydrophobic residues from the C-terminus of the protein just after the end of the homeodomain resulted in a significant improvement in activity. Hydrophobic residues at the C-terminus of a protein can lead to lower levels of functional expression in bacteria (Parsell et al., 1990).

Construction of the master alignment of sites:

The master alignment contains 1860 binding sites for 83 of the 84 Drosophila homeodomain proteins as well as Oct1 (Lag1 was excluded because it lacks Asn51, which makes the alignment of its sites to all others within the dataset problematic). CONSENSUS selected substrings from 1868 of the 2211 input sequences and took the reverse complement of 657 of these substrings, wherein alignments were generated for each factor independently(Hertz and Stormo, 1999). In every case, the length of the motif was selected by varying the motif length parameter (-L) and selecting the alignment with the smallest e-value. These 84 separate alignments formed the basis for the construction of the master alignment. The orientation of the alignments for individual factors produced by CONSENSUS was somewhat arbitrary; consequently, we manually reversed the orientation of 35 sets of sequences (about 588 sites). As described previously, the high information content Ade (recognized by Asn51) was used as an anchor to help align the sets of sites (notice that all but 4 sites contain Ade at position 6 in the master alignment below). Information about the probable orientation of each individual site gleaned from the observed site biases (described above) led us to manually 'flip' the orientation of some individual sites (47), overriding some orientation decisions made by CONSENSUS. For the master alignment of all 84 sets of sites we used the entire sequence of each aligned site, not just the 1868 substrings returned by CONSENUS (8 problematic sites were removed; Supplementary Table 3). At this point, the alignment contained 15 columns as the registers of the aligned subsites in each sequence varied, so the 5' and 3' flanking columns 1, 2, 13, 14, and 15 were removed to generate a master alignment with 10 columns because from 57 to 99 percent of these columns were comprised of gaps as the library sequence elements are only 10 bp in length.

Count matrix for the entire master alignment:

```
         1      2      3      4      5      6      7      8
9      10
A |    297    435    269     78   1511   1856     48    326
983    260
C |    310    396    419     66      1      0    219    143
109    254
G |    242    446    161     19    213      2     84    328
512    345
T |    358    542   1006   1697    135      2   1509   1063
60    119
- |    653     41      5      0      0      0      0      0
196    882
```

All Sequence logos (Schneider and Stephens, 1990) for these factors were generated using WebLogo (Crooks et al., 2004). We note that the number of selected binding sites that comprise a particular logo are modest (typically 20 to 40) and consequently, the significance of bases that are absent or occur infrequently in a motif cannot be fully assessed.

Clustering of binding site motifs

The master alignment of sites was used to determine the pairwise global alignments between every set of homeodomain binding sites. The aligned sites for each homeodomain protein were converted to count matrices. Pairwise distances between all matrices were calculated based on average log likelihood ratio (ALLR) similarity scores (Wang and Stormo, 2003). When calculating the ALLR scores, gaps were treated as missing data and ignored. The formula for the ALLR score was modified slightly:

instead of using the natural logarithm function (log base e), log base 2 was used. The Neighbor program from the Phylip phylogenetic analysis package (Felsenstein, 2005) was used to cluster the motifs using the neighbor joining method. The input to Neighbor was a pairwise distance table based on the master alignment of sites. The radial logarithmic neighbor joining tree of the motifs in Figure 2 was produced using the TreeIllustrator program (Trooskens et al., 2005). The branch lengths displayed in this image are logarithmically proportional to the actual branch lengths calculated by the Neighbor program. The phylogram of the homeodomain amino acid sequences in Figure 2 was produced using TreeIllustrator with the pairwise distances determined by ClustalW (Thompson et al., 1994).

MI analysis

MI analysis was performed on the dataset using the Master alignment of binding sites as previously described (Gutell et al., 1992). The MI plot was transformed into a joint rank product matrix by transforming each element in the MI matrix by calculating the rank of each element's MI value in that column (the column-wise rank) and the rank of each element's MI value in that row (the row-wise rank). The column-wise rank and row-wise rank for each element were multiplied to yield the joint rank product matrix. The product matrix was transformed to generate a heat plot using the following formula:

$$\frac{\max(Ln(X)) - Ln(X_{ij})}{\max(Ln(X))}$$

where X_{ij} is the joint rank product matrix element ij and max(X) is the maximum value in X (600).

G-test significance analysis

The significance of an apparent difference between motifs for two groups of homeodomains was estimated using a G-test (Sokal and Rohlf, 1995). Aligned binding sites for each group of factors were pooled and one position (column) in the DNA binding motif was analyzed by generating a 2 by 4 contingency table, where rows contained the 2 classes and columns 4 DNA bases. Small pseudo counts (0.01) were added to each value and the G-test statistic was calculated allowing 3 degrees of freedom for each base, unless a base was not observed in both of the two classes, in which case 1 degree of freedom was subtracted.

Specificity Predictions for the human homeodomain set

193 homeodomains containing proteins were annotated in the SMART human genome database and 175 of these were independent homeodomains containing Asn51. To predict the DNA-binding specificity of this set we used the DNA-binding specificity of up to 3 of the fly homeodomains with the highest BLOSUM45 similarity scores (calculated from a sequence-to-profile multiple sequence alignment (Edgar, 2004) between the query sequence and the 84 fly homeodomain profiles) provided that: 1) they contained Asn51; 2) they contained identical residues at the other 5 key recognition positions (5, 47, 50, 54 and 55); and 3) they passed a BLOSUM45 similarity score

threshold. The similarity score threshold was set to 200, based on a cross validation analysis of the fly homeodomain set (data not shown). Additionally, once a reference protein passed all of our filters, additional reference proteins were only added to the predictive set if their similarity score was within 40 similarity score units of the most similar reference protein. If no reference homeodomain passed these three criteria, we considered up to 3 homeodomains within the set that contained identical residues at 4 of the 5 key recognition positions, as long as they also passed the similarity score threshold. Specificity predictions comprise all of the selected binding sites for all of the reference homeodomains that passed the filters. In some cases no fly homeodomains met these criteria and consequently no prediction was made.

Cross-validation analysis and prediction of the Transfac homeodomains

To assess the accuracy of the specificity predictions we performed a cross-validation analysis where the binding specificity of each fly homeodomain was predicted based on the information of all of the other homeodomain proteins. All TRANSFAC 10.2 datasets associated with proteins classified as homeodomains (TRANSFAC classes C0006, C0027, C0047, C0053) and that contain at least 20 binding sites were extracted from the database (Matys et al., 2006). The 47 groups of binding sites that met these requirements were reanalyzed with CONSENSUS to generate new motifs. 27 of these 47 transcription factors were sufficiently similar to a *D. melanogaster* homeodomain to make a prediction based on our criteria (described in the text). In some cases (8), multiple homeodomains were associated with one dataset in TRANSFAC and vice versa (5). In these cases, we compared the predicted matrix for a factor to each of the CONSENSUS matrices associated with it. We used the ALLR score to determine the best local alignment (Matalign-v2a, Wang, T & Stormo, G. D. *unpublished*) between the predicted and CONSENSUS matrices. Based on these alignments, we assessed the degree of similarity using the ALLR similarity score, the ALLR based distance and the e-value computed by Matalign.

Competition Gel Shift Assay

Oligonucleotides.

The Oligonucleotides used for this assay were designed to have a single, central homeodomain binding site that represents the consensus recognition sequence of one of 7 core specificity groups (Engrailed, Bar, Abd-B, Bicoid, NK-2, Six, and TGIF) as well as one outlier (CG11617). Once annealed, the resulting duplex oligonucleotides contain a 5' GG overhang at each end that can be used to radiolabel the DNA. These sequences of the oligonucleotides are listed below with the recognition sequence in **bold**. Where multiple binding sites were examined for a single specificity group, the differences within these sequences are <u>underlined</u>:

Engrailed Top

GGGCAGGCAG***TAATTA***GGACGTCG

Engrailed Bottom

GGCGACGTCC***TAATTA***CTGCCTGC

Bar Top

GGGCAGGCAG***TAATTG***GGACGTCG

Bar Bottom

GGCGACGTCC***CAATTA***CTGCCTGC

Abd-B-A Top

GGGCAGGCAG***TTATTA***GGACGTCG

Abd-B-A Bottom

GGCGACGTCC***TAATAA***CTGCCTGC

Abd-B-G Top

GGGCAGGCAG***TTATTG***GGACGTCG

Abd-B-G Bottom

GGCGACGTCC***CAATAA***CTGCCTGC

Bicoid Top

GGGCAGGCAG***TAATCC***GGACGTCG

Bicoid Bottom

GGCGACGTCC***GGATTA***CTGCCTGC

NK-2 Top

GGGCAGGCAG***CAAGTG***GGACGTCG

NK-2 Bottom

GGCGACGTCC***CACTTG***CTGCCTGC

CG11617-A Top

GGGCAGGCAG***TTAACA***GGACGTCG

CG11617-A Bottom

GGCGACGTCC***TGTTAA***CTGCCTGC

CG11617-C Top

GGGCAGGCAG***TTCACA***GGACGTCG

CG11617-C Bottom

GGCGACGTCC***TGTGAA***CTGCCTGC

CG11617-T Top

GGGCAGGCAG***TTTACA***GGACGTCG

CG11617-T Bottom

GGCGACGTCC***TGTAAA***CTGCCTGC

Six Top

GGGCAGGCAG***TGATA***CGGACGTCG

Six Bottom

GGCGACGTCCG***TATCA***CTGCCTGC

TGIF Top

GGGCAGGCAG***TTGACA***GGACGTCG

TGIF Bottom

GGCGACGTCC***TGTCAA***CTGCCTGC

Expression and Purification of Proteins.

Each homeodomain was expressed as a C-terminal fusion to maltose binding protein (MBP) from pJH196 (a generous gift from Keith Joung, {Hurt, 2003}) using an *in vitro* transcription-translation system (Expressway^{TM} Cell-Free E. coli Expression System, Invitrogen). The zinc fingers utilized for the binding site selection in the B1H system were not incorporated into these constructs. Each MBP-HD construct was expressed in two 100μl reactions from approximately 2μg of plasmid DNA per reaction. These reactions were incubated while rotating for 6.5 hours at 37^{o}C. The reactions for each construct were combined together with one of the binding buffers listed below (900 μl final volume) and the MBP-HD proteins were captured on 100μl of Amylose Resin (New England BioLabs) by incubation at 4°C for 1.5 hours while rotating. The resin-bound MBP-HD proteins were washed 4 times with 1ml binding buffer. Finally, the protein was eluted from the resin by incubation with 50ml of binding buffer supplemented with 40mM maltose at room temperature, while rotating for 30 minutes. Aliquots of protein were stored at -80^{o}C.

Two different binding buffers were used in these experiments. The majority of the gel shift assays utilized a binding buffer consisting of 10mM Tris-HCl pH 7.5, 0.1mM EDTA, 25mM NaCl, 1mM DTT and 5% glycerol. The more complex binding buffer for Ptx and Caudal consisted of 15mM HEPES pH 7.8, 50mM KCl, 50mM KGlutamate , 50mM KOAc, 5mM $MgCl_2$, 1mM DTT and 5% glycerol. All binding buffers were supplemented with 0.1mg/ml BSA and 0.1% IGEPAL CA-630 for the gel-shift assays.

Gel Shift Competitions.

To perform the competition gel shift assays, oligonucleotides for the consensus binding site corresponding to each homeodomain were annealed and then endlabeled. 30 ml labeling reactions were done using 200ng of annealed oligonucleotide, 40mCi a-^{32}P dCTP, 5 units Klenow (exo-), and a final concentration of 3.33mM dNTPs minus dCTP. These reactions were incubated at 37^{o}C for 30 minutes and then chased with 3.33mM dNTPs (including dCTP) for an additional 30 minutes at 37^{o}C. The labeled oligonucleotides were recovered from free radionucleotides using a G-25 spin column (BioRAD).

The optimal amount of protein for each homeodomain to be used in the DNA-binding reaction for the competition assay was determined by performing shift assays with a titration of protein on its labeled consensus site. Titrations of both protein and DNA were performed to ensure that binding reactions were under K_d conditions with [labeled DNA] $<< K_d$ and that the amount of HD-DNA complex formation was not saturated (data not shown). The appropriate concentration of cold consensus binding site needed to effectively compete the majority of the HD-DNA* complex at the optimal protein concentration was determined by titration of competitor.

The competition assays were then performed by equilibrating the homeodomain with 80pM of its labeled target site and one of the cold competitor duplexes in a 20μl reaction for at least 2 hours at room temperature. Each protein was challenged in a separate reaction with each of the eight of the specificity group binding site oligonucleotides and one control reaction without competitor. The final concentrations of cold competitor DNA used for each homeodomain is as follows:

CG11617 used 5.625nM

En, Vis, and Tin used 9nM

Lbe, Optix, CG34031, Ptx, and Cad used 90nM

10μl of each reaction was then run on a pre-run (35 minutes at 300V) 7.5% native polyacrylamide gel (0.5xTBE) for 35 minutes at 300V. The gels were dried and then exposed phosphoimaging plates for 8-12 hours. These plates were scanned with a FUJIFILM FLA-5000 and the percentage of protein-DNA complex in each reaction was determined by quantifying the free DNA and bound DNA bands with FUJIFILM's program, Image Gauge 4.22.

ACKNOWLEDGEMENTS

We would like to thank Xiangdong Meng for his valuable advice and technical support. We would like to thank the Berkeley Drosophila Genome Project (BDGP) for producing the cDNA clones used in this study, the Drosophila Genomics Resource Center (DGRC) for distributing the clones, and Mark Stapleton and Susan Celniker for sharing unpublished results. Some of these ORFs were obtained from clones produced by BDGP under National Institutes of Health grant (HG002673 to S. E. Celniker). We would like to thank Adam Richards for technical support. S.A.W. and M.B.N. were supported by NIH grant 1R21HG003721 from NHGRI. A.W. was supported in part by NIH grant 1R21HG003721 from NHGRI. M.H.B. and A.W. were supported in part by a New Scholar in Aging Award from the Ellison Medical Foundation and American Cancer Society grant RSG-05-026-01-CCG. R.G.C. was supported by training grant T32 GM08802. G.D.S. was supported by NIH grant HG00249 from NHGRI.

CHAPTER V: GENERAL DISCUSSION

The Bacterial One-Hybrid System

We have developed an omega-based B1H system that allows the high throughput determination of TF DNA-binding specificity. This system has several advantages over other techniques for characterizing DNA-binding specificity. First, the use of *E. coli* as our platform allows the isolation of complementary TF - binding site combination *in vivo* in a single round of selection using relatively simple techniques. Because *E. coli* demonstrate an extremely high transformation efficiency, randomized binding site libraries with complexity greater than 10^8 members can be utilized. Perhaps the greatest advantage realized by this system is the flexibility provided by utilizing omega-TF hybrids, as the absence of competition from endogenous omega has resulted in an extremely sensitive selection system with a much greater dynamic range than previous systems (Durai et al., 2006; Meng et al., 2005). This sensitivity has allowed us to successfully characterize TFs that failed to generate motifs in the alpha-based B1H system.

Using this system we have determined recognition motifs for ~14% of the predicted *D. melanogaster* TFs. For comparison the FlyREG database contains motifs for 53 TFs constructed from 5 or more identified binding sites (Bergman et al., 2005); thus our database doubles the number of specificities that are available, and in cases where these databases overlap, our data is typically of higher quality. The rate of successful TF characterization within this system (101 of 102) makes it amenable to perform comprehensive surveys of TF specificity in complex organisms: once cloned, ten or more factors can be analyzed in parallel in the B1H system in a manner of days. Moreover, the success rate and high throughput nature of this system has allowed us to efficiently analyze a complete set of DNA-binding domains. The B1H system offers two additional advantages for the analysis of transcription factor specificity. Binding sites are assayed for the ability to activate a biological response in the context of competition from a pool of potential sites in the *E. coli* genome. For some factors, this assay may provide more relevant measure of specificity than the off-rate measurements obtained in *in vitro* assays using oligonucleotides (Berger et al., 2006). Perhaps more importantly, the ability to determine the orientation of the homeodomain on its binding site allows even partially symmetric sites to be properly aligned when constructing recognition motifs. If it is possible to determine the orientation of other domains characterized with the B1H system, it may be possible to draw the same types of predictive conclusions about the amino acid sequence-DNA-binding specificity correlations. At the very least, it would provide a much greater understanding of the protein-DNA interaction for several DBD families.

Our current dataset is focused primarily on monomeric DNA-binding domains, but also includes homodimers and heterodimers. This reductionist approach overlooks the potential for sets of factors to cooperatively recognize motifs that are not a simple composite sites formed from their individual motifs, such as the Exd-Hox combinations that play critical roles in specification during development (Pearson et al., 2005; Ryoo and Mann, 1999; Wilson and Desplan, 1999). In addition, Young's analysis of the yeast regulatory networks uncovered one primary group of regulators considered ‘condition altered’. In other words, their promoter preference was altered under different growth conditions (Harbison et al., 2004). The difference in promoter preference is likely due to interactions with different cofactors or dimerization partners available in one condition versus another, which may alter the sequence specificity. Certainly these alterations are lost in a set of monomeric specificities. However, these types of combinations can potentially be characterized using the B1H system, as a set of complementary expression plasmids for the characterization of heterodimers have been developed (Meng et al., 2005; Meng and Wolfe, 2006). Still, a great number of questions remain unanswered for the characterization of a heterodimeric B1H systems such as what promoter strength(s) to use, which monomer is the optimal omega fusion partner, and how to decipher the possibility of recovering monomeric, homodimeric and heterodimeric binding sites. In addition, some criteria for choosing sets of factors to be evaluated must be applied because of the combinatorial issues involved with testing all possible pair-wise combinations.

The Genome Surveyor

The PWMs generated from our B1H data when used in combination with Genome Surveyor provide a fast, flexible and accessible platform for user-guided prediction of CRMs in the fly genome. This type of PWM-guided CRM discovery has been previously accomplished with a set of maternal and gap TFs by several groups (Berman et al., 2002; Berman et al., 2004; Rajewsky et al., 2002; Schroeder et al., 2004) using different computational approaches. Both demonstrated that known CRMs and novel CRMs could be successfully identified within the genome based on the presence of clusters of binding sites for factors that function in a common regulatory pathway. These studies demonstrated that even relatively crude representations of the DNA-binding specificity of a TF, typically constructed from DNaseI footprinting on a limited number of sites (Adryan and Teichmann, 2006), could help identify CRMs and that these predictions could be improved by using two related fly genomes (Berman et al., 2004; Sinha et al., 2004). These computational approaches, as well as an additional method (Sosinsky et al., 2003), differ in the tactics used for CRM identification, but share a common strategy with Genome Surveyor of identifying clusters of overrepresented binding sites.

The key features of Genome Surveyor are that it evaluates the quality of each binding site as well as over-representation of binding sites relative to the genome average, but is still rapid enough to allow genome-wide searches to be performed on a web-server. Thus, Genome Surveyor, which is integrated within the GBrowse software interface, provides a particularly powerful platform for gene-specific or genome-wide searches for CRMs regulated by a combination of factors. Users can rapidly perform genome-wide searches with any combination of 100 factors over the *D. melanogaster* and *D. pseudoobscura* genomes and then investigate the locations of peaks of interest within the genome using the GBrowse tools. Peaks that overlap with previously identified CRMs can be easily identified by uploading annotations for these elements from the REDfly website (redfly.ccr.buffalo.edu)(Gallo et al., 2006). The number and quality of PWMs available for these searches will increase with the adoption of new, high-depth sequencing and barcoding technologies such 454 (Hoffmann et al., 2007; Margulies et al., 2005) and SOLEXA-based sequencing (Barski et al., 2007; Johnson et al., 2007) for the analysis of the B1H-selected binding sites.

Future Directions for CRM Prediction

As the number of factors with high quality PWMs increases, it should be feasible to annotate most potential CRMs using combinations of factors that function in common regulatory networks. Cooperating TFs could be identified based on common expression patterns, phenotypes, or physical interactions. Because Genome Surveyor is built into the GBrowse webtool format (Stein et al., 2002), it will also be possible to incorporate other corroborating datasets into these tools, such as genome wide ChIP analysis of TF binding or chromatin structure. The combination of these experimental and computation approaches for the identification of CRMs should provide the most robust method for the functional annotation of these elements throughout eukaryotic genomes.

A Catalog of Homeodomains Specificities

Using the Omega-based B1H system we characterized the specificity of 84 homeodomains from *D. melanogaster* and 16 additional specificity mutants. This dataset dramatically increases the number of characterized fly homeodomains; for example, only 18 of the homeodomains in this study have binding site information present in the FlyREG database (Bergman et al., 2005) and the specificity of two of these factors is described by only a single binding site. B1H binding site data is now available for all the independent homeodomains in *D. melanogaster*. Perhaps the greatest advantage of the B1H determined homeodomain specificities has been our ability to determine the orientation of the homeodomain on its binding site. This orientation data allows even partially symmetric sites to be properly aligned when constructing recognition motifs. Correct alignment of selected sites is not only critical for a precise ranking of predicted binding sites in genomic DNA sequences, but it is also required to construct an accurate recognition code.

In principal, technical factors such as the use of an N-terminal fusion partner or the number of selected sequences analyzed could have limited the accuracy or precision of our dataset. However, several observations argue against these concerns. First, the ability to cluster factors based on their observed binding specificities and to identify correlations in DNA contact residues demonstrates a striking internal consistency within the dataset. The ability to use this data to rationally alter DNA-binding specificity in mutagenesis experiments provides an essential confirmation of the significance of these correlations. Second, gel shift competition assays of homeodomains expressed independently of the zinc finger fusion behave as the B1H data would predict. Third, comparisons with external data sources - the previously determined *in vitro* and *in vivo* specificity data for fly homeodomains - provides clear evidence of the accuracy of our data. For example, a comparison with the detailed characterization of Ubx, Dfd and Abd-B by Beachy and colleagues (Ekker et al., 1994; Ekker et al., 1992), confirms the high degree of precision observed within some of the specificity groups. Combined with the advantages discussed above, this validation of the homeodomain dataset confirms the utility of the B1H system.

This study provides the first global analysis of homeodomain specificities in an organism. We find that the homeodomain family displays a greater than expected range of specificities. Within the core 6 bp binding site, a wide variety of bases can be preferred at most positions. The majority of homeodomains (93%) in our dataset can be clustered into eleven different specificity groups. Factors that fall within a particular group often share common residues at the DNA recognition positions. Besides the eleven specificity clusters, there are 6 individual homeodomains with divergent specificities. In general, these orphan factors display different combinations of amino acids at the key DNA-recognition positions than are found in the specificity groups. Thus, while the 43 homeodomains in the Antp and En groups all have related binding specificities, we find evidence for at least 17 different-DNA binding specificities within the entire *D. melanogaster* homeodomain set.

Specificity Determinants

An important conclusion from our analysis is that overall sequence similarity between two homeodomains is a useful, but sometimes misleading indicator of the degree of similarity in their DNA-binding specificities. Once factors are clustered into specificity groups, it is possible to overlay this information with their degree of sequence homology.

A substantial correlation between sequence similarity and preferred recognition motif is observed as expected since proteins with greater overall homology are also more likely to be similar at the DNA-recognition positions. However, we observed multiple examples where pairs of closely related homeodomains cluster into different specificity groups. In both naturally-occurring and mutant homeodomains, single amino acid changes at putative DNA recognition positions appear sufficient to significantly alter specificity. Conversely, alteration of only 5 recognition positions in En is sufficient to change its specificity to that of the TGIF family member Achi, although these homeodomains only share 17 common residues. These observations illustrate the importance of defining which amino acid positions contribute to variations in binding site specificity in order to make accurate predictions.

Among the specificity groups, there are clear correlations between amino acid sequence and specificity. Using this data we have produced a qualitative recognition code for homeodomains based on our computational and mutagenic analysis of specificity combined with previous biochemical and structural data. Our recognition code clarifies the key specificity determinants that define variation in sequence recognition preferences among homeodomains. The origin of specificity is much clearer for residues that are found in the recognition helix than for those in the N-terminal arm. At the 3’ end of the binding site, positions 4, 5 and 6 are specified directly by the combination of residues at positions 47, 50 and 54. However, the influence of each of these residues does not map to a single position within the binding site; their influences are complex, where multiple residues can simultaneously affect specificity at multiple binding site positions. The mutational analysis of Bcd highlights this complexity, as residues at all three of these positions impinge upon the nucleotide preference at position 4. Consequently, our recognition code describes how specific residues bias the specificity of a homeodomain towards a particular base or bases, but global specificity preferences are dependent on a constellation of primary and secondary effects at the protein-DNA interface. In some cases, the specificity preferences dictated by one residue are subordinate to a residue at a second position. Arg54 is usually associated with a strong preference for Cyt at binding site position 4, but not in Bcd where Ile47 supercedes Arg54, resulting in a preference for Thy. Even more dramatic is the effect of Ala8 on recognition in the Iroquois group, which disrupts the ability of Arg5 and Arg55 to dictate strong specificity preferences at binding site positions 1 and 2. Interestingly, although Ala8 is highly conserved within this class, introducing a hydrophobic residue (Ala8Phe) reestablishes the expected specificity preferences at binding sites positions 1 and 2. Thus, a binding pocket complementary to an aromatic residue at residue 8 is present in Caup despite the presence of Ala; this pocket may mediate important protein-protein interactions, such as homodimerization observed within this family (Bilioni et al., 2005).

Despite the ability of the B1H homeodomain dataset to identify a large number of homeodomain specificity determinants, there are clearly aspects of DNA-recognition that remain inadequately described. Precisely how variation in N-terminal arm residues contribute to specificity remains mysterious. In addition, the numerous examples of homeodomains with unique specificity strongly suggest that additional combinations of recognition residues can occur that will generate specificities not present in the *D. melanogaster* set. Nonetheless, this dataset is likely to contain examples of the most widely occurring homeodomain specificities in metazoans and thus will provide a powerful tool to predict the specificities of homeodomains in any organism. Because our recognition code is qualitative and incomplete, the sequence similarity throughout the homeodomain as well as at the key recognition positions are important factors for accurate specificity predictions. Taking both of these elements into account, we have predicted the DNA-binding specificities of 79% of the independent homeodomains in the human genome (http://ural.wustl.edu/flyhd/). These specificities should provide an important resource for predicting functional binding sites in regulatory regions of genes that are controlled by these factors. The same principles that were employed to predict the specificity of the human set should be applicable to the prediction of homeodomain specificities in other organisms.

CRM Prediction with Homeodomain Specificities

This dataset should also be useful to help predict target sites for these factors in the *D. melanogaster* genome (Berman et al., 2004; Schroeder et al., 2004; Sinha et al., 2003). The reductionist approach of analyzing isolated homeodomains has resulted in robust determination of monomeric homeodomain specificities and identification of clear correlations between amino acid sequence and binding site preference. However, our dataset will not always provide a complete representation of the “functional” DNA-binding specificity of each factor, as many are known to recognize DNA not only as monomers, but also as heterodimers or higher order complexes within which their preferred target sequence may be modified (Pearson et al., 2005; Ryoo and Mann, 1999; Wilson and Desplan, 1999). A full description of the protein-protein interactions among homeodomains and between homeodomains and other transcription factors will further improve the value of this dataset.

Continued B1H experiments should lead to more detailed understanding of homeodomain recognition. Our current experiments have led to a recognition code that can be used to rationally reengineer the DNA-binding specificity of homeodomains; this is most clearly demonstrated by the successful conversion of Engrailed into a TGIF-like factor. Intermediate specificity alterations were also obtained in these experiments including homeodomains with specificities (TAACA and TGATTA) not observed for any of the natural homeodomains characterized in this study. The throughput of the B1H system will facilitate the synthesis of a more comprehensive recognition model by allowing the characterization of additional homeodomains with differ combinations of recognition residues obtained from other species or mutagenesis experiments. With the B1H system, it should also be straightforward to perform selections on pools of mutagenized homeodomains to select proteins with novel DNA-binding specificity, providing a relatively unbiased assessment of the range of residues that are compatible with recognition of a given motif. As more specificity data is determined for unique homeodomains from disparate species, homodimeric and heterodimeric specificities, and potentially engineered homeodomain specificities, the recognition code for this domain will become more refined. If so, a refined recognition code may make it possible to accurately predict CRMs using predicted homeodomain specificities and therefore alleviate the need to characterize homeodomains in every genome.

Artificial Domains

In addition to the characterization of TF DNA-binding specificity, the omega-based bacterial one-hybrid system provides an intriguing platform for the creation of artificial DNA binding proteins with unique specificities. We reasoned that the same sensitivity and dynamic range demonstrated by the characterization of TF specificity would likely provide a responsive assay for the survey of the protein DNA interface and selection of DBD's with novel specificity. In fact, we have already used this system to engineer Cys_2His_2 Zinc fingers with novel specificity by simply by swapping the bait and prey combination in the selection system (Meng et al., 2008). In other words, by installing a fixed sequence of interest upstream of the reporter, we have selected zinc fingers able to bind that fixed sequence from a library of randomized zinc finger proteins.

Cys_2His_2 Zinc fingers have been the focus of much research since it was first demonstrated that the specificity of this domain could be easily modified (Rebar and Pabo, 1994). The Pabo lab first established the fundamental doctrine of this field: by focusing randomization on positions that structures indicate are able to make base specifying contacts (typically positions -1, 2, 3 and 6 of the recognition helix) and then selecting functional interactions with the target site, zinc fingers that specify a wide range of sequences can be engineered (Rebar and Pabo, 1994; Wolfe et al., 1999; Joung et al., 2000). This general approach has produced zinc fingers that have successfully targeted attached auxiliary domains such as activators, repressors, methyltransferases, and nucleases to specified sequences in multiple genomes (Urnov and Rebar, 2002; Dreier et al., 2005; Smith and Ford, 2007; Perez et al., 2008). Fused to nucleases, zinc fingers have now been used to target double strand breaks in species from *Drosophila* to zebrafish to human cell lines, which have in turn resulted in targeted knockouts and knock-ins (Beumer et al., 2006; Meng et al, 2008; Lombardo et al., 2007).

Though zinc finger nucleases (ZFNs) have been successfully used to cleave their genomic targets, the common designs of ZFNs have limitations that stand in the way their use as a therapeutic. First and perhaps most importantly, the platforms currently in use to engineer zinc fingers result in imperfect specificities that lead to off target lesions. Second, zinc fingers appear to have difficulty specifying A-T rich sequences; limiting the number of potential targets. Third, potential targets are limited to the current architecture of a ZFN, two monomers each consisting of 3 or 4 N-terminal zinc fingers fused to a C-terminal nuclease. These monomers are targeted to binding sites spaced by 5 or 6 base pairs (Cathomen and Joung, 2008). Any potential targets that would require a different architecture of domains, such as an N-terminal nuclease, or different spacings are not currently possible. Finally, the improved efficiency of homologous recombination due to a double strand break decreases dramatically with the distance between the cut site and desired genetic conversion. In fact, a distance of only 45 bases will reduce the efficiency of repair by over 60% (Elliot et al, 1998). Therefore, the ability to target an exact sequence, despite the failings of current designs and selection procedures, may be absolutely critical to the development of these therapeutic agents.

Previous studies have focused on the utility of zinc fingers as a modular, flexible system to create artificial DNA binding domains for genetic engineering of cell lines or animals (Beumer et al., 2006; Urnov et al., 2005). To account for the GC bias of zinc fingers, target sequences are typically chosen by their GC content to accommodate this bias. However, our zinc finger-homeodomain (ZFHD) selection framework, which is derived from ZFHD1(Pomerantz et al., 1995), provides a system in which the specificity of both components can, in principle, be engineered to create hybrid transcription factors that can be used for targeted gene regulation or modification. Incorporation of the homeodomain into the artificial DBD arsenal has the potential to provide a set of modules able to strongly specify A-T rich sequences. Together, zinc fingers and homeodomains would provide a more complete index of sites that are able to be specified efficiently by artificial DBDs. Our analysis of the full complement of *D. melanogaster* homeodomains has begun to catalog the versatility of this tool utilized by nature to recognize a variety of specific sequence elements and provides a blueprint for their future utilization by scientists to manipulate natural systems.

REFERENCES

Adryan, B., and Teichmann, S.A. (2006). FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. Bioinformatics *22*, 1532-1533.

Ades, S.E., and Sauer, R.T. (1994). Differential DNA-binding specificity of the engrailed homeodomain: the role of residue 50. Biochemistry *33*, 9187-9194.

Ades, S.E., and Sauer, R.T. (1995). Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex. Biochemistry *34*, 14601-14608.

Armache K J, Kettenberger H and Cramer P (2003) Architecture of initiation-competent 12-subunit RNA polymerase II; *Proc. Natl.Acad. Sci. USA,* **100,** 6964–68.

Amoutzias GD, Robertson DL, Van de Peer Y, and Oliver SG (2008) Choose your partner: dimerization in eukaryotic transcription factors. Trend in Biochem Sci, 33, 220-29.

Amoutzias GD, Robertson DL, Oliver SG, and Bornberg-Bauer E (2004) Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. EMBO Rep, 5, 274-79.

Amoutzias GD, Pichler EE, Mian N, De Graaf D, Imsiridou a, Robinson-Rechavi M, Bornberg-Bauer E, Roberson DL, and Oliver SG. (2007) A protein interaction atlas for the nuclear receptors: properties and quality of a hub-based dimerization network. BMC Syst Biol, 1, 34-46.

Arnosti, D.N. (2003). Analysis and function of transcriptional regulatory elements: insights from Drosophila. Annu Rev Entomol *48*, 579-602.

Arnosti, D.N., Barolo, S., Levine, M., and Small, S. (1996). The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development *122*, 205-214.

Bhaumik SR, Raha T, Aiello DP, and Green MR (2004) In vivo target of a transcriptional activator revealed by fluorescence resonance energy transfer. Genes & Dev, 18, 333-43.

Bailey, T.L., and 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.

Banerjee-Basu, S., and Baxevanis, A.D. (2001). Molecular evolution of the homeodomain family of transcription factors. Nucl Acids Res *29*, 3258-3269.

Barberis A, Pearlberg J, Simkovich N, Farrell S, Relnagel P, Bamdad C, Sigal G, and Ptashne M. (1995) Contact with a Component of the Polymerase II Holoenzyme Suffices for Gene Activation. *Cell*, **81**, 359-368.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823-837.

Benos, P.V., Lapedes, A.S., and Stormo, G.D. (2002). Probabilistic code for DNA recognition by proteins of the EGR family. Journal of molecular biology *323*, 701-727.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotech *24*, 1429-1435.

Bergman, C.M., Carlson, J.W., and Celniker, S.E. (2005). Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics *21*, 1747-1749.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A *99*, 757-762.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. Genome Biol *5*, R61.

Bertolino, E., Reimund, B., Wildt-Perinic, D., and Clerc, R.G. (1995). A novel homeobox protein which recognizes a TGT core and functionally interferes with a retinoid-responsive motif. J Biol Chem *270*, 31178-31188.

Beumer, K., Bhattacharyya, G., Bibikova, M., Trautman, J.K., and Carroll, D. (2006). Efficient gene targeting in Drosophila with zinc-finger nucleases. Genetics *172*, 2391-2403.

Bilioni, A., Craig, G., Hill, C., and McNeill, H. (2005). Iroquois transcription factors recognize a unique motif to mediate transcriptional repression in vivo. Proceedings of the National Academy of Sciences of the United States of America *102*, 14671-14676.

Borukhov S, and Nudler E (2003) RNA polymerase holoenzyme: structure, function and biological implications. *Curr Opin Microbiol*, **6**, 93-100.

Brent R, and Ptashne M (1985) A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. Cell, 1985, 729-36.

Brown RS, Sander C, and Argos P. (1985). Theprimary structure of transcription factor TFIIIA has 12 consecutive repeats. *FEBS Lett.,* 186, 271-74.

Bucher P (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J *Mol Biol*, **212**, 563-78.

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

Carroll SB (1990) Zebra patterns in fly embryos: activation of stripes or repression of interstripes. *Cell*, **60**, 9-16.

Chang, C.P., Brocchieri, L., Shen, W.F., Largman, C., and Cleary, M.L. (1996). Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. Mol Cell Biol *16*, 1734-1745.

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, **387**, 67-73.

Clyde DE, Corado MSG, Wu X, Pare A, Papatsenko D, and Small S (2003) A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. Nature, 426, 849-53.

Coleman RA and Pugh BF (1995) Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA. *J Biol Chem*, **270**, 13850-59.

Conaway JW, Florens L, Sato S, Tomomori-Sato C, Parmely TJ, Yao T, Swanson SK, Banks CA, Washburn MP, and Conaway RC. (2005) The mammalian Mediator complex. *FEBS Lett.,* **579**, 904-8.

Connolly, J.P., Augustine, J.G., and Francklyn, C. (1999). Mutational analysis of the engrailed homeodomain recognition helix by phage display. Nucl Acids Res *27*, 1182-1189.

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., National Institutes of Health Intramural Sequencing, C., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E.D., *et al.* (2004). From the Cover: Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proceedings of the National Academy of Sciences *101*, 992-997.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome research *14*, 1188-1190.

Damante, G., Pellizzari, L., Esposito, G., Fogolari, F., Viglino, P., Fabbro, D., Tell, G., Formisano, S., and Di Lauro, R. (1996). A molecular code dictates sequence-specific DNA recognition by homeodomains. Embo J *15*, 4992-5000.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A *97*, 6640-6645.

Dave, V., Zhao, C., Yang, F., Tung, C.S., and Ma, J. (2000). Reprogrammable recognition codes in bicoid homeodomain-DNA interaction. Mol Cell Biol *20*, 7673-7684.

Davidson EH (2001) *Genomic regulatory systems: development and evolution*. Academic press. San Diego.

Dearolf, C.R., Topol, J., and Parker, C.S. (1989). The caudal gene product is a direct activator of fushi tarazu transcription during Drosophila embryogenesis. Nature *341*, 340-343.

Deplancke B, Dupuy D, Vidal M and Walhout M (2004) A Gateway-compatible yeast one-hybrid system. Genome Res., 14, 2093-2101.

Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Reece-Hoyes JS, Hope IA, Tissenbaum HA, Mango SE, and Walhout AJM (2006) A Gene-Centered C. elegans Protein-DNA Interaction Network. Cell, 125, 1193-1205.

Dostie, J., and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. Nature protocols *2*, 988-1002.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., *et al.* (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research *16*, 1299-1309.

Dove, S.L., and Hochschild, A. (1998). Conversion of the omega subunit of *Escherichia coli* RNA polymerase into a transcriptional activator or an activation target. Genes Dev *12*, 745-754.

Dove, S.L., Joung, J.K., and Hochschild, A. (1997). Activation of prokaryotic transcription through arbitrary protein-protein contacts. Nature *386*, 627-630.

Dreier B, Fuller RP, Segal DJ, Lund CV, Blancafort P, Huber A, Koksch B, and Barbas CF 3rd (2005) Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors. *J. Biol Chem*, **280**, 35588-97.

Drosophila 12 Genomes Consortium, et al. (2008) Evolution of genese and genomes on the Drosophila phylogeny. *Nature*, **450**, 203-18.

Durai, S., Bosley, A., Abulencia, A.B., Chandrasegaran, S., and Ostermeier, M. (2006). A bacterial one-hybrid selection system for interrogating zinc finger-DNA interactions. Comb Chem High Throughput Screen *9*, 301-311.

Edgar, R. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Ekker, S.C., Jackson, D.G., von Kessler, D.P., Sun, B.I., Young, K.E., and Beachy, P.A. (1994). The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. EMBO J *13*, 3551-3560.

Ekker, S.C., von Kessler, D.P., and Beachy, P.A. (1992). Differential DNA sequence recognition is a determinant of specificity in homeotic gene action. EMBO J *11*, 4059-4072.

Ellington, A.D., and Szostak, J.W. (1990). *In vitro* selection of RNA molecules that bind specific ligands. Nature *346*, 818-822.

Elliot B., Richardson C., Winderbaum J., Nickoloff J.A., and Jasin M. (1998) Gene conversion tracts from double-strand break repair in mammalian cells. *Mol. Cell. Biol.*, 18, 93-101.

Elrod-Erickson M, and Pabo CO (1999). Binding studies with mutants of Zif268. *J. Biol. Chem.* 274:19281-85.

Elrod-Erickson M, Rould MA, Nekludova L, and Pabo CO (1996). Zif268 protein-DNA complex refined at 1.6 A° : a model system for understanding zinc finger-DNA interactions. *Structure, 4,* 1171-80.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author Department of Genome Sciences, University of Washington, Seattle.

Foucher I, Montesinos ML, Volovitch M, Prochiantz A and Trembleau A (2003) Joint regulation of the MAP1B promoter by HNF3b/Foxa2 and engrailed is the result of a highly conserver mechanism for direct interaction of homeoproteins and Fox Transcription factors. Development, 130, 1867-76.

Fraenkel, E., and Pabo, C.O. (1998). Comparison of X-ray and NMR structures for the Antennapedia homeodomain-DNA complex. Nat Struct Biol *5*, 692-697.

Fraenkel, E., Rould, M.A., Chambers, K.A., and Pabo, C.O. (1998). Engrailed homeodomain-DNA complex at 2.2 A resolution: a detailed view of the interface and comparison with other engrailed structures. Journal of molecular biology *284*, 351-361.

Gallo, S.M., Li, L., Hu, Z., and Halfon, M.S. (2006). REDfly: a Regulatory Element Database for Drosophila. Bioinformatics *22*, 381-383.

Gao, F., Foat, B., and Bussemaker, H. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics *5*, 31.

Gentry, D.R., and Burgess, R.R. (1989). rpoZ, encoding the omega subunit of Escherichia coli RNA polymerase, is in the same operon as spoT. Journal of bacteriology *171*, 1271-1277.

Gehring, W.J., Affolter, M., and Burglin, T. (1994a). Homeodomain proteins. Annu Rev Biochem *63*, 487-526.

Gehring, W.J., Qian, Y.Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A.F., Resendez-Perez, D., Affolter, M., Otting, G., and Wuthrich, K. (1994b). Homeodomain-DNA recognition. Cell *78*, 211-223.

Gentry, D.R., and Burgess, R.R. (1989). rpoZ, encoding the omega subunit of Escherichia coli RNA polymerase, is in the same operon as spoT. Journal of bacteriology *171*, 1271-1277.

Georgakopoulos T, Gounalaki N, and Thireos G (1995) Genetic evidence for the interaction of the yeast transcriptional co-activator proteins GCN5 and ADA2. *Mol Gen Genet.*, **246**, 723-8.

Giniger E., Varnum SM, and Ptashne M (1985) Specific DNA binding of GAL4, a positive regulatory protein in yeast. *Cell*, **40**, 767-74.

Goto T, Macdonald P and Maniatis T (1989) Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. Cell, 57, 431-22.

Grad YH, Roth FP, Halfon MS, and Church GM (2004) Prediciton of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D. pseudoobscura. *Bioinformatics*, **20**, 2738-50.

Grandori C, Cowley SM, James LP, and Eisenman RN (2000) The Myc/Max/Mad network and the transcriptional control of cell behavior. Annu Rev Cell Dev Biol., 16, 653-99.

Grant, R.A., Rould, M.A., Klemm, J.D., and Pabo, C.O. (2000). Exploring the role of glutamine 50 in the homeodomain-DNA interface: crystal structure of engrailed (Gln50 --> ala) complex at 2.0 A. Biochemistry *39*, 8187-8192.

Green MR (2005) Eukaryotic Transcription Activation: Right on Target. *Molecular Cell*, **18**, 399-402.

Greisman HA and Pabo CO (1997) Sequential optimization strategy yields high-affinity zinc finger proteins for diverse DNA target sites. Science, 275, 657-61.

Grewal SI, and Elgin SC (2002) Heterochromatin: new possibilities for the inheritance of structure. Curr. Opin. Genet. Dev., 12, 178-87.

Grigoryan G and Keating AE (2006) Structure-based prediction of bZIP partnering specificity. J. Mol. Biol., 355, 1125-42.

Gross CA, Chan C, Dombrowski A, Gruber T, Sharp M, Tupy J, Young B (1998) The functional and regulatory roles of sigma factors intranscription. *Cold Spring Harb Symp Quant Biol*, **63**, 141-155.

Gruschus, J.M., Tsao, D.H., Wang, L.H., Nirenberg, M., and Ferretti, J.A. (1997). Interactions of the vnd/NK-2 homeodomain with DNA by nuclear magnetic resonance spectroscopy: basis of binding specificity. Biochemistry *36*, 5372-5380.

Gutell, R.R., Power, A., Hertz, G.Z., Putz, E.J., and Stormo, G.D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucl Acids Res *20*, 5785-5795.

Hanes, S.D., and Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. Cell *57*, 1275-1283.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J*., et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99-104.

Harding K, Joey T, Warrior R and Levine M (1989) Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila. EMBO J, 8, 1205-12.

Harvey, R.P. (1996). NK-2 homeobox genes and heart development. Dev Biol *178*, 203-216.

Hazbun, T.R., Stahura, F.L., and Mossing, M.C. (1997). Site-specific recognition by an isolated DNA-binding domain of the sine oculis protein. Biochemistry *36*, 3680-3686.

Hertz, G.Z., and Stormo, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics *15*, 563-577.

Hochshild A and Dove SL (1998) Protein-Protein Contacts that activate and Repress Prokaryotic Transcription. *Cell*, **92**, 597-600.

Hoffmann, C., Minkah, N., Leipzig, J., Wang, G., Arens, M.Q., Tebas, P., and Bushman, F.D. (2007). DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic acids research *35*, e91.

Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, and Young RA (1998) Dissection the Regulatory Circuitry of a Eukaryotic Genome. *Cell*, **95**, 717-28.

Hovde, S., Abate-Shen, C., and Geiger, J.H. (2001). Crystal structure of the Msx-1 homeodomain/DNA complex. Biochemistry *40*, 12013-12021.

Hu, J.C., Kornacker, M.G., and Hochschild, A. (2000). *Escherichia coli* one- and two-hybrid systems for the analysis and identification of protein-protein interactions. Methods *20*, 80-94.

Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E, and Stubbs L (2006) A comprehensive catalog of human KRAB-associated zinc figer genes: Insight into the evolutionary history of a large family of transcriptional repressors. Genome Research, 16, 669-77.

Ingham PW (1988) The molecular genetics of embryonic pattern foramtion in Drosophila. *Nature*, 335, 25-34.

Jaeger, J., and Reinitz, J. (2006). On the dynamic nature of positional information. BioEssays *28*, 1102-1111.

Jenuwein T (2001) Re-SET-ting heterochromatin by histone methyltransferases. *Trends Cell. Biol*., **11** 266-73.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497-1502.

Jones S (2004) An overview of the basic helix-loop-helix proteins. Genome Biology, 5, 226-32.

Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., and Mann, R.S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell *131*, 530-543.

Joung JK, Koepp DM, and Hochschild A (1994) Synergistic Activation of Transcription by Bacteriophage λcI Protein and *E. coli* cAMP Receptor Protein. *Science*, **265**, 1863-66.

Joung JK, Le JU, and Hochschild A (1993) Synergistic activation of transcription by *Escherichia coli* cAMP receptor protein. *Proc. Natl. Acad. Sci. USA*, **90**, 3083-87.

Joung JK, Ramm, EI, and Pabo CO (2000) A bacterial two-hybrid selection system for the studying protein-DNA and protein-protein interactions. *Proc. Natl. Acad. Sci. USA,* **97***,* 7382-87.

Kehle J, Beuchle D, Treuheit S, Christen B, Kennison JA, Bienz M, and Muller J (1998) dMi-2 a Hunchback interaction protein that functions in *polycomb* repression. Science, 282, 1897-1900.

Khattak S, Lee BR, Cho SH, Ahnn J, and Spoerel NA (2002) Genetic characterization of Drosophila Mi-2 ATPase. Gene, 293, 107-114.

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 Drosophila genomes. Genome research *17*, 1919-1931.

Kuhn DT, Chaverri JM, Persaud DA, and Madjidi A (2000) Pair-rule genes cooperate to activate en stripe 15 and refine its margins during germ band elongation in the D. melanogaster embryo. Mech Dev, 95, 297-300.

Kim, C.A., and Berg, J.M. (1996). A 2.2 A resolution crystal structure of a designed zinc finger protein bound to DNA. Nat Struct Biol *3*, 940-945.

Kim SY and Kim Y (2006) Genome-wide prediction of transcriptinal regulatory elements of human promoters using gene epxression and promoter analysis data. BMC Bioinformatics, 7, 330.

Kim, D.W., Kempf, H., Chen, R.E., and Lassar, A.B. (2003). Characterization of Nkx3.2 DNA binding specificity and its requirement for somitic chondrogenesis. J Biol Chem *278*, 27532-27539.

Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B., and Pabo, C.O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: A framework for understanding homeodomain-DNA interactions. Cell *63*, 579-590.

Klemm, J.D., and Pabo, C.O. (1996). Oct-1 POU domain-DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. Genes Dev *10*, 27-36.

Klug A (1995) Gene regulatory proteins and ther interaction with DNA. *Ann NY Acad. Sci.* **758**, 143-60.

Kohler JJ and Schepartz A (2001) Kinetic Studies of Fos-June-DNA Complex Formation: DNA Binding Prior to Dimerization. Biochem., 40, 130-142.

Kuras L, and Struhl K (1999) Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. *Nature*, **399**, 609-13.

Lander, E. S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I*., et al.* (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science *298*, 799-804.

Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, and Wasserman WW (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol*., **2**, 13-17.

Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., and Bork, P. (2006). SMART 5: domains in the context of genomes and networks. Nucl Acids Res *34*, D257-260.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147-151.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276,* 565-570.

Li, X.Y., Macarthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C.L*., et al.* (2008). Transcription Factors Bind

Thousands of Active and Inactive Regions in the Drosophila Blastoderm. PLoS biology *6*, e27.

Lieb, J.D., Liu, X., Botstein, D., and Brown, P.O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet *28*, 327-334.

Lifanov, A.P., Makeev, V.J., Nazina, A.G., and Papatsenko, D.A. (2003). Homotypic regulatory clusters in *Drosophila*. Genome research *13*, 579-588.

Linnell, J., Mott, R., Field, S., Kwiatkowski, D.P., Ragoussis, J., and Udalova, I.A. (2004). Quantitative high-throughput analysis of transcription factor binding specificities. Nucleic acids research *32*, e44.

Liu, X., Brutlag, D.L., and Liu, J.S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput *6*, 127-138.

Lombardo A. et al. (2007) Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nature Biotechnology*, **25**, 1298-1306.

Lonetto M, Gribskov M, and Gross CA (1992) The σ70 family: sequence conservation and evolutionary relationships. *J Bacteriol*,**174**, 3843-49.

Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. Nature *403*, 564-567.

Ludwig, M.Z., Patel, N.H., and Kreitman, M. (1998). Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development *125*, 949-958.

Luscher B (2001) Function and regulation of the transcription factors of the Myc/Max/Mad network. Gene, 277, 1-14.

Luscombe NM, Austin SE, Berman HM, and Thorton JM (2000) An overview of the structures of protein-DNA complexes. Genome Biol., 1, 1-37.

Ma J, and Ptashne M (1987) Deletion analysis of Gal4 defines two transcriptional activating segments. Cell, 48, 847-53.

Mahony, S., Auron, P.E., and Benos, P.V. (2007). Inferring protein-DNA dependencies using motif alignments and mutual information. Bioinformatics *23*, i297-304.

Maldonado E, Ha I, Cortes P, Weis L, and Reinberg D (1990) Factors involved in specific transcription by mammalian RNA polymerase II: role of transcription factors IIA, IID, and IIB during formation of a transcription-competent complex. *Mol Cell Biol*, **10**, 6335-47.

Manoukian AS, and Krause HM (1992) Concentration-dependent activities of the even skipped protein in Drosophila embryos. Genes & Dev, 6, 1740-51.

Manoukian AS, and Krause HM (1993) Control of segmental asymmetry in Drosophila embryos. Development,118, 785-96.

Margalit, Y., Yarus, S., Shapira, E., Gruenbaum, Y., and Fainsod, A. (1993). Isolation and characterization of target sequences of the chicken CdxA homeobox gene. Nucleic acids research *21*, 4915-4922.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.*, et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature.

Markstein, M., Markstein, P., Markstein, V., and Levine, M.S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc Natl Acad Sci U S A *99*, 763-768.

Marshak S, Benshushan E, Shoshkes M, Havin L, Cerasi E and Melloui D (2000) Functional conservation of regulatory elements in the pdx-1 gene; PDX-1 and hepatocyte nuclear factor 3β transcription factors mediate B-cell-specific expression. Mol. Cell Biol, 20-7583-90.
Martinowich K, Hattori D, Wu H, Fouse S, He F, Hu Y, Fan G, and Sun YE (2003) DNA-methylation-related chromatin remodeling in activity-dependent *Bdnf* gene regulation. *Science*, **302**, 890-93.

Maston GA, Evans SK, and Green MR (2006) Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet.*, **7**, 29-59.

Matsui T, Segall J, Weil PA, and Roeder RG (1980) Multipole factors required for accurate initiation of transcription by purified RNA polymerase II. *J. Biol. Chem.*, **255**, 11992-6.

Mathias, J.R., Zhong, H., Jin, Y., and Vershon, A.K. (2001). Altering the DNA-binding specificity of the yeast Matalpha 2 homeodomain protein. J Biol Chem *276*, 32696-32703.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. Nucl Acids Res *34*, D108-110.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V.*, et al.* (2003). TRANSFAC(R): transcriptional regulation, from patterns to profiles. Nucl Acids Res *31*, 374-378.

McCue LA, Thompson W, Carmack CS, and Lawrence CE (2002) Factors influencing the identification of transcription factor binding specificity by cross species comparison. *Genome Res*., **12**, 1523-32.

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

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

Meng, X., Smith, R.M., Giesecke, A.G., Joung, J.K., and Wolfe, S.A. (2006). Counter-selectable marker for bacterial-based interaction trap systems. Biotechniques *40*, 179-184.

Meng, X., and Wolfe, S.A. (2006). Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. Nat protocols *1*, 30-45.

Miller J, McLachlan AD, and Klug A (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J, 4, 1609-14.

Miller, J.C., and Pabo, C.O. (2001). Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. Journal of molecular biology *313*, 309-315.

Mooney RA and Landick R (1999) RNA Polymerase Unveiled. *Cell*, **98**, 687-690. Cramper P (2004) Structure and Function of RNA Polymerase. Adv. Protein Chem, 67, 1-42.

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

Mukherjee, K., and Burglin, T.R. (2007). Comprehensive Analysis of Animal TALE Homeobox Genes: New Conserved Motifs and Cases of Accelerated Evolution. J Mol Evol *65*, 137-153.

Myasnikova E, Samsonova A, Kozlov K, Samsonova M, and Reinitz J (2001) Registration of the expression patterns of Drosophila segmentation genes by two independent methods. Bioinformatics, 17, 3-12.

Newman JR and Keating AE (2003) Comprehensive indentification of human bZIP interactions with coiled-coil arrays. Science, 300, 2097-2101.

Ng HH, and Bird A (2000) Histone deacetylases: silencers to hire. *Trends Biochem. Sci*, **25**, 121-26.

Nielsen AL, Oulad-Abdelghani M, Ortiz JA, Remboutsika E, Chambon P, and Losson R (2001) Heterochromatin formation in mammalian cells; interaction between histones and HP1 proteins. *Mol. Cell*, **7,** 729-39.

Oikawa T and Yamada T (2002) Molecular biology of the Ets family of transcription factors. Gene, 303, 11-34.

Parsell, D.A., Silber, K.R., and Sauer, R.T. (1990). Carboxy-terminal determinants of intracellular protein degradation. Genes Dev *4*, 277-286.

Passner, J.M., Ryoo, H.D., Shen, L., Mann, R.S., and Aggarwal, A.K. (1999). Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. Nature *397*, 714-719.

Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, and Burley SK (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes & Dev*, **13**, 3217-30.

Pavletich NP and Pabo CO (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1A. Science, 252, 809-17.

Pearson, J.C., Lemons, D., and McGinnis, W. (2005). Modulating Hox gene functions during animal body patterning. Nat Rev Genet *6*, 893-904.

Peel, A.D., Chipman, A.D., and Akam, M. (2005). Arthropod Segmentation: beyond the Drosophila paradigm. Nat Rev Genet *6*, 905-916.

Pellizzari, L., Tell, G., Fabbro, D., Pucillo, C., and Damante, G. (1997). Functional interference between contacting amino acids of homeodomains. FEBS Lett *407*, 320-324.

Percival-Smith, A., Müller, M., Affolter, M., and Gehring, W.J. (1990). The interaction with DNA of wild-type and mutant fushi tarazu homeodomains. EMBO J *9*, 3967-3974.

Perez E.E., et al. (2008) Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature Biotechnology*, **26**, 808-816.

Pick, L. (1998). Segmentation: Painting stripes from flies to vertebrates. Developmental Genetics *23*, 1-10.

Piper, D.E., Batchelor, A.H., Chang, C.P., Cleary, M.L., and 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*, 587-597.

Pomerantz, J.L., Sharp, P.A., and Pabo, C.O. (1995). Structure-based design of transcription factors. Science *267*, 93-96.

Ptasne M and Gann A. (1997) Transcriptional activation by recruitment. Nature, 386, 569-77.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E.D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics *3*, 30.

Rea S, Eisenhaber F, O'Carroll D, Strahl BD, Sun ZW, Schmid M, Opravil S, Mechtler K, Ponting CP, Allis CD, and Jenuwein T (2000) Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, **406**, 593-99.

Rebar E.J. and Pabo C.O. (1994) Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science*, **263**, 671-73.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, and Young RA (2000) Genome-Wide Location and Function of DNA Binding Proteins. Science, 290, 2306-09.

Rothe M, Nauber U, and Jackle H (1989) Three hormone receptor-like Drosophila genes encode an identical DNA-binding finger. EMBO J., 8, 3087-94.

Roulet, E., Busso, S., Camargo, A.A., Simpson, A.J., Mermod, N., and Bucher, P. (2002). High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol *20*, 831-835.

Ryoo, H.D., and Mann, R.S. (1999). The control of trunk Hox specificity and activity by Extradenticle. Genes Dev *13*, 1704-1716.

Sabo, P.J., Humbert, R., Hawrylycz, M., Wallace, J.C., Dorschner, M.O., McArthur, M., and Stamatoyannopoulos, J.A. (2004). Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proceedings of the National Academy of Sciences *101*, 4537-4542.

Sauer F, Hansen SK, and Tjian (1995)a. DNA Template and Activator-Coactivator Requirements for Transcriptional Synergism by Drosophila Bicoid. Science, 270, 1825-28.

Sauer F. Hansen SK, and Tjian (1995)b. Multiple TAFIIs Directing Synergistic Activation of Transcription. Science, 270, 1783-88.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucl Acids Res *18*, 6097-6100.

Schroeder, M.D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E.D., and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. PLoS Biol *2*, E271.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. Proceedings of the National Academy of Sciences of the United States of America *95*, 5857-5864.

Sharrocks AD, Brown AL, Ling Y, and Yates PR (1997) The ETS-doman Transcription factor family. Int. J. Biochem. Cell Biol., 29, 1371-1387.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S*., et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.

Sinha, S., Liang, Y., and Siggia, E. (2006). Stubb: a program for discovery and analysis of cis-regulatory modules. Nucleic acids research *34*, W555-559.

Sinha, S., Schroeder, M.D., Unnerstall, U., Gaul, U., and Siggia, E.D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC Bioinformatics *5*, 129.

Sinha, S., van Nimwegen, E., and Siggia, E.D. (2003). A probabilistic method to detect regulatory modules. Bioinformatics *19 Suppl 1*, i292-301.

Smale ST, and Kadonaga JT (2003) The RNA polymerase II core promoter. *Annu Rev Biochem*, **72**, 449-79.

Small S, Kraut R, Hoey T, Warrior R, and Levine M (1991) Transcriptional regulation of a pair-rule stripe in Drosophila. Genes & Dev., 5, 827-839.

Small S., Blair A, and Levine M (1992) Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO, 11, 4047-57.

Small S., Blair A, and Levine M (1996) Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. Dev Biol, 175, 314-324.

Smith A.E. and Ford K.G. (2007) Specificity targeting of cytosine methylation to DNA sequences in vivo. *Nucleic Acids Research*, **35**, 740-754.

Sokal, R.R., and Rohlf, F.J. (1995). Biometry: the principles and practice of statistics in biological research., 3rd edition edn (New York, W. H. Freeman and Co.).

Sosinsky, A., Bonin, C.P., Mann, R.S., and Honig, B. (2003). Target Explorer: An automated tool for the identification of new target genes for a specified set of transcription factors. Nucleic acids research *31*, 3589-3592.

Spiegelman BM and Heinrich R (2004) Biological control through regulated transcriptional coactivators. Cell, 119, 157-67.

St Johnston, D., and Nusslein-Volhard, C. (1992). The origin of pattern and polarity in the Drosophila embryo. Cell *68*, 201-219.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A.*, et al.* (2002). The Generic Genome Browser: A Building Block for a Model Organism System Database. Genome Res *12*, 1599-1610.

Stormo GD and Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. TIBS, 23, 109-13.

Stormo GD (1998) Information content and Free Energy in DNA-Protein Interactions. J. theor. Biol, 195, 135-37.

Subramaniam, V., Jovin, T.M., and Rivera-Pomar, R.V. (2001). Aromatic amino acids are critical for stability of the bicoid homeodomain. J Biol Chem *276*, 21506-21511.

Sweetser D, Nonet M, Young RA (1987) Prokaryotic and eukaryotic RNA polymerase have homologous core subunits. *Proc Natl Acad Sci USA*, **84**, 1192-96.

Tautz D (1988) Regulation of the Drosophila segmentation gene hunchback by two maternal morphogenetic centres. Nature, 332, 281-84.

Thiel G, Lietz M, and Hohl M (2004) How mammalian transcriptional repressors work. *Eur. J. Biochem*, **271**, 2855-62.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res *22*, 4673-4680.

Tolon RM, Castillo AI, Jimenez-Lara AM, and Aranda A (2000) Association with Ets-1 causes ligand and Af2 independent activation of the nuclear receptor. Mol. Cell Biol, 20, 8793-8802.

Tomancak, P., Berman, B.P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., Celniker, S.E., and Rubin, G.M. (2007). Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol *8*, R145.

Treisman, J., Gönczy, P., Vashishtha, M., Harris, E., and Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. Cell *59*, 553-562.

Trooskens, G., De Beule, D., Decouttere, F., and Van Criekinge, W. (2005). Phylogenetic trees: visualizing, customizing and detecting incongruence. Bioinformatics *21*, 3801-3802.

Tsai FT and Sigler PB (2000) Structural basis of preinitiation complex assembly on human pol II promoters. *EMBO J*, **19**, 25-36.

Tucker-Kellogg, L., Rould, M.A., Chambers, K.A., Ades, S.E., Sauer, R.T., and Pabo, C.O. (1997). Engrailed (Gln50-->Lys) homeodomain-DNA complex at 1.9 A resolution: structural basis for enhanced affinity and altered specificity. Structure *5*, 1047-1054.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505-510.

Tupler, R., Perini, G., and Green, M.R. (2001). Expressing the human genome. Nature *409*, 832-833.

Urnov, F.D., Miller, J.C., Lee, Y.L., Beausejour, C.M., Rock, J.M., Augustus, S., Jamieson, A.C., Porteus, M.H., Gregory, P.D., and Holmes, M.C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature *435*, 646-651.

Urnov F.D. and Rebar E.J. (2002) Designed transcription factors as tools for therapeutics and functional genomics. *Biochem. Pharmacol.*, **64**, 919-923.

Vassylyev DG, Sekine S, Laptenko O, Lee J, Vassylyeva MN,Borukhov S, Yokoyama S: Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A ° resolution. (2002) *Nature*,**417**, 712-19.

Verrijzer, C.P., Alkema, M.J., van Weperen, W.W., Van Leeuwen, H.C., Strating, M.J., and van der Vliet, P.C. (1992). The DNA binding specificity of the bipartite POU domain and its subdomains. EMBO J *11*, 4993-5003.

Wade PA, Gegonne A, Jones, PL, Ballestar E, Aubry F, and Wolffe AP (1999) Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. Nat Genet., 23, 62-6.

Wade PA, Pruss D, and Wolffe AP (1997) Histone acetylation: chromatin in action. *Trends Biochem. Sci*, **22**, 128-132.

Wang X, Lee C., Gilmour DS, and Gergen JP (2007) Transcription elongation controls cell fate specification in Drosophila embryo. *Genes & Dev.,* **21**, 1031-36.

Wang, T., and Stormo, G.D. (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics *19*, 2369-2380.

Widom J (1998) Structure, dynamics, and function of chromatin in vitro. *Annu Rev Biophys Biomol Struct*, **27**, 285-327.

Wijchers PJEC, Burbach PH, and Smidt MP (2006) In control of biology: of mice, men and Foxes. J. Biochem, 397, 233-246.

Wilson D., Charoensawan V, Kummerfeld SK, and Teichmann SA (2008) DBD Taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.,* **36**, 88-92.

Wilson, D.S., and Desplan, C. (1999). Structural basis of Hox specificity. Nat Struct Biol *6*, 297-300.

Wilson, D.S., Sheng, G., Jun, S., and Desplan, C. (1996). Conservation and diversification in homeodomain-DNA interactions: a comparative genetic analysis. Proc Natl Acad Sci U S A *93*, 6886-6891.

Wolberger, C. (1996). Homeodomain interactions. Curr Opin Struct Biol *6*, 62-68.

Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D., and Pabo, C.O. (1991). Crystal structure of a MATalpha2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell *67*, 517-536.

Wolfe, S.A., Greisman, H.A., Ramm, E.I., and Pabo, C.O. (1999). Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. J Mol Biol *285*, 1917-1934.

Wolfe SA, Nekludova L, and Pabo CO (2000) DNA Recognition by Cys_2His_2 Zinc Finger Proteins. Annu Rev Biophys Biomol Struct, 3, 183-212.

Woodhead L, and Johns EW (1976) The isolation of nucleosomes from Saline-washed chromatin. *FEBS Lett*., **62**, 115-117.

Wright, W.E., and Funk, W.D. (1993). CASTing for multicomponent DNA-binding complexes. Trends Biochem Sci *18*, 77-80.

Zeitlinger, J., Zinzen, R.P., Stark, A., Kellis, M., Zhang, H., Young, R.A., and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev *21*, 385-390.

Zhu, C.C., Dyer, M.A., Uchikawa, M., Kondoh, H., Lagutin, O.V., and Oliver, G. (2002). Six3-mediated auto repression and eye development requires its interaction with members of the Groucho-related family of co-repressors. Development *129*, 2835-2849.

Appendix

Table A.1

Amino acid sequence, selection promoter strength/stringency and the binding sites recovered for each AP factor assayed.

Abd-A				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RRRGRQTYTRFQTLELEKEFHFNHYLTRRRRIEIAHALCLTERQIKIWFQNRRMKLKKELRAVKEINEQAR			
Selected sequences				
>AbdAG04	>AbdAG09	>AbdAH02	>AbdAH07	>AbdAH12
TTTTTAATTA	TACCAAACCC	TACGTAACTT	CGTTCTTTAA	ATCTTAATTAC
>AbdAG05	>AbdAG10	>AbdAH03	>AbdAH08	>AbdAG02
CTACCATTTT	CACTAATTA	GGTCATTAAA	GATTTAATTA	GGTAATTAAA
>AbdAG06	>AbdAG11	>AbdAH04	>AbdAH09	>AbdAG03
TTAATTAC	CATAATTA	TTTTTTATGA	GCGCTAATGA	TCGTTAATGA
>AbdAG07	>AbdAG12	>AbdAH05	>AbdAH10	
TTCTTTATTA	GTTTTAATTA	CTACTAATTC	TGTTTAATGA	
>AbdAG08	>AbdAH01	>AbdAH06	>AbdAH11	
GGACCCACAT	GGTTGCGGCC	TGCAATTAAA	GGCAATTAAG	

Abd-B				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	VRKKRKPYSKFQTLELEKEFLFNAYVSKQKRWELARNLQLTERQVKIWFQNRRMKNKKNSQRQANQQNNNN			
Selected sequences				
>abdb2	>abdb7	>abdb13	>abdb19	>abdb24
GGGTTTATAG	GGTTTACAAC	TTTTTATAAC	GCTTTTATTA	TGTTTTATGA
>abdb3	>abdb8	>abdb14	>abdb20	
GTTTTATTGT	CGTTTAATGT	TTATTAATTA	ACTTTTACGA	
>abdb4	>abdb10	>abdb15	>abdb21	
TTTTTTATGG	TGATTTATGT	GTTTTATGA	TGATTTATTA	
>abdb5	>abdb11	>abdb17	>abdb22	
TGATTAATGG	CATATTATGA	AGTTTTATGG	TGATTTATTA	
>abdb6	>abdb12	>abdb18	>abdb23	
CGCTTTATGT	GCATTTATTA	TCTTTAACGA	TCTTTAATTA	

Antp	
Promoter-Stringency	UV5m-10mM

Amino Acid Sequence	RKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQ NRRMKWKKENKTKGEPGSGGEGD			
Selected sequences				
>AntpA02	>AntpA07	>AntpB01	>AntpB08	
GCCTTAATTA	GGTTTAATGA	TTCATAATTA	GGCAATTAAG	
>AntpA03	>AntpA08	>AntpB02	>AntpB09	
AATTTAATTA	GGC TTAATGA	GTGTTAATTA	CGTTTAATTA	
>AntpA04	>AntpA10	>AntpB04	>AntpB11	
AGCTTAATGA	CTACTAATTA	GATTTAATTA	TTTTTAATGA	
>AntpA05	>AntpA11	>AntpB06	>AntpB12	
TTGTTAATGA	CCCTTAATGG	TGTTTAATGA	CCTTTAATGA	
>AntpA06	>AntpA12	>AntpB07		
GGTAATTAAA	TAGCACTTTT	TTTTTAATGA		

Bcd				
Promoter-Stringency	UV5m-5mM			
Amino Acid Sequence	PRRTRTTFTSSQIAELEQHFLQGRYLTAPRLADLSAKLALGTAQVKIWFK NRRRRHKIQSDQHKDQSYEG			
Selected sequences				
>Bcd21G1	>Bcd21G6	>Bcd21G12	>Bcd21H5	>Bcd21H10
CTGAAAATCT	CTGGTTTAAC	AGGACTAAGC	CCTTAAATCT	CTATTAAGCT
>Bcd21G2	>Bcd21G7	>Bcd21H1	>Bcd21H6	>Bcd21H11
CCTTAAGTCG	CTTCTAATCC	TGGGTAATCT	ATTCTAATCT	ATCCAAATCC
>Bcd21G3	>Bcd21G9	>Bcd21H2	>Bcd21H7	
AGATAAGTCA	CAATCAATCC	GGGAATTAGA	TCTTCAATCC	
>Bcd21G4	>Bcd21G10	>Bcd21H3	>Bcd21H8	
TTGTAAGCTG	AACAAATCCT	ACCGCTAAGC	GTTTAAGCCC	
>Bcd21G5	>Bcd21G11	>Bcd21H4	>Bcd21H9	
TTCCATAATCT	TAGATTAATG	CGACCTAATC	CCTTAAGCTA	

Bcd				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PRRTRTTFTSSQIAELEQHFLQGRYLTAPRLADLSAKLALGTAQVKIWFK NRRRRHKIQSDQHKDQSYEG			
Selected sequences				
>bcd1	>bcd6	>bcd11	>bcd17	>bcd23
TGTTAATCCG	TCTTAATCCC	CGGGTAATCC	GGTTATCCG	GGTTAATCCG
>bcd2	>bcd7	>bcd13	>bcd18	>bcd24
ATGGATTAGA	GCTTAATCCG	TGTTAATCC	TGTTAATCCC	ATGGATTAGA

>bcd3	>bcd8	>bcd14	>bcd20	
CGTTAATCTC	GGGTTAATCC	TGGGATTATA	CGCTTAATCC	
>bcd4	>bcd9	>bcd15	>bcd21	
GGTTTAATCC	GAGATAATCC	GCGTAATCCA	TTACTAATCC	
>bcd5	>bcd10	>bcd16	>bcd22	
TCTATAATCC	AGCTTATCC	GGCTTAAGCC	GTCCTAATCC	

Blimp-1	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	GKMHYECNVCCKTFGQLSNLKVHLRTHSGERPFKCNVCTKSFTQLAHLQKHHLVHTGEKPHQCDICKKRFSSTSNLKTHLRLHSGQKPYACDLCPQKFTQFVHLKLHKRLHTNDRPYVCQGCDKKYISASGLRTHWKTTSCKPNNLEEE
Selected Sequences	
>CG5249C3	
CCATGGCCCGTAAGAGAAAGTGAGAGTG	
>CG5249C4	
GCATGAGAGTGAAAGTTAGCTCAACTAG	
>CG5249C6	
GCAAAATCAGTGAAAGTGGCGGGCGCAT	
>CG5249C7	
GCGCCGACGAGACTAAAAGTGAAAGTTC	
>CG5249C9	
GGCCACTGAAAGTGAAAGCTGGCCACCA	
>CG5249C11	
CCTACACTGGCTGAACGAAAGCGAAAGT	
>CG5249D1	
GCCGTGTAATCTGAAGAAAGTGAAAACA	
>CG5249D2	
TACGGGACGAACGAAAGTGAAAGCAAGT	
>CG5249D3	
CTAAAAAGTGAAAGTCCTGCTCTGGATG	
>CG5249D4	
CAGCTTCAAGTCCCGAAAGGGAAAGTT	
>CG5249D5	
GCTGCAAAAGTGAAAGTAGCCAAAAACG	
>CG5249D6	
GAAAATGAAAACGAAAGTGCGCGCATCC	

Btd	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	QHICHIPGCERLYGKASHLKTHLRWHTGERPFLCLTCGKRFSRSDELQRHGRTHTNYRPYACPICSKKFSRSDHLSKHKKTHFKDKK
Selected Sequences	
>5BtdE2	>5BtdF7
GCCCCCCCTCCTAGATACGCCCCCCTAA	AGGCCATCGAAAGTCAGGAGGGCGGACA

>5BtdE3	>5BtdF8
ACTAGGTGCACTGGGAAGTAGGCGGACA	ACGACGCGGAGGGGGCGTGACTATTACA
>5BtdE4	>5BtdF9
GAGAATACGCCCACATGCACGAACTCTC	CGAAACTCGCAGATAAGAGGGGCGGATT
>5BtdE5	>5BtdF10
CCTAACCTCTATGGACCGTGGGCGGTAG	CGAAGATAACCCCCCGAAGGGGCGGCAC
>5BtdE6	>5BtdF11
TTGAAAGCCGACAACGAGAAGGCGTATG	AAGGTAGACGTAAACTTTGGGGGCGGAGT
>5BtdE7	>5Btd2G1
CCAACCCAAGACGAAAAGGGGCGGAACG	GAACAGTGGGCGGGTACGAAGTCACTAA
>5BtdE8	>5Btd2G2
GAAGCATAAGAAAAAAAAGGGGAGGATG	ACCCCCGCAAGTATTGACAGTTGCCATG
>5BtdE9	>5Btd2G3
GCATGCAGTTACACAGAAGGGGGCGGGT	ACCCCCGCAAGTATTGACAGTTGCCATG
>5BtdE10	>5Btd2G4
GATAAGTGAAATATGGAGAGGGCGGGTG	GAAGCATAAGAAAAAAAAGGGGAGGATG
>5BtdE11	>5Btd2G5
AAGTATCGCCCACAACCCCGCCCCTCAAA	ACGCCCGTACGCTTACGCCCCCTTACAG
>5BtdE12	>5Btd2G6
AACCCAGACAGCAAAAACGCCCCCCAAT	TAACCGGCGACGGCTGAAAAGGGGCGGG
>5BtdF1	>5Btd2G7
ACACCCTTGCTACCATACGCCCACGAAA	GGCGTAAGCATCATTGTAGGGGCGGTAC
>5BtdF2	>5Btd2G9
GCCCACGTCCTTCAACACAGGGGCGGACA	ATGTACCCCTCGCCATCCGCCCCCCACA
>5BtdF3	>5Btd2G10
GCGACAGAGTGGGCGGATGTGAAAGACA	GACCTATCCGAGGGGCGGGGAAATACGA
>5BtdF4	>5Btd2G11
CGATGACCTTGGGGGCGGGGCCAAAACACC	AGAAGTTGTCAACTCACGCCCACACCAA
>5BtdF5	>5Btd2G12
AATTTAGAGGGCGTTTCATATTAATGCG	AGAAGTTGTCAACTCACGCCCACACCAA

Cad				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KDKYRVVYTDFQRLELEKEYCTSRYITIRRKSELAQTLSLSERQVKIWFQ NRRAKERKQNKKGSDPNVMGVG			
Selected sequences				
>CadE1	>CadE11	>CadF8	>Cad2E7	>Cad2F4
CCACAAATTA	AGCAATTAAG	CTCAATAAAA	ACTCTAATTG	ACCATAATTA
>CadE2	>CadE12	>CadF9	>Cad2E8	>Cad2F5
CTAATCAACA	CTAATAAAA	ATCATAAAAC	CCAATAAACT	GCAATAAAAA
>CadE3	>CadF1	>CadF10	>Cad2E9	>Cad2F6
GTAATAACTT	ACCGTAATTA	ATTCCGCTCT	GCAATCATTA	CTTTTTATTG
>CadE4	>CadF2	>CadF11	>Cad2E10	>Cad2F7
GAATTAATAG	CCAATAAATG	GTAATAAAGT	CCTTAAATTA	CCCATAAATT
>CadE5	>CadF3	>Cad2E1	>Cad2E11	>Cad2F9
GCTTAAATGA	GCCATTAAAG	CCAATAAAGG	AAAAGGATTC	GTTTTTATGA

>CadE6	>CadF4	>Cad2E2	>Cad2E12	>Cad2F10
ATGATTTATTT	AGTTTAATAA	GACATTATTA	AGGCACTACG	CCCATATAAT
>CadE8	>CadF5	>Cad2E3	>Cad2F1	>Cad2F11
CTTATAAAAT	CTATTTATTA	GCTAATAAAT	GTTCTAATTA	GCAATAAAAA
>CadE9	>CadF6	>Cad2E4	>Cad2F2	
ACAGTAATTA	TATTTTATTA	TTATTTATTA	CTCATAAACA	
>CadE10	>CadF7	>Cad2E5	>Cad2F3	
GAACACTACT	GCAATAAACA	TCGAGCATGT	GGATTTATAA	

D	
Promoter-Stringency	UV5m-5mM
Amino Acid Sequence	MHSLATSPGQEGHIKRPMNAFMVWSRLQRRQIAKDNPKMHNSEISKR LGAEWKLLAESEKRPFIDEAKRLRALHMKEHPDYKYRPRRKPKNPLT AGPQGGLQMQ
Selected Sequences	
>DykD1	>DykE6
AATCCCATTGTTATACATGCTGCAATAT	GAAAAATAAACACAAAAGGATACTATAA
>DykD2	>DykE7
CGTCCCACGCAACACAATGGACAACATA	CTACGCGTCAAACAAAAGGCGCGAAGAA
>DykD3	>DykE8
TCGAGCCACGTGTACCATTGTTGTAGGA	TAGACAACCATGACTGGACATTCAAGAG
>DykD4	>DykE9
TAGACAACCATGACTGGACATTCAAGAG	GCGGCATAAGAACAAAGGATTTCTAGCT
>DykD5	>DykE10
TTCCACATTGACCATAGCCAGCCACTCC	ACCCCCGCAGTATTGACAGTTGCCATG
>DykD6	>DykE11
CAATGGGCCTTGATCCATTGTTCACGAC	CGCATACTAGAACAATAGGCCACTACGAA
>DykD7	>DykF1
CTGTCACAAATAACAATAGGAGGCGAAC	GAACCCATCGTGTCCATTGTTCACCATT
>DykD8	>DykF2
CAATGGGCCTTGATCCATTGTTCACGAC	AACGCCCCGAGAACAATAGGGAACCATA
>DykD9	>DykF3
AGCAAAGTACAACAATGGAAGGCTCAAT	GCCCAGACAATAAAAGCCCTTTAGAGTC
>DykD10	>DykF4
ACCCCCGCAAGTATTGACAGTTGCCATG	AAAAAGACATAACAATAGAGCTCGGTTG
>DykD11	>DykF5
AACGCCCCGAGAACAATAGGGAACCATA	AGTACAATGAAAACAAAGAAACCCCCAA
>DykD12	>DykF6
ACGACAAAACATAAAAGCGTCACACATG	ACCCCCGCAAGTATTGACAGTTGCCATG
>DykE1	>DykF7
ACCCCCGCAAGTATTGACAGTTGCCATG	ACTCCCTTAAGGGCCCATTGTTCTCCCC
>DykE2	>DykF8
CAGACGCTTACACAAAGAAACGGTGAAG	TAACCCAAAAGAACAAAGGATACAATGG
>DykE3	>DykF10
CAACATCGCTAAACAATAGCCTAAAGTA	GTTCACACAGAACAATGGCCCCGACAAC
>DykE4	>DykF11
ATCCCCTCGAAACAATAGAAGCGACATC	CCGCACTGCAATACAAAGGAATACAGAT

>DykE5	>DykF12
ACAACATCAACGACAATGGAAGAACCAA	CAATGGGCCTTGATCCATTGTTCACGAC

Dfd				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PKRQRTAYTRHQILELEKEFHYNRYLTRRRRIEIAHTLVLSERQIKIWFQNRRMKWKKDNKLPNTKNVRKKT			
Selected sequences				
>dfd1	>dfd6	>dfd11	>dfd16	>dfd21
CTTCATTAAG	CCTAATTAAG	AGCTATTAAA	CTCATTACT	CGACTAATGA
>dfd2	>dfd7	>dfd12	>dfd17	>dfd22
GGTCATTAAT	GATAATTAAT	GCACTAATGA	CTTCATTAAG	TATCATTAAC
>dfd3	>dfd8	>dfd13	>dfd18	>dfd23
TATCATTAAA	CCTAATTAAG	TCGTAATGA	AGTCATTAGG	CCGTTAATGA
>dfd4	>dfd9	>dfd14	>dfd19	>dfd24
GGTCATTAAT	CCCCATTAAT	TGCTTAATGG	TACCTAATGA	CAATTAATGA
>dfd5	>dfd10	>dfd15	>dfd20	
GTCATTAACA	TTTTTAATGA	ATCGTAATTA	TGGATAATGA	

ems				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PKRIRTAFSPSQLLKLEHAFESNQYVVGAERKALAQNLNLSETQVKVWFQNRRTKHKRMQQEDEKGGEGGSQR			
Selected sequences				
>emsA1	>emsA6	>emsA11	>emsB5	>emsB10
TTAATTATA	GGTCATTACT	CCAATTATTG	TTTCTAATGA	CTAATTAGAG
>emsA2	>emsA7	>emsA12	>emsB6	>emsB11
CCATTTATGT	GTCCATTAAT	CCATAAATTA	TGCCTAATGA	CTAATTAGCG
>emsA3	>emsA8	>emsB1	>emsB7	
CATTTTATGA	TGTGATTAAC	TCTGGAGAGG	TCACTAATTA	
>emsA4	>emsA9	>emsB2	>emsB8	
GCCATGGACC	ACATAAATGA	GCCAATTATA	GGTCTAATGA	
>emsA5	>emsA10	>emsB3	>emsB9	
TTCACTAATA	ACTAATTAAA	CTCCATTAAA	CCAATTAGAG	

en	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	EKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKSTGSKNPLALQLMAQ
Selected sequences	

>eng1	>eng7	>eng12	>eng17	>eng22
CTAATTAGCG	GTGCTAATTA	CAATTAAAA	CGACTAATTA	CCAATTAAAC
>eng2	>eng8	>eng13	>eng18	>eng23
TATTTAATTA	TCAATTAACC	CCAATTAAAA	CCAATTAAAA	TCAATTAAG
>eng3	>eng9	>eng14	>eng19	>eng24
CTCATTAGTG	ACGTTAATTA	TAACTAATTA	CTCAATTAAG	CGGCTAATTA
>eng4	>eng10	>eng15	>eng20	
AGGGTAATTA	TAGGTAATTA	CTCTTAATTG	GCGTTAATGA	
>eng5	>eng11	>eng16	>eng21	
GCAATTATCA	CGGCTAATTA	CCGATAATTG	GCTAATTAAG	

eve				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	VRRYRTAFTRDQLGRLEKEFYKENYVSRPRRCELAAQLNLPESTIKVWFQNRRMKDKRQRIAVAWPYAAVYSD			
Selected sequences				
>Eve-G1	>Eve-G6	>Eve-G11	>Eve-H4	>Eve-H9
TCCGACATAA	CTTCTAACGA	ACACATTAAC	TGTTTAATGA	TTGCTAATGA
>Eve-G2	>Eve-G7	>Eve-G12	>Eve-H5	>Eve-H10
TTACTTAATT	CATCATTATA	CCTCATTATG	CGGCTAATTA	CCTCATTAAT
>Eve-G3	>Eve-G8	>Eve-H1	>Eve-H6	>Eve-H11
TCGATTATTA	GTCGTTAGTA	GTTAATTAAA	TTGCTAATTA	GGTCATTAAC
>Eve-G4	>Eve-G9	>Eve-H2	>Eve-H7	
CTTCTAATCA	TGGCTAATTG	TCCCATTAAC	ACACTAATTA	
>Eve-G5	>Eve-G10	>Eve-H3	>Eve-H8	
TAGTAAATTA	TCAATTAGAC	AGTCATTAAA	GTAATTAGTA	

fkh	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	GRSRVDKPTTYRRSYTHAKPPYSYISLITMAIQNNPTRMLTLSEIYQFIMDLFPFYRQNQQRWQNSIRHSLSFNDCFVKIPRTPDKPGKGSFWTLHPDSGNMFENGCYLRRQKRFKDEKKEAIRQLHKSPSHSSLEATSPGKKD
Selected Sequences	
>FkhA1	>Fkh2A3
CCAAGTTTGCTTTGTGCAAGAAGTTAAG	TGGTCGTTTACTAACAATAAATTATTTG
>FkhA3	>Fkh2A4
CGGCACAGTACACTTTATTTACTCAACG	ATTGTTTGTACAAAGCAACCCTGACGA
>FkhA4	>Fkh2A6
CCTATGCCGCCAAAAATGAAAACAGATT	AACTGAGGCCAGGTGTTTATCTATCAAC
>FkhA5	>Fkh2A7
CCACTTTAAAATGCAAATAGACTAAACA	AATGTTGACCTAAAGAAAACGATCAACA
>FkhA6	>Fkh2A8

CTTATGAGAAAACGAGATTAAATAAACA	AAAACCCACCCACAATATTTACGCAAGAC
>FkhA7	>Fkh2A9
GCAGGGCAAAACCTGTTTGTTGAAGAGC	GCAGGGCAAAACCTGTTTGTTGAAGAGC
>FkhA8	>Fkh2A10
ACAGATACAAGTCTAAATGGACTGTTTGC	TCAAGGACCCCATTGTTTACCTTAGATG
>FkhA9	>Fkh2A11
ACTTTTAGAAATGTAAATAAACAACACCG	AAATTGTTATATCCAAACATAAACAAA
>FkhA10	>Fkh2A12
GATTCTGGCAGAGATAACCCCCAGACGTG	GAGGCCCCCGCCAATATTTGACCAAAGG
>FkhA12	>Fkh2B1
CCGCAAACCCCCTTTGTTTACCTACTCT	CACATCTTTGAGATAAGTTTACACAGAGA
>FkhB2	>Fkh2B2
CTTTCGAACCCAAAGGCACGCGATCAAA	GGAATCCATAAATAAACAACAAGGATGA
>FkhB3	>Fkh2B5
GGCGCGGGGAATAATGAAGGACCTCCAA	GCAGAAGAATGACTATTTGCTCATCTAC
>FkhB6	>Fkh2B6
TGCAAAGAGACGTGTGTTTACCCAAAGC	AAGCGGCGTGCGGTGGAATCGGCCCAAA
>FkhB7	>Fkh2B7
GCAGGGATGCGTGTAAACACAAGCAAAA	TCGTCCCAACCTCGTGTTTGAATAATGA
>FkhB8	>Fkh2B8
CCGGAAATAATAGAGCCGCCCGTGTTTG	ACGGGGAACAAGGAAGTTTGTTTAAGTT
>FkhB9	>Fkh2B9
TGGACCCGCTCGGATGTTTGCCTAAGCT	CGGTGCCCTACGCAAACAATGAAGCTAC
>FkhB10	>Fkh2B10
CCCCCCGGCGGATTGTTTGAGTAAGGTG	AAGAGCTGCTGATCAAATAGTTCACTGA
>FkhB11	>Fkh2B11
GCATCCCTGGTGTTACCACAGTCACAAA	CAGCTACCGCGAACTGTTTGCACACAAC

ftz				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	SKRTRQTYTRYQTLELEKEFHFNRYITRRRRIDIANALSLSERQIKIWFQN RRMKSKKDRTLDSSPEHCGAG			
Selected sequences				
>FtzG1	>FtzG6	>FtzG11	>FtzH5	>FtzH10
ATCATAATTG	TTACTAATGA	CATCATTAAC	GAGTTAATGA	ATATTAATTA
>FtzG2	>FtzG7	>FtzG12	>FtzH6	>FtzH11
CAGCCGCCC	TTATAAATGA	TGCTTAATTA	GCCCTAAGAT	TTGTTAATGA
>FtzG3	>FtzG8	>FtzH1	>FtzH7	
TCCCATTAAC	TTTTTAATTG	GGCTTAATGG	CTGTTAATTA	
>FtzG4	>FtzG9	>FtzH2	>FtzH8	
CTCTTAATTA	CGCCTAATGA	GAACCTACTT	GGGTTAATTA	
>FtzG5	>FtzG10	>FtzH3	>FtzH9	
ATGCTCCCGC	TAGTTAATTA	CCGTTAATTA	TTTTTAATGA	

Gsc	
Promoter-Stringency	UV5m-10mM

Amino Acid Sequence	KRRHRTIFTEEQLEQLEATFDKTHYPDVVLREQLALKVDLKEERVEVWF KNRRAKWRKQKREEQERLRKLQEE			
Selected sequences				
>GscA1	>GscA6	>GscA11	>GscB5	>GscB10
CCGATTACGA	TCAGATTATC	TAGGATTATG	TCGGATTAAG	ATGGATTAGT
>GscA2	>GscA7	>GscA12	>GscB6	>GscB11
CAAGTAATCC	TAGGATTACT	ACAATAATCC	ACGGATTAAA	GGATTAATG
>GscA3	>GscA8	>GscB1	>GscB7	
AAGATTAGTC	TGCGATTAAG	TCGTTAATCT	CCCCTAATCC	
>GscA4	>GscA9	>GscB3	>GscB8	
TAGGATTATT	ATCGTAATCC	GGGATTAACA	GGGATTAACA	
>GscA5	>GscA10	>GscB4	>GscB9	
AAGATTAGTA	TCGTTAATCT	CGAGATTAAG	TGGATTAGGA	

gt	
Promoter-Stringency	lppC-5mM
Amino Acid Sequence	KRVLEQIRSSNGGSRTVTNPKMRRTNSRSGSVNEGSSSNNNSESEDRA AAEESSDCDSQAGNFESKSATSSSSNLANATAANSGISSGSQVKDAAY YERRRKNNAAAKKSRDRRRIKEDEIAIRAAYLERQNIELLCQIDALKV QLAAFTSAKVTTA
Selected Sequences	
>LGiantH1	>LGiantA7
GCATCACATAAATACGCCAAAACAGAAA	TCCCTTACGTAACAAAGTAACGAGGAAG
>LGiantH2	>LGiantA8
ATATTGCGTAACAGGGTAGTGCCACC	GGCGGACGGTCATTACGTCATCACGATG
>LGiantH3	>LGiantA9
TAAACAAAGATAGAAGCGTTATATTAAA	TATAACTATGACGTAACCGCAGATAGCT
>LGiantH4	>LGiantA10
CGAAGCATTATGTAACACGCGTAACAAC	TCATTGCACACGGTTATGTAATACCACT
>LGiantH5	>LGiantA11
GCCAATACGCCGGAGCTTTACAGTGGTT	AAAAAATTACGCAACACCGTTTTGGTAA
>LGiantH6	>LGiantA12
ATTACGTAACTTAAGGGCAAGTTACTCT	ATTACGTAACTTAAGGGCAAGTTACTCT
>LGiantH7	>LGiantB1
AAGGGATATTGCGTAACCGTAATCCTTC	CCAACGTCCGTGTTGCGTAACTTCGAAC
>LGiantH8	>LGiantB2
CAAAGCGCCATTACGCAATCACGCGAAC	TGGCGATTACGTAAGTCAAGACGAGGTT
>LGiantH9	>LGiantB3
GCTCACAACAGCGTTACGTAACATACAG	ATACCCGCCATTACGTAATCGCCACTCC
>LGiantH10	>LGiantB4
ATGTAGATTACGTAACAGGGGACCATGA	TCTTCATGTTACGTAAGATAACCCAGGT
>LGiantH11	>LGiantB5
CATCAAGTGTTACGCAATCGGGACCAAC	GTGCGGTATTACATAACTTCACCATTTG
>LGiantH12	>LGiantB6
TCATCGAAACCTCATTACGTAATAGCAT	GGCTCTTGCATATTACGTAAGCACTATA

>LGiantA1	>LGiantB7
TTCCCTCCGTACCTTACGTAACGCAACA	TAGGCACCAAACATTACGTAATACAAGG
>LGiantA2	>LGiantB8
TAGGCTACAACTAATGAAAATAAAAAAA	TATAAGTGTGACGTAATACACGAGATAC
>LGiantA3	>LGiantB9
TAGGCTACAACTAATGAAAATAAAAAAA	GGAATTACGTAACAACATAGTGATCTCT
>LGiantA4	>LGiantB10
TAGGTTATTACATAACAGGATCATAGCT	GTGCGGTATTACATAACTTCACCATTTG
>LGiantA5	>LGiantB11
GACAATAGGGTTCAGTCCATACTCGCAAA	CCAATTACGTAACAAGTACTACAGGTAG
>LGiantA6	
GGACGGAAGCATAATGTGTTACGTAACT	

<table>
<tr><td colspan="2">h</td></tr>
<tr><td>Promoter-Stringency</td><td>lppC-5mM</td></tr>
<tr><td>Amino Acid Sequence</td><td>VTGVTAANMTNVLGTAVVPAQLKETPLKSDRRSNKPIMEKRRRARINNCLNELKTLILDATKKDPARHSKLEKADILEKTVKHLQELQRQQAAMQQAADPKIVNKFKAGFAD</td></tr>
<tr><td colspan="2"></td></tr>
<tr><td colspan="2">Selected Sequences</td></tr>
<tr><td>>LHairyA1</td><td>>LHairyB11</td></tr>
<tr><td>ACGCGCCACATCAAACTCTCCAATTGAA</td><td>CCCCACAGTCTTTGGCACGTGCCAGATC</td></tr>
<tr><td>>LHairyA2</td><td>>LHairy2E1</td></tr>
<tr><td>GACAATAGGGTTCAGTCCATACTCGCAAA</td><td>GAAGCGCGCCAGCGAAAAAAGATCACCAC</td></tr>
<tr><td>>LHairyA3</td><td>>LHairy2E2</td></tr>
<tr><td>CCCCGTCCTGACACGCGCCGGACCCAAG</td><td>ACGCGCCACACGCAGGGACTTTAGAGGG</td></tr>
<tr><td>>LHairyA4</td><td>>LHairy2E3</td></tr>
<tr><td>TGATCAGCCACGTGGCCAAAACCCCACG</td><td>GAGTCATAAAGCTGCCCGCGAGCCCTGT</td></tr>
<tr><td>>LHairyA5</td><td>>LHairy2E4</td></tr>
<tr><td>ACGTGTGAACTTGAAAATTAATTCATTG</td><td>ACCTTGCCCCCACGCGCAGACTACAAAA</td></tr>
<tr><td>>LHairyA6</td><td>>LHairy2E5</td></tr>
<tr><td>ACGTGCCAGGAGGGCCCCTTGAACCAAT</td><td>GACCCAACGCGAAGCAATTGGCACGTGCC</td></tr>
<tr><td>>LHairyA7</td><td>>LHairy2E6</td></tr>
<tr><td>GAAAAAGCAGAATCGTAATTACGCAAGT</td><td>ACGTGTCACTGGGATGATAATCTCCACA</td></tr>
<tr><td>>LHairyA8</td><td>>LHairy2E7</td></tr>
<tr><td>ACGCGCCACATCAAACTCTCCAATTGAA</td><td>ACGCGCCATACATTCGCAAAATAAAACG</td></tr>
<tr><td>>LHairyA9</td><td>>LHairy2E8</td></tr>
<tr><td>ACGCGACCCCCCAACTACCATGCTACGGC</td><td>GGAACTAAATAAATGGCACGTGCCCATT</td></tr>
<tr><td>>LHairyA10</td><td>>LHairy2E9</td></tr>
<tr><td>CGGCTTGCCACGCGCCAACGCCCGATGC</td><td>GGCACCGGCACGCGACGTTTTACCCAGC</td></tr>
<tr><td>>LHairyA11</td><td>>LHairy2E11</td></tr>
<tr><td>ACGCGACCCGTACAACAAATTTCACCAG</td><td>TCTAGCAGCGGTCTAGCCACGCGACCAC</td></tr>
<tr><td>>LHairyA12</td><td>>LHairy2F1</td></tr>
<tr><td>GGGACACCCGGGCAACCGCAAATGATAA</td><td>GGGAAGACACGCGACACCAAAACGAATC</td></tr>
<tr><td>>LHairyB1</td><td>>LHairy2F2</td></tr>
<tr><td>GAATATCGGCACGTGGCCCTAGAGACCA</td><td>ACGCGCCACATCAAACTCTCCAATTGAA</td></tr>
<tr><td>>LHairyB2</td><td>>LHairy2F3</td></tr>
<tr><td>ACGCGCCACATCAAACTCTCCAATTGAA</td><td>CTGACCGGCACGCGACATTCGCCACACT</td></tr>
<tr><td>>LHairyB3</td><td>>LHairy2F4</td></tr>
<tr><td>TGTCTACACGCCCTGTTCTAGCTACGCC</td><td>CACGTGCCAGATATTACCGCCCACCAGA</td></tr>
<tr><td>>LHairyB4</td><td>>LHairy2F5</td></tr>
</table>

AGAAATGGCACACGCCATGAGTGCACTA	ATATTCGCCACGCGACCACCGACCCCAC
>LHairyB5	>LHairy2F6
CACGTGGCATACCGGTAGAAAAGCCGAC	TGAATGAGGACACGTGCCGTGAAAACGC
>LHairyB6	>LHairy2F7
AATAGCCACGCGCCAGACGTATATTCAC	GATCGAAAAACTACAAAATTAATTAACA
>LHairyB8	>LHairy2F9
AGCGTGAAGGCACGTGCCACCGACAGCC	CTATGGCACGTGCGCCCCCTCAGCCGAA
>LHairyB9	>LHairy2F10
CAAGGACACGCGCAACGGTGCCCGGACA	CTGACCGGCACGCGACATTCGCCACACT
>LHairyB10	>LHairy2F11
CACGTCAGGCACGCGACCCCTTCTCGCA	GGAGCACGCAGTGTGGCATGGCACGTGCC

hb	
Promoter-Stringency	UV5-5mM
Amino Acid Sequence	KMKNYKCKTCGVVAITKVDFWAHTRTHMKPDKILQCPKCPFVTEFKHHLEYHIRKHHKNQKPFQCDKCSYTCVNKSMLNSHRKSHSSVYQYRCADCDYATKYCHSFKLHLRKYGHKPGMVLD
Selected Sequences	
>5HbG1	>5HbH7
AACAATACAAGTAATAAAAAAATAAGGA	TTCAATCTCCGTGAAAAACGTTCATGTG
>5HbG2	>5HbH8
CCAGCTCAAACCTTTTTCCCTTTGCCAC	GTTGCCAAAAAAAAAACGTCATCACAAA
>5HbG3	>5HbH9
ATGGATAGCGCAAACGAAAAAAAAAGTG	GAAGTATCTCCGAACCCAAAAAAACACG
>5HbG4	>5HbH10
TGAAGAAGAACACTGAAGAAAAAAATCG	TAAAAGATAAATCCCACAAAAAACCAGA
>5HbG6	>5HbH11
CGTGACGTCAGTGTGAAAAAAAAGGTCA	GCTGACCCAATTCACACAAAAAACGAAC
>5HbG7	>5HbC1
CCGTCCCACAGATGATAAAAAAAACTTC	AGGAGCCCACAACCCAACATTAACTCAC
>5HbG8	>5HbC2
CGATAGTGTCCCTATCAAAAAAACATTT	ACCGGATAAATACCCGTACCAACCATGC
>5HbG9	>5HbC3
ACATGAAAAGCAAAAAAACGAGT	CACACGACGAAAGTGCACAAAAAAATTC
>5HbG10	>5HbC4
CCGTTATGACCGCGATCAAAAAAACCAT	GCATATCCCCGTGGCTATGAAGCAAACT
>5HbG11	>5HbC5
TACCGGAGCGATACACAAAAAAACATGC	ACTAACACATCGCAACGCAAAAAACGCA
>5HbG12	>5HbC6
TGAATTTGGAGTGGAGTAAAAAAACGCT	AATGCCAAAAAGAGCAAAAAAACACCAA
>5HbH1	>5HbC7
ACATCAAGCGAGATCCACAAAAAACTAG	AATATGAAAGAACAAGCAAAAAAATAGC
>5HbH2	>5HbC8
CCCACGATCGTCTACAACAAAAAACACA	ATATAAAAAATTAAACATAAAAAAATAC
>5HbH3	>5HbC9
CACCCTGTCGCAACCCCAAAAAAACATC	CACAGCTACTCACCCAAAAAAAACACAT
>5HbH4	>5HbC10
GAGGGCTCACCGCTGCAAAAAAACACCC	ACTAGTCCAGAGAGAACAGCATTCTGGC
>5HbH5	>5HbC11
GCAATGAACATTGCGTCCCCAATAAATC	CAACCCGCCAGCATCAAAAAAAACAGGC

>5HbH6	>5HbC12
ACATCAAGGCTGAAGGCCCCTGGACGTC	GCACGCGCCATAGTCAAAAAAACCACAA

hkb	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	KFKCPNCDVAFSNNGQLKGHIRIHTGERPFKCDVNTCGKTFTRNEELTRHKRIHTGLRPYPCSACGKKFGRRDHLKKHMKTHMPQE
** alternative linker: AAADYKDDDDKFRTGSTSLYKKAGS	
Selected Sequences	
>Hkb2B1	>HKB5G1
TTGATCAACTGCAGTGCGAACCGGATCCC	CCACCAGCTTCCAAGTAGGCAACCCAGAC
>Hkb2B2	>HKB5G2
ATGACATTTGACTCAGCATCTACCCCCC	CAAGAACCGATTACTGACAACATCGAAA
>Hkb2B3	>HKB5G3
ACAGACAACACAGCAGGCTTGGGCGTGAC	ACATACCAGCCTGTAATCACGCCTATGT
>Hkb2B4	>HKB5G4
CAGCCGCCGCCCATAGACTGCCGGCAA	AGGAGCCCACAACCCAACATTAACTCAC
>Hkb2B6	>HKB5G5
AGACCCCTCCAACAGGAAGAAGGCGTGA	GACTCGGGCCGTAGACACGCCCACCACG
>Hkb2B7	>HKB5G6
ACCTCCGTCGGTGAGCATCACGCCTACC	CACACTTAATCACTACTCACGCCCCTCG
>Hkb2B8	>HKB5G7
ACAGGGGGCGTGAGTACGACTACCCACC	TCAACACACTGATGCGTATCACGCCTCT
>Hkb2B9	>HKB5G8
CTATAACTTACCCCCCCCTCACGCCCCC	AGTTTCGTACATATTCCCGCCCCTCAAT
>Hkb2B10	>HKB5G9
TGATGGTGGGTGCTACACTCACGCCCCC	TACGAATGATGTTAGTAGGGGCGTGATG
>Hkb2B11	>HKB5G10
GATCTAAGCCATAAAAAGGGGCGTGAAC	GCGACCGCGAAGCAAGGGGGCGTGACGCG
>HKBB2	>HKB5G12
TCACATCACGCCCCCTCCCGACCCACCC	CCCAGCAAATCCCTGATGTCACGCCCCCCG
>HKBB3	>HKB10H2
AAGGATCGGGTGAACGCCCCTTTACCGG	TACAAAATAGAGCAACCCGCCCCCCATC
>HKBB4	>HKB10H3
GACTGCCACAGTGCTCCGATAGGCGTGA	ACGACAGGGAGGAGCGTGGGCGTGCACT
>HKBB5	>HKB10H4
CCCACAGATAACCTCGCTCCTCACGCCTA	CACCCCGGTCAAACAGCACGCCCCCCAC
>HKBB6	>HKB10H5
GTGAATTACAGCAAATACACCTAGCATT	CTAACGCGTCGACACACACGCCCCCTTC
>HKBB7	>HKB10H6
TGCGTTAACAACGGCAAATCACGCCTTC	ACTTGGGAACCAAAATATCACGCCCAGT
>HKBB8	>HKB10H7
GCGCTTGCATCCCCCCACGCCCACATAA	TTCTTTCCTGATTCCGTGAAAGGCGTGA
>HKBB10	>HKB10H8
ACGGGAACAACCCTAGAGGGGCGTGAGG	GGCCTGGAGTGGGCGGGGAGAACACAAC
>HKBB11	>HKB10H12
GCAAGTGATCTGCTAGCCCTCACGCCCCC	AGGAGTAGCCATGACGTGGGCGTGAACC
>HKBB12	
CTGCGCGAACCCCCCCCTCACGCCCCCT	

Inv

Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	EDKRPRTAFSGTQLARLKHEFNENRYLTEKRRQQLSGELGLNEAQIKIWF QNKRAKLKKSSGTKNPLALQLMAQ			
Selected sequences				
>InvC1	>InvC7	>InvC12	>InvD5	>InvD10
ATAATTAACC	GGTAATTATA	CCTAATTAAA	GTAATTAGTA	TCACGCCGAG
>InvC2	>InvC8	>InvD1	>InvD6	>InvD11
ACTAATTAAT	TCTAATTAAA	CCAATTAAAT	GATGCTAAAC	GGTAATTAAC
>InvC3	>InvC9	>InvD2	>InvD7	
ATAATTAGCA	AGCCCTCGCA	TCAATTAGAG	CCAATTAGTT	
>InvC5	>InvC10	>InvD3	>InvD8	
CCACTAATTA	AGTCAGCATG	TCAATTAAAA	TCAATTAAAA	
>InvC6	>InvC11	>InvD4	>InvD9	
CTTCACTGAA	TAATTAGAG	CTAATTAGAA	ATTCCGCTCT	

kni	
Promoter-Stringency	UV5-10mM
Amino Acid Sequence	NQTCKVCGEPAAGFHFGAFTCEGCKSFFGRSYNNISTISECKNEGKCII DKKNRTTCKACRLRKCYNVGMSKGGSRYGRRSNWFKIHCLLQEHEQ AAAAAG
Selected Sequences	
>KnirpsE1	>Knirps7
AGAGTAAAACCCTTGCACCCGGAGATC	GCCAAAGGAGAAATTAGAGCAGACTAAG
>KnirpsE2	>Knirps2F1
AACATCGGCGGAATGAGAGCAACGAATA	GTCTCACCGGAATTAGAGCACGCAGAAC
>KnirpsE3	>Knirps2F2
TAGAAGAGCTACATGTAGGTCACCACACG	TTCAACCTACCAAATATAGGCCACCTAA
>KnirpsE4	>Knirps2F3
AAACCGCCATTGCACCAGTTTCTAGCAG	AAACTGGAGCAGATACGACTATGACGCG
>KnirpsE6	>Knirps2F4
CCTAACGTGTCAAAAGTAGAACACAAGC	AGATCGTCCAAATCAGGGCACCCGCCCA
>KnirpsE7	>Knirps2F5
CCACGGGGAAACATGCTCCAGATAAATC	GACGGCTGCCGACTAGACACGACTAGCA
>KnirpsE8	>Knirps2F6
AACACGGAGAAAATTTAGAGCGGCGACG	GGCGCGTTTGAACAGGACTATAGACCAC
>KnirpsE10	>Knirps2F7
GCATGGCGGTTAAACCAGGTCAATAAAA	GTGCCTAGCAATGTGCCCTAGATACTAC
>KnirpsF1	>Knirps2F8
GTATGCCCGAAAAATAGAGCACTAGGGGA	AGCCAGTGCCATTAAACCAGGTCACGCAC
>KnirpsF3	>Knirps2F9
AATTGCAACCGAAAGTAGAGCAGGAATA	TAGAAAGCAGACAAAGTAGAGCACGATT
>KnirpsF4	>Knirps2F10
GCCCAGCCAGAAAAATAGAGCAGTACAC	TAAATTAAGCAATCCAGGGCAAGGTGAA
>Knirps1	>Knirps2F11
TCCTTTCTACTTGCTCCATATAAAATCA	GGCACGCTCAGTAAGTAGAGCAGACTAT

>Knirps3	>Knirps2F12
CAGCCATTTCGATGGTCTAGTTTTAAGA	ACTACTTAAAATCTAGAGCAGTTTGAAC
>Knirps5	>Knirps2F12
AGCCCGTCCGAATGTGGAGCACAAAACT	ACTACTTAAAATCTAGAGCAGTTTGAAC
>Knirps6	
CTGCGACGGTAAATTAGGTCACGTAATC	

Kr	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	KSFTCKICSRSFGYKHVLQNHERTHTGEKPFECPECHKRFTRDHHLKTHMRLHTGEKPYHCSHCDRQFVQVANLRRHLRVHTGERPYTCEICDGKFSDSNQLKSHMLVHNGEK
** alternative linker: AAADYKDDDDKFRTGSTSLYKKAGS	
Selected Sequences	
>KrE1	>KrF7
GACAATAGGGTTCAGTCCATACTCGCAAA	ATAAAAACTAGGCCGCTAACGAGTTAAC
>KrE2	>KrF8
CAGAACGCAACAAGTCGAAAGAACAGCC	AAATCGACGGCTATCCAAAAGGGGTAGA
>KrE3	>KrF9
GCAAGCTACATACTACGAGAAGAGACAG	CCATTCGAAAGGGTGAAGACAGAACAGA
>KrE4	>KrF10
AAGAACACAGAAGAGGATCAAAGGGTGT	GCTCAGAGCAAATGCACCAAAGGGGTTT
>KrE5	>KrF11
CGCTCTATCAAAAGGATTAGTTTAAATC	CAACTCTAGCAAGGGGTAAAGATACAAC
>KrE6	>KrD1
GCTCGCACGCCACAAGAAAGGATTCACG	CAACTTAACGGGTGAACCACTAGCAAGA
>KrE7	>KrD2
GCAACTTACGAAAGGGTAAGCAACCTCT	GTACCCCCCTCCGTGAGGTTAAAGGGTA
>KrE8	>KrD3
CGTACGAACGGGTTCAGCCCGTGAGCGG	GACATTCAAAAGGGTAATGTGGTCATTGC
>KrE9	>KrD4
CAACGTCTGAAGGGGTAAACGGAGCTAG	GGACCCAACCCGCCCACAGGAAGGGGTA
>KrE10	>KrD5
TCTACTGGCTCAAGAACGAACGAGTTAT	GGCACGAACAGAAATAAAGGGGGTAAGC
>KrE12	>KrD6
AAACCTAACCCTTCAACCCACTCTCCAA	GGAAAAGGGTGAAAAAGCTCATTCAATC
>KrF1	>KrD7
TGATAACACCGACATAACGAACGGGTTT	CAACCCTTCCGTCCCTAACCCTACCAGA
>KrF2	>KrD8
GCCTTTCGCATAGTACCAAAAGGGCTAG	TGCCCACCAAGACATGTAACCCCTTACCC
>KrF3	>KrD9
GTGGCCGCCAGGTGGCCAACGGGGTAAC	GCCTCCCTCACCTTTCCGAAAAGGGGTA
>KrF4	>KrD10
CACGGCGCCAATACCTCATCCAGTTAAT	TCAGATGCCGGTGCCGCAAAGGGTAACA
>KrF5	>KrD11
CAGCACCACGTACCCAATAAAGGGGTTC	TACGCAAAGGGGTTGGAGAAACACTAAT
>KrF6	>KrD12
ACAATCATAGACCACACAACCCTTCCAG	AAGTCATGAAGGGTTAAAGAACAATGAC

lab	
Promoter-Stringency	UV5m-10mM

Amino Acid Sequence	NNSGRTNFTNKQLTELEKEFHFNRYLTRARRIEIANTLQLNETQVKIWFQ NRRMKQKKRVKEGLIPADILTQH			
Selected sequences				
>LabE1	>LabE8	>LabF4	>LabF11	
TCATTAACGA	TAGTTAATTA	CTTCACTGAA	AGTCTAATGA	>LabH10
>LabE2	>LabE9	>LabF5	>LabH3	TATCGCCCAC
GCCTTAATTA	ATACTAATTA	CTGTTAATTA	GCTGTTATTT	>LabH11
>LabE3	>LabE10	>LabF6	>LabH5	AATGATCGTC
TGGCTAATTA	CGTTCTTTAA	AATGATCGTC	GTGCGCGCAG	>LabH12
>LabE4	>LabE11	>LabF7	>LabH6	TACATAATGA
GATAATTAAT	GCTTGATGCG	GTCCAGATTG	GTGTTAATTA	
>LabE5	>LabF1	>LabF8	>LabH7	
CATACCCAGA	AGTCATTAAG	CGTTAATT	GTCTTAATTA	
>LabE6	>LabF2	>LabF9	>LabH8	
AATTTAATTA	CTACCAGATT	GGTCATTAAT	TCACGCCGAG	
>LabE7	>LabF3	>LabF10	>LabH9	
GCTAATTAAT	GTAGCCAATG	CAGGCACCCA	CTACTAAATT	

nub	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	RAAGEPSPEETTDLEELEQFAKTFKQRRIKLGFTQGDVGLAMGKLYGN DFSQTTISRFEALNLSFKNMCKLKPLLQKWLDDADRTIQATGGVFDPA ALQATVSTPEIIGRRRKKRTSIETTIRGALEKAFLANQKPTSEEITQLAD RLSMEKEVVRVWFCNRRQKEKRINPSLDS
Selected Sequences	
>2NubG1	>2NubC4
TGCATGGGTCAAAATCTGCATAAACAAT	AATATGCAAAACAGAGCTAAACGCGGGG
>2NubG2	>2NubC5
ACAGGATGCAAATTAGTCGACAGCCCTA	ATATTCAAATTAGAACGGACATACCCCC
>2NubG3	>2NubC6
ACCCTTGACGCACCCATGCAAATGAGGG	GACCAACCCTCATTACCATATCCCATCC
>2NubG4	>2NubC7
GATTGGATCACAATTAACATATAACCCT	GAGCTCCGAGTATTATGTAAATACGTCT
>2NubG5	>2NubC8
GTAGCTTAATTATGCAATACATAGTGC	GCTTATGCAAATACAAACCCCTCTCAAG
>2NubG6	>2NubC9
TTACTGGTAACCAAATTCAAATCAAAAA	AAAGATGCATATGCTAATTAGCACTACG
>2NubG7	>2NubC10
GTGAATTACAGCAAATACACCTAGCATT	TCATATTTAAATGAGTTTAGGCCACAAA
>2NubG8	>2NubC11
GCAATGTAATGATATGCAAGGTGACCGC	ATTATGCAAATACGGTTAACCGTTCTGA
>2NubG9	>2NubC12
TTAATGTTCAAATTTACATAATGCCTTA	ATACTGACACGAATGCAAATCAGGATAC
>2NubG10	>2NubD1

ACTCGCCACCGCATATGTAAAACAGATAC	CACACCCCCAGTATGCTAATGTGAGATA
>2NubG11	>2NubD4
ACGATATGCAAATGAGGCTCCCCACATA	ATTAGTTCATTAATATTCATCCAAATCC
>2NubG12	>2NubD5
GGGTGCTCATTACGTATGTAAACACTCC	ACAAGCACAAATATGCAAATGATGGCTT
>2NubH1	>2NubD6
CCGGTTATACGTATGCAAATGCCGTAGA	TGCGCGTTCAAAATATGTTAATGACTAA
>2NubH5	>2NubD9
AAGACATCCATTATGCAAATAATGGTTA	TACATTTAATTTACATATAGTAGCATCA
>2NubC1	>2NubD10
TTATTTAAATATTAGAGTTTCCTAAATA	GAATCTCTCAATATGCAAATTAACTTC
>2NubC2	>2NubD11
CGATATGCAAAATACCCGGAACCCACTA	TCGAAACCCCGTATGCAAATTAGCTTTA

oc				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	QRRERTTFTRAQLDVLEALFGKTRYPDIFMREEVALKINLPESRVQVWF KNRRAKCRQQLQQQQSNSLSSSK			
Selected sequences				
>OcA1	>OcA6	>OcA11	>OcB5	>OcB10
CCACTAATC	TATATAATCC	GCAGATTAAC	CTGGATTAAG	TCGGATTAAG
>OcA2	>OcA7	>OcB1	>OcB6	>OcB11
AGCTTAAGCC	TTTGCTAATC	GTGGATTAAT	AGGGATTATA	CCGGATTAAC
>OcA3	>OcA8	>OcB2	>OcB7	
CGATAATCCC	CATTAATAAC	TTCATAATCC	GAGTTAATCC	
>OcA4	>OcA9	>OcB3	>OcB8	
GGGGCTTAAA	CGCGGATTAG	GAGGATTACG	CTGGATTAGT	
>OcA5	>OcA10	>OcB4	>OcB9	
GAGGATTATT	AGGATTAAGG	AGCGATTAAG	AGGATTAAT	

odd	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	SSRPKKQFICKYCNRQFTKSYNLLIHERTHTDERPYSCDICGKAFRRQD HLRDHRYIHSKDKPFKCSDCGKGFCQSRTLAVHKVTHLEEGPHKCPIC QRSFNQRANLKSHLQSHSEQS
** alternative linker: AAADYKDDDDKFRTGSTSLYKKAGS	
Selected Sequences	
>Odd1	>OddB1
TCACCGTACGCACCTGGTCATCGAAGAC	ACGCCTTCTACAGTAGCAAGATGTCGCA
>Odd3	>OddB3
ATCAAATTACAGTAGCACTAGACACGCG	CACTGCAATACAGTAGCAAACCAGTTTC
>Odd4	>OddB5
CGCGCTCTACCGGTAGCACTAGTAATAT	GCGCTGCAACCAGTAGCCGTAATGCGAC
>OddA2	>OddB6
GGCGCGTACTGGAAGCAACACTTGACCC	CCTATTCACACAGTAGCACGAATCCTCA
>OddA4	>OddB7

ATTTCGTAAACGGTAGCAGTTTTGCGCG	GCCCCCCGCACAGTAGCAACGTTGGACA
>OddA5	>OddB8
AAGACTTCCCGGTAGCAGTTGCTGCGAT	TGATGCCACACAGTAGCGGAAGAGATTA
>OddA6	>OddB9
TACAAAAAACAGTAGCAGCAACAGGAGC	GCTGCACGAGTAGCCAAAGTCAGACACA
>OddA8	>OddB10
CATAAAAACCAGTAGCAGGTCCAAATAA	GAATTTCCGCGCGCTAGACAGTCTCACG
>OddA11	>OddB11
CCTCATGCACAGTAGCACGCAAGGCGAG	GATTTCCACCCCCATTGCAAAGACTCGA
>OddA12	>OddB12
TGTTAATTGAGTAGTGGCATCTTGCACC	AGACGGGAACAGTAGCCACGAGAACGCA

opa	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	QCLWIDPDQPGLVPPGGRKTCNKVFHSMHEIVTHLTVEHVGGPECTTHACFWVGCSRNGRPFKAKYKLVNHIRVHTGEKPFACPHPGCGKVFARSENLKIHKRTHTGEKPFKCEHEGCDRRFANSSDRKKHSHVHTSDKPYNCRINGCDKSYTHPSSLRKHMKVHGNVDEKSPSH
Selected Sequences	
>Opa2G9	>Opa2H11
CCCCCGCCGTGAATTTCAGATAGTAGGCT	GATCCAGACTCGGTCCGACCCCCCGCTG
>Opa2G10	>OPAC2
CCATTCACCTCCCGCAGAAGCAGCACACC	CAAAAGACTGCTGTAGGCCCCCCCATGG
>Opa2G11	>OPAC3
AGAATCTTACCGCCCCGACCCCCCGCCA	TATTGTGCGAGTCCCAGCTCCCCTACAG
>Opa2G12	>OPAC4
AACCTGGACCCGCTATGCCCCCCCACGA	TACAACCTCACCACAGCAGGGAGCCCTC
>Opa2H1	>OPAC5
GCTTCCCCCCCGCTGCGCAGGAACCATT	CTCTCCTCCCGATTCGCCGCCCTGCTG
>Opa2H2	>OPAC6
GCAACTATAAGCCAAGCCCCCCCGCTGG	AGCACAACGACTAAGTTGGCACCCGCTG
>Opa2H3	>OPAC7
TAAACTATACCCCCTGCTGGCACCCTCA	GAAGCACCCGCCCAGAGAACCCCCGGTG
>Opa2H4	>OPAC8
GCACCGGCCCAGTACGGACCGCCCGTTG	GTGGACCAGGCTAAGACACCCCGCGGAG
>Opa2H5	>OPAC9
GCCTCCAGTAGCCCCTGCCCCCCCGCTG	GAAGGGAAACTAGTCCACAGGTACACAA
>Opa2H6	>OPAC12
GCGTATACGGCGCAGAAGACCCCCCTGGGG	CCCGCTACGAACCGAAGACCCCCCGCTG
>Opa2H10	
GCCACAGAGGGGAGGGGACCCCCCACAG	

Optix				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	GEQKTHCFKERTRSLLREWYLQDPYPNPTKKRELAKATGLNPTQVGNWFKNRRQRDRAAAAKNRIQHSQNSSG			
Selected sequences				
>11617C2	>11617D1	>OptixE2	>Optix2A2	>Optix2B2

GAACCTACTT	GTAGTGCTAG	GTAGCCAATG	GCCCATGATA	TTTCTGCGTG
>11617C3	>11617D2	>OptixE3	>Optix2A3	>Optix2B3
GCGCATGAGA	CCGTCTAAAC	CGTTCTTTAA	ACATGTGATA	CATTGCGATA
>11617C4	>11617D3	>OptixE4	>Optix2A4	>Optix2B4
TGAAGTGATA	TAATGCACAC	CCCTAACATG	GGAGCTGATA	GTGATTGATA
>11617C5	>11617D5	>OptixE5	>Optix2A5	>Optix2B5
CAACGTATT	CTTTTCATCT	GAACCTACTT	CAATCTGATA	CCAAGTGATA
>11617C6	>11617D7	>OptixE6	>Optix2A6	>Optix2B6
CAATGTGATA	AGAAACTATG	ATCTTAATTAC	GTAGCTGATA	ATAAGTGATA
>11617C7	>11617D8	>OptixE7	>Optix2A7	>Optix2B7
ACCAGTGATA	ATAAATGATA	GTGCGTACTG	TACCCACGCC	CGTTATGATA
>11617C8	>11617D9	>OptixE8	>Optix2A9	>Optix2B8
GATTGCGATA	GTTAGTGATA	GGAAGTGATA	GAAGTGATAG	CTTTCTGATA
>11617C9	>11617D10	>OptixE9	>Optix2A10	>Optix2B9
CATCGCTATG	ATAAGTGATA	ATCTTATTAC	AACCGCGATA	ATTCAAACA
>11617C10	>11617D11	>OptixE10	>Optix2A11	
ACGTATTGGT	CGTAGTGATA	CTACTAAATT	GTGATTGATA	
>11617C11	>11617D12	>OptixE11	>Optix2A12	
CGTAGTGATA	AGAGATGATA	ATCAGTCCTT	TCCCTTGATA	
>11617C12	>OptixE1	>OptixE12	>Optix2B1	
ATCTTATTAC	TCTTCCATTA	CTACCAGATC	ATTCGCGATA	

pb				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIAASLDLTERQVKVWFQ NRRMKHKRQTLSKTDDEDNKDS			
Selected sequences				
>Pb1	>pBG3	>pBG8	>pBH3	>pBH8
GGTCATTAGA	CTCATTAAA	TGACTAATGA	GGTAATTATA	AGGTTAATGA
>Pb2	>pBG4	>pBG9	>pBH4	>pBH9
GGTAATTAAC	GGTAATTATA	GCGTTAATGA	GATAATTATC	TCATTAATGA
>Pb3	>pBG5	>pBG11	>pBH5	>pBH10
TGTAATTAAA	CTGTTAATTA	GCTCATTAAG	CCAGCAAGAT	CCTCATTAGA
>Pb4	>pBG6	>pBG12	>pBH6	>pBH11
GGTCATTAAC	CGTAATTAAT	GCTAATTAAT	TTGCTAATGA	GGTAATTAGA
>pBG2	>pBG7	>pBH1	>pBH7	>pBH12
ATCCTAATTA	TACATAATGA	TGTAATTAAA	GCTAATTAAG	TTGCTAATTA

prd

Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	NSGQGRVNQLGGVFINGRPLPNNIRLKIVEMAADGIRPCVISRQLRVSHGCVSKILNRYQETGSIRPGVIGGSKPRIATPEIENRIEEYKRSSPGMFSWEIREKLIREGVCDRSTAPSVSAISRLVRGRDAPLDNDMSSASGSPAGDGTKASSSCGSDVSGGHHNNGKPSDEDISDCESEPGIALKRKQRRCRTTFSASQLDELERAFERTQYPDIYTREELAQRTNLTEARIQVWFSNRRARLRKQHTSVSGGAPGGAAASVSH
Selected Sequences	
>1prd2A12	>21PrdD6
GACTCGAATAACAATTGGTCACGCTTCG	AGAGTAAAACCCTTGCACCCGGAGATC
>2PrdD5	>22PrdD7
ACTACGTAAACAATACAGTCACGCCTGT	TTTCACAACTAATTAACCACGCCACCAA
>3prd2A8	>23PrdD8
TACAAGGACTAAGCCGACACGGTAGGGA	CGAAACCTATTAGCGTCACGCCCCCCAC
>5PrdD2	>24PrdD9
GCGCCCACGAGGCAGTCCGACACGCTCA	AAACCGAGTTCCGATCCGTCACGGCTTT
>6prd8	>25prd2A2
CACGACGTCACGGTAGGAAATCCCGTCC	TCACCGCGGAGGTCTCGTTCCGGTCTGT
>7prd2A3	>26prd2A5
GATTTCGTCACGCTCCGTAACGATCCTG	ATTAGGGATTAATTCCGTTACGGCCACC
>8prd2A7	>27prd2A10
TCTGCTCTGAACCACACGGTCACGGTGA	ATTAGTCACGCTCTCAATTAACTCATGC
>9prd13	>29prd6
GTTAATTACCGTGACGCGCTTGAGAGAT	CTAAAAAGCAATCCTCCGACACGCCCCA
>10PrdD4	>30prd7
CCCGGACCCGCTAAACGTTTCACGGTTC	CAATTAGGCACGGGAGGCTGAGACATAA
>11prd17	>31prd9
CACCTGACAGCGTCCCGTCACGCTGCCCC	ATAATTAGAAAACCAGATGAAACCGTGA
>12prd2	>32prd10
CTGATCCTCAACCCCTCACGGTGAGCA	TGCGAGCACATAATACTTTCACGCATGA
>13prd15	>33prd11
GGTCTGGAATATCGGCATTCACGCTTGA	CAAAGTGTCTTCATTGCCGGTAAGCATA
>14prd12	>34prd14
ACCCCCGCAAGTATTGACAGTTGCCATG	GGCGCGTTTGAACAGGACTATAGACCAC
>15prd16	>35prd19
GTACCTGACCCGCATTTCACGGTGGGCC	GCAGCCCAACCCCCTCCGTCACGCCACC
>16PrdD3	>36prd20
ACACATCACAACCAACGTTACGCTCCCC	GAGTCTCACATACATCCGTCACGCCACC
>17prd2A4	>37prd21
CACGTAAAAAGCCAGACACCGTGACGCA	CACACTAACACGGATTAGTCACACAGTC
>18prd24	>38prd22
CTAAAAACCCAATTAGACTCGGTACCAG	AATTAGTCACACCGGTCCAAAAGTAATG
>19PrdD11	>39prd23
TCACGTTAATTAATGCGGTCACGCATGG	ACCGCGCGCGTGACATTAGTCACGCAACA
>20PrdD12	
TTACAGGTAACCCATTAGTCACGCATCA	

Scr	
Promoter-Stringency	UV5m-10mM
Amino Acid	TKRQRTSYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQN

Sequence	RRMKWKKEHKMASMNIVPYHMG			
Selected sequences				
>ScrE04	>ScrF01	>ScrF11	>ScrG7	>ScrH3
CCTTAATGA	GGAAAGTGGA	TCATTAATGA	CTGTTAATTA	ACGTTAATGA
>ScrE05	>ScrF02	>ScrF12	>ScrG8	>ScrH4
TACATAATGA	CGATAATGA	TCGTTAATGA	CACTAATTA	CTATTAATGA
>ScrE06	>ScrF04	>ScrG1	>ScrG9	>ScrH7
TTGGGTACAA	GACTTAATGA	GCAATTAAAG	CCGTTAATTA	TTGGGTACAA
>ScrE08	>ScrF06	>ScrG2	>ScrG10	>ScrH9
CTACTAATTA	ACGCTAATGA	GAATTAATGA	GTCCAGATTG	GTACTAATGA
>ScrE09	>ScrF07	>ScrG3	>ScrG12	>ScrH10
ATAGGTCCGT	TACACACAGC	GGTTTAATGA	GCGCTAATGA	TATTTAATGA
>ScrE11	>ScrF09	>ScrG5	>ScrH1	>ScrH11
AGTGCTTCAC	GGAAATACGC	CCACTAATTA	TGTTAATGA	TCACTAATGA
>ScrE12	>ScrF10	>ScrG6	>ScrH2	
TCCTTAATGA	CTACTAACTT	GCAATTAACG	GACATAATGA	

Slp1	
Promoter-Stringency	UV5-10mM
Amino Acid Sequence	DKLDVEFDDELEDQLDEDQESEDGNPSKKQKMTAGSDTKKPPYSYNALIMMAIQDSPEQRLTLNGIYQYLINRFPYFKANKRGWQNSIRHNLSLNKCFTKIPRSYDDPGKGNYWILDPSAEEVFIGETTGKLRRKNPGASRTRLAAYRQAIFSPMMAASPYGAPASSYGYPA
Selected Sequences	
>5SlpE1	>5SlpG1
GCTCACTACAATTTGTGTTTGACCAAGA	GAACGCCTCAACGTAAATACATGACAGC
>5SlpE2	>5SlpG2CG
CATAAAAGTACCACAACAAAGCGCATT	GGTGCATGCATATCCCCGGGTATAAACACG
>5SlpE3	>5SlpG3
GGGAATCTACCAAAACATGAACCCAAAC	CAGGAAAAGACAACAACAACACTACCAT
>5SlpE4	>5SlpG4
AATTAAAAAACAAAAACAATCAAAAAAC	CCATTGCGCCATTTATTTACAAAATCC
>5SlpE5	>5SlpG5
GCCGTCAGAGCATAAACATGGCCACTGT	ATAAAGCCCGCAATGAAAACAATACGAA
>5SlpE6	>5SlpG6
CTCCCATATCAGATGTTGGAACACAAAA	CCCCCCAAGAGTAAACAGACTTGGGAGG
>5SlpE7	>5SlpG7
CCCTCATCTAACGGCTGATACGGTGAAA	GCGTAGAGCACCAAAACATACCCGCGTT
>5SlpE8	>5SlpG8
CGCACTGTTTATCTACCGCAGACTAACC	ATGGCGTGGAGTCCAAACACTGATCTGA
>5SlpE10	>5SlpG9
AGCCACCCGATGGTAAACAATGAAAATT	AATAAGACCTGTGTAAATACAATGAAGA
>5SlpE11	>5SlpG10

ACAGAATAATGTAAACACTCGGCAAACC	CCAACGCGAATGTAAACACCTGATTAAC
>5SlpE12CG	>5SlpG11
AAACAATCCGCATAGTTTGTGTAAACACG	GATCTCCCCGCCGTAAACATTACTCAAG
>5SlpF1	>5SlpG12
TTACCCATTAACAAAAACAAAGGACAAG	ACACCAGACAGGTAAACAAGAGACAGAT
>5SlpF2	>5SlpH2
TGAGGCGCCTGTGTAAACAGACACCATT	AGCCTGGACCCGCAAACAAAAGGACTAA
>5SlpF3	>5SlpH3
GCAGCCAAACTGTAAACATACGTCAATA	GACACTTGAACATAAACACCCCCATATA
>5SlpF4	>5SlpH4
GATGTTTACATAACCCAAACAAGACGGT	ATACGCCCTAACCTAAACACAGGGCTAG
>5SlpF5	>5SlpH5
GCCACCTAGATATAAACAAATAACCCAA	GCCTGACAATGTAAACAATACATAAGTG
>5SlpF6	>5SlpH6
ACCAGTAGGACTGGATGTTTACACATCG	TACACTGAGGCGGTGTTTACGCAAGGCG
>5SlpF7	>5SlpH7
AAATAACTCCTTGTAAACAATCCGAGCA	AACTAGTCCCGCTTGTTTATCCACCTTA
>5SlpF9	>5SlpH8
CGTACAGAGTGTGTTTGTGTAACGAAAC	TTGCCGTACTATGTAAACAAGCAAACTC
>5SlpF10	>5SlpH9
GCGTGTGCCGGAAATGTTTTGCCTGAGTG	CGTGCACCCAACCAAAACAATCCTAACA
>5SlpF11	>5SlpH11
TCAAGGACCCCATTGTTTACCTTAGATG	GCACAGCGCAGTATAAACACAATTACCT

tll	
Promoter-Stringency	UV5-10mM
Amino Acid Sequence	SSRILYHVPCKVCRDHSSGKHYGIYACDGCAGFFKRSIRRSRQYVCKS QKQGLCVVDKTHRNQCRACRLRKCFEVGMNKDAVQHERGPRNSTLR RHMAMYKDAMMGAGEMPQIPAEILMNTAALTGFPGVPMPMPGLPQR AGHHPAHMAAFQPPPSAAAVLDLSVPRVPHHPVHQGHHGFFSPTAAY MNALATRALPPTPPLMAAEHIKETAAEHLFKN
Selected Sequences	
>5TllA1	>5TllB11
ACACACAACGCCAAACACACTCCCACCC	GCTAGGACATTTCTAAAGTCAACCCTAA
>5TllA2	>5TllE1
ACGGCGACCTGATCTACCATCAAACACT	CTTACGCTCCCAGAAAAGTCAAAACCAC
>5TllA3	>5TllE2
CAAACCAGCACGCCTTAAAGTCAACGAT	GTCATCCAAACCCCAAAGTCAAAATGTA
>5TllA4	>5TllE3
GCGCCAGGCTCCGCTAAAGTCAGGTTAT	GCGGACGAAGGTTTGCAAAGTCAATTAA
>5TllA6	>5TllE4
ACATCCACCGGGATGAAAGTCAAAACACT	AGGAGCACTCTCATGAAGTCAAATAAAC
>5TllA8	>5TllE5
TATCCCAGCCGAGAGAAAGTCAAAATCA	CTACCCCGAAGCTAAAAGTCAAAGCAAC
>5TllA9	>5TllE6
CCCCAGCCAAGGTGTAAAGTCAACTTGA	GAAGCAACCAGCCAAAAGTCAAACCTTC
>5TllA10	>5TllE8
AACAGCGCCCGATTAAAAGTCAACCGTT	CAACGCTACAAGCAGAAAGTCAACCTAG
>5TllA11	>5TllE9
CGGAGTTTGAGGACGGAAGGCCAAATAA	TCGGGGCACAAGTAAAAGTCAAACCACG

>5TllA12	>5TllF1
ACACTGACCGGGCACGAAGTCAACCTAA	TCCATTCACATAAATAAAAGTCAACTAG
>5TllB2	>5TllF2
ATGAGTGCATGGCAAAAGTCAAACAAGG	TCGATCACACAACTAAAAGTCAATAACC
>5TllB5	>5TllF5
GAGCACCAGTGACAAGAAGTCAACCAAA	ACTCTTGGGACCGTAAAAGTCAACATTA
>5TllB6	>5TllF7
AAACACTGGTAAAGAAAAGTCAACAAAC	AGACCTACCCAAATAGAAGTCAAATTAG
>5TllB7	>5TllF8
GGAAAGTGGCCGCTAAAAGTCAAACCGC	CTGCCCTTATGGGTAAAAGTCAAAGTCA
>5TllB8	>5TllF9
CGACGCCCGAGCAAAAAGTCAATCCAAG	ATTAGATACCAGGAAAAGTCATATCTAA
>5TllB9	>5TllF10
CTCTCGACCTGGCAAAAGTCAAACCCAG	CTCCACAGGGCCGAAAAGTCAATAACTA
>5TllB10	>5TllF11
AATGCTCAGCACTAAAAGTCAAGTCTCA	ATCCCCAACTTACTAAAAGTCAAAACAC

ttk	
Promoter-Stringency	UV5-10mM
Amino Acid Sequence	DYCTKEGEHTYRCKVCSRVYTHISNFCRHYVTSHKRNVKVYPCPFCF KEFTRKDNMTAHVKIIHKIE
Selected Sequences	
>5TtkG1	>5TtkH1
ACGCACTGTAAACAGGATAAGCAAAAAC	ACATACAAGGCAAGGATAACTATTCCAC
>5TtkG2	>5TtkH2
AGAAGGAAAGACCAAGGATAAGCCTCTC	CAGGATAATGGGAAACTAGACATTAATT
>5TtkG3	>5TtkH3
CAAAGGATAATAAACTCTGCCCCAAGTC	GTATCAGGTCTAGGATAATCAGAGCAGT
>5TtkG4	>5TtkH4
CAGTGAGCCTTCGAGGACAATCCTATCC	ATCCCAACAACAGGACAATGGAAACTCG
>5TtkG5	>5TtkH5
GTGCCTCGGGCAGGATAAGCGCACGTGA	CCTCGCTACACAGGATAATTTGTAAATT
>5TtkG7	>5TtkH6
CGACCGAGCCGAGGATAATCTTTCATCT	TTACTCTACTAAGAGGATAATCAACGGC
>5TtkG8	>5TtkH7
CAGAACGCACCGCAAGGATTACAGAACT	TAGTGTGGCAACCAGGACAATATACGCA
>5TtkG9	>5TtkH8
ATGACTATAAGAAAGGACAAACTAGTTT	GGGGCACGACGCCAGGATAATGAGATCT
>5TtkG10	>5TtkH9
AAGGATAACACATATAATTAAGACGGGC	CAAATCAAACGAAGGATAATCCAAAAGA
>5TtkG11	>5TtkH10
GTATCAGCCACCTAAAGGATAACCCATCC	GGTATATCCACCAGGATAATGCAAAGTT
>5TtkG12	>5TtkH11
AGCGGAGCCACAAAAGGACAACTGCAATG	GACACATCCGCCCAAGGATAATCAGACC

Ubx	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	RRRGRQTYTRYQTLELEKEFHTNHYLTRRRRIEMAHALCLTERQIKIWF QNRRMKLKKEIQAIKELNEQEKQA

Selected sequences				
>UbxC02	>UbxC07	>UbxD02	>UbxD08	
TAATTAATTA	TGCAATAAAA	GCGTTAATTA	TCCTTAATGA	
>UbxC03	>UbxC08	>UbxD04	>UbxD09	
AATTTTATTA	GGCAATTAAG	GCCTTAATTA	GCCTTAATTA	
>UbxC04	>UbxC11	>UbxD05	>UbxD10	
GCTTTAATTA	GTATTAATGA	GACAATTAAA	CTTTTTATGA	
>UbxC05	>UbxC12	>UbxD06	>UbxD11	
GCTTTAATTA	AATTTAATGG	CCGTTAATTA	GGTAATTAAC	
>UbxC06	>UbxD01	>UbxD07	>UbxD12	
CTATTAATTA	TCTTAATGA	GCCCATTAAA	TGCAATTAAA	

Table A.2

Amino acid sequence, selection promoter strength/stringency and the binding sites recovered for each homeodomain assayed.

Abd-A				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RRRGRQTYTRFQTLELEKEFHFNHYLTRRRRIEIAHALCLTERQIKIWFQNRRMKLKKELRAVKEINEQAR			
Selected sequences				
>AbdAG04	>AbdAG09	>AbdAH02	>AbdAH07	>AbdAH12
TTTTTAATTA	TACCAAACCC	TACGTAACTT	CGTTCTTTAA	ATCTTAATTAC
>AbdAG05	>AbdAG10	>AbdAH03	>AbdAH08	>AbdAG02
CTACCATTTT	CACTAATTA	GGTCATTAAA	GATTTAATTA	GGTAATTAAA
>AbdAG06	>AbdAG11	>AbdAH04	>AbdAH09	>AbdAG03
TTAATTAC	CATAATTA	TTTTTTATGA	GCGCTAATGA	TCGTTAATGA
>AbdAG07	>AbdAG12	>AbdAH05	>AbdAH10	
TTCTTTATTA	GTTTTAATTA	CTACTAATTC	TGTTTAATGA	
>AbdAG08	>AbdAH01	>AbdAH06	>AbdAH11	
GGACCCACAT	GGTTGCGGCC	TGCAATTAAA	GGCAATTAAG	

Abd-B				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	VRKKRKPYSKFQTLELEKEFLFNAYVSKQKRWELARNLQLTERQVKIWFQNRRMKNKKNSQRQANQQNNNN			
Selected sequences				
>abdb2	>abdb7	>abdb13	>abdb19	>abdb24
GGGTTTATAG	GGTTTACAAC	TTTTTATAAC	GCTTTTATTA	TGTTTTATGA
>abdb3	>abdb8	>abdb14	>abdb20	
GTTTTATTGT	CGTTTAATGT	TTATTAATTA	ACTTTTACGA	
>abdb4	>abdb10	>abdb15	>abdb21	
TTTTTTATGG	TGATTTATGT	GTTTTATGA	TGATTTATTA	
>abdb5	>abdb11	>abdb17	>abdb22	
TGATTAATGG	CATATTATGA	AGTTTTATGG	TGATTTATTA	
>abdb6	>abdb12	>abdb18	>abdb23	
CGCTTTATGT	GCATTTATTA	TCTTTAACGA	TCTTTAATTA	

Achi	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	LRKRRGNLPKTSVKILKRWLYEHRYNAYPSDAEKFTLSQEANLTVLQVCNWFINARRRILPEMIRREGNDPLHFT

Selected sequences				
>AchiG1	>AchiG8	>AchiH5	>achi2F3	>achi2F11
TGTCAAG	GCCTGTCATA	AATGGATATT	CTGTCAATC	TGCTGTCAAA
>AchiG2	>AchiG9	>AchiH6	>achi2F4	>achi2F12
AATGTGACA	GTTGTCAAAA	GCTGTCAAAA	AGCTGTCAAA	CCTGTCAAAC
>AchiG3	>AchiG10	>AchiH8	>achi2F5	
AGCTGTCATG	AGATCTGACA	CTCTTTAGCC	AAACTAAGAT	
>AchiG4	>AchiH1	>AchiH10	>achi2F6	
TTATCTGACA	ATATCAAATG	ATTCCGCTCT	TAAGATGACA	
>AchiG5	>AchiH2	>AchiH11	>achi2F8	
AGCATGACAG	AGCTGTCAGA	GCTGTCAAAG	CCTGTCAAA	
>AchiG6	>AchiH3	>achi2F1	>achi2F9	
TGTTTGACAT	GAAATGACA	CGAATGACAA	GTCCAGATTG	
>AchiG7	>AchiH4	>achi2F2	>achi2F10	
AAGGCAGAGA	ATCTGTCAGT	TGTTGTCAAA	GTCCAGATTG	

Al				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	QRRYRTTFTSFQLEELEKAFSRTHYPDVFTREELAMKIGLTEARIQVWFQ NRRAKWRKQEKVGPQSHPYN			
Selected sequences				
>AlE2	>AlE9	>AlF2	>AlF7	>AlF12
GCTAATTAAT	GATTAATTAA	GACTAATTAA	TAATTAATT	GTTAATTAAA
>AlE5	>AlE10	>AlF3	>AlF8	
TTTTAATTAA	TTCTAATTAA	CGCTAATTGA	CACTAATTAC	
>AlE6	>AlE11	>AlF4	>AlF9	
TGCTAATTAA	TATAATTAA	TCATAATTAA	GTCTAATTAA	
>AlE7	>AlE12	>AlF5	>AlF10	
GCATAATTAA	TGCTAATTAA	CGCTAATTGG	GGCTAATTAA	
>AlE8	>AlF1	>AlF6	>AlF11	
GACTAATTAA	ACACGCTAAT	ACCTAATTAA	CCCTAATTGA	

Antp				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQ NRRMKWKKENKTKGEPGSGGEGD			
Selected sequences				
>AntpA02	>AntpA07	>AntpB01	>AntpB08	

GCCTTAATTA	GGTTTAATGA	TTCATAATTA	GGCAATTAAG	
>AntpA03	>AntpA08	>AntpB02	>AntpB09	
AATTTAATTA	GGC TTAATGA	GTGTTAATTA	CGTTTAATTA	
>AntpA04	>AntpA10	>AntpB04	>AntpB11	
AGCTTAATGA	CTACTAATTA	GATTTAATTA	TTTTTAATGA	
>AntpA05	>AntpA11	>AntpB06	>AntpB12	
TTGTTAATGA	CCCTTAATGG	TGTTTAATGA	CCTTTAATGA	
>AntpA06	>AntpA12	>AntpB07		
GGTAATTAAA	TAGCACTTTT	TTTTTAATGA		

Ap				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	TKRMRTSFKHHQLRTMKSYFAINHNPDAKDLKQLSQKTGLPKRVLQVW FQNARAKWRRMMMKQDGSGLLEKGE			
Selected sequences				
>ApA2	>ApA7	>ApB2	>ApB7	
ATCATTAACC	GTGCTAATTG	TGCCAATATA	TCGTTAATGA	
>ApA3	>ApA8	>ApB3	>ApB8	
AGACTAATTG	GCTAATTAAT	TCCTTAATGA	TCGCTAATTA	
>ApA4	>ApA9	>ApB4	>ApB9	
AGGTTAATTA	CCAATTATGA	CTAATTAAGT	CGACTAATTA	
>ApA5	>ApA10	>ApB5	>ApB10	
TAAATAATGA	CTAATTAAGG	TCTCATTAAA	CGCTAATGA	
>ApA6	>ApA11	>ApB6	>ApB11	
CTAATTAGCG	ACAAATTAAG	ACTAATTAAT	CTAATTAGCA	

Ara				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	LAARRKNATRESTATLKAWLNEHKKNPYPTKGEKIMLAIITKMTLTQVS TWFANARRRLKKENKMTWEPKNRTDDD			
Selected sequences				
>AraG1	>AraG11	>AraH10	>Ara2G7	>Ara2H4
GGGTATTACA	GTGATATACA	TGCAAAAACA	GGTAATAACA	TCAAATTACA
>AraG2	>AraG12	>AraH11	>Ara2G8	>Ara2H6
TGTACTTACA	GTTTAGAACA	AGCTTTTACA	GTGTAGAACA	TGGAAAAACA
>AraG4	>AraH3	>Ara2G1	>Ara2G10	>Ara2H8
GCAGTTTACA	TCGAATAACA	ACCCCAAACA	CCTGAAAACA	TAACATTACA
>AraG5	>AraH5	>Ara2G2	>Ara2G11	>Ara2H9
CGTAATTACA	AAACATAACA	GCAATTAACA	GAAAGAACA	CTACAAAACA

>AraG6	>AraH6	>Ara2G3	>Ara2G12	>Ara2H10
TCTGTTGAGT	CATGTAAACA	GCAAGTTACA	AGAATTAACA	AGAGAAAACA
>AraG7	>AraH7	>Ara2G4	>Ara2H1	>Ara2H11
CAGGAAAACA	TGTTCAGCTA	TAAGAGAACA	GGAAAAAACA	ATGAATTACA
>AraG8	>AraH8	>Ara2G5	>Ara2H2	
TTGTATTACA	TTCAAAAACA	CATATAAACA	TCCCAAAACA	
>AraG9	>AraH9	>Ara2G6	>Ara2H3	
GTCCAGATTG	AAACATAACA	GGGAAAAACA	TGAAGCCTTG	

<table>
<tr><td colspan="5">Awh</td></tr>
<tr><td>Promoter-Stringency</td><td colspan="4">UV5m-10mM</td></tr>
<tr><td>Amino Acid Sequence</td><td colspan="4">KTKRVRTTFTEEQLQVLQANFQIDSNPDGQDLERIASVTGLSKRVTQVWFQNSRARQKKHIHAGKNKIREPEGS</td></tr>
<tr><td colspan="5"></td></tr>
<tr><td colspan="5">Selected sequences</td></tr>
<tr><td>>AwhC1</td><td>>AwhC10</td><td>>AwhD8</td><td>>Awh2A9</td><td>>Awh2B6</td></tr>
<tr><td>GTAATCAAAT</td><td>TTCATTAATA</td><td>CAATTAGCG</td><td>TCTAATTAAA</td><td>ATCCTAATTA</td></tr>
<tr><td>>AwhC2</td><td>>AwhC11</td><td>>AwhD9</td><td>>Awh2A10</td><td>>Awh2B7</td></tr>
<tr><td>CCAATCAGCC</td><td>TCCTGACAGT</td><td>TAGTTAATTA</td><td>CCAATTAGCA</td><td>TCTAATTAAA</td></tr>
<tr><td>>AwhC3</td><td>>AwhC12</td><td>>AwhD10</td><td>>Awh2A11</td><td>>Awh2B8</td></tr>
<tr><td>GTTCATTAAA</td><td>CACTTGATTA</td><td>GTACTAATGA</td><td>CATAATTA</td><td>CTATTAATTA</td></tr>
<tr><td>>AwhC4</td><td>>AwhD1</td><td>>AwhD11</td><td>>Awh2A12</td><td>>Awh2B9</td></tr>
<tr><td>CTGATTACGC</td><td>CTCATTAGCG</td><td>TACCTAATGA</td><td>TTAATTAGGC</td><td>ACCCTAATGA</td></tr>
<tr><td>>AwhC5</td><td>>AwhD2</td><td>>Awh2A1</td><td>>Awh2B1</td><td>>Awh2B10</td></tr>
<tr><td>GATTTGATTA</td><td>TGTCTAATTA</td><td>GCCTTAATTT</td><td>TAGTTAATTA</td><td>TAGTTAATGA</td></tr>
<tr><td>>AwhC6</td><td>>AwhD3</td><td>>Awh2A2</td><td>>Awh2B2</td><td>>Awh2B11</td></tr>
<tr><td>TCAGCTAATT</td><td>GTAATTAGGT</td><td>CTAATCAAAT</td><td>AAGATAATTA</td><td>CCTCTAATTG</td></tr>
<tr><td>>AwhC7</td><td>>AwhD5</td><td>>Awh2A3</td><td>>Awh2B3</td><td></td></tr>
<tr><td>CTACTTAATT</td><td>TCACTAATTA</td><td>TTAATTATC</td><td>ATAATTAAG</td><td></td></tr>
<tr><td>>AwhC8</td><td>>AwhD6</td><td>>Awh2A4</td><td>>Awh2B4</td><td></td></tr>
<tr><td>CTAATTACTC</td><td>GTAATTAAAG</td><td>CTTTTGATTA</td><td>TAACTAATTA</td><td></td></tr>
<tr><td>>AwhC9</td><td>>AwhD7</td><td>>Awh2A8</td><td>>Awh2B5</td><td></td></tr>
<tr><td>ATAATTATGG</td><td>TTTTTAATTA</td><td>ACCTAAATGA</td><td>CCACTAATTT</td><td></td></tr>
</table>

<table>
<tr><td colspan="5">Bap</td></tr>
<tr><td>Promoter-Stringency</td><td colspan="4">UV5m-10mM</td></tr>
<tr><td>Amino Acid Sequence</td><td colspan="4">KKRSRAAFSHAQVFELERRFAQQRYLSGPERSEMAKSLRLTETQVKIWFQNRRYKTKRKQIQQHEAALLGASK</td></tr>
<tr><td colspan="5"></td></tr>
<tr><td colspan="5">Selected sequences</td></tr>
<tr><td>>BapA2</td><td>>BapA11</td><td>>BapB8</td><td>>Bap2C5</td><td>>Bap2D3</td></tr>
</table>

GCTTAAGTGG	ATCTTATTAC	GGACCCACAT	GACTTAAGTG	GGTTAAGTGG
>BapA3	>BapA12	>BapB9	>Bap2C6	>Bap2D4
TCTTAAGTGG	AGTTAAGTGG	TGTTAAGTGG	GCTTAAGTGC	CTTTATAACT
>BapA4	>BapB1	>BapB10	>Bap2C7	>Bap2D5
TAACGCTGCA	GTTTAAGTGG	TCTTCCATTA	ACTTAAGAAC	CTTTATAACT
>BapA5	>BapB3	>BapB11	>Bap2C8	>Bap2D6
ATCAGTCCTT	ACTTATCTGA	CGTTAAGTGG	TCTTAAGTAC	ACTTAACGT
>BapA6	>BapB4	>Bap2C1	>Bap2C9	>Bap2D9
AGTTCTTAAG	TCACGCCGAG	ACACTTAAAG	ATTTAAGTGA	TTAAGTACC
>BapA7	>BapB5	>Bap2C2	>Bap2C10	>Bap2D10
ATCTTATTAC	TGTTAAGTGG	ACTTAAGTAC	TATTAAGTAC	ATTTAAGTGA
>BapA9	>BapB6	>Bap2C3	>Bap2C12	>Bap2D11
ACTTATCTGA	TCTTAAGTGG	TTTTAAGTGT	TTTTAAGTGA	ACTTAAGTAC
>BapA10	>BapB7	>Bap2C4	>Bap2D2	
CATTAATAAC	ACTTAAGTAC	CTGTAAGTGT	CTTTAAGTAC	

Bcd				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PRRTRTTFTSSQIAELEQHFLQGRYLTAPRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYEG			
Selected sequences				
>bcd1	>bcd6	>bcd11	>bcd17	>bcd23
TGTTAATCCG	TCTTAATCCC	CGGGTAATCC	GGTTATCCG	GGTTAATCCG
>bcd2	>bcd7	>bcd13	>bcd18	>bcd24
ATGGATTAGA	GCTTAATCCG	TGTTAATCC	TGTTAATCCC	ATGGATTAGA
>bcd3	>bcd8	>bcd14	>bcd20	
CGTTAATCTC	GGGTTAATCC	TGGGATTATA	CGCTTAATCC	
>bcd4	>bcd9	>bcd15	>bcd21	
GGTTTAATCC	GAGATAATCC	GCGTAATCCA	TTACTAATCC	
>bcd5	>bcd10	>bcd16	>bcd22	
TCTATAATCC	AGCTTATCC	GGCTTAAGCC	GTCCTAATCC	

BH1				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	QRKARTAFTDHQLQTLEKSFERQKYLSVQERQELAHKLDLSDCQVKTWYQNRRTKWKRQTAVGLELLAEAGN			
Selected sequences				
>B1HG1	>B1HG10	>B1HH8	>BH12E8	>BH12F5
GCAAACCCCT	TGTTAAACGG	GCTGTTATTT	AGCAATTATG	ATTGGCCACC

>B1HG2	>B1HG11	>B1HH9	>BH12E9	>BH12F6
TCGTTTCCCT	TCACGCCGAG	GGCAATTAAG	TGTTAAACGG	AACATTTAAT
>B1HG3	>B1HH1	>B1HH10	>BH12E10	>BH12F8
TCACGCCGAG	CCCTAACATG	CCCGTTCGTG	TCTTAAACGG	TTGGGTACAA
>B1HG4	>B1HH2	>BH12E2	>BH12E11	>BH12F9
GCCTAAGAGA	CATTGTACTA	TTCTAAACGG	TCTTAAACGG	GTCGGTACCC
>B1HG5	>B1HH3	>BH12E3	>BH12E12	>BH12F10
GGTAAAGCAT	GGTAAAGCAT	GCCAATTAAC	AGATAATTGC	CTACCAGATT
>B1HG6	>B1HH4	>BH12E4	>BH12F1	>BH12F11
CGCTAAAGTG	TCTGTTGAGT	GGCTAATTGA	GGATAATTGA	TTCTAATTGA
>B1HG7	>B1HH5	>BH12E5	>BH12F2	
ATCTTATTAC	CGTTAATT	AGTTAATAGG	ACTTAAACGT	
>B1HG8	>B1HH6	>BH12E6	>BH12F3	
GCTAATTGA	AGTTGACCAC	AACAATTAAC	GATAATTAA	
>B1HG9	>B1HH7	>BH12E7	>BH12F4	
GTTAATTGA	TCACGCCGAG	ATCAATTAAC	CTTTAAACGG	

BH2				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KQRKARTAFTDHQLQTLEKSFERQKYLSVQDRMELANKLELSDCQVKT WYQNRRTKWKRQTAVGLELLAEAGN			
Selected sequences				
>BH2E1	>BH2E7	>BH2E12	>BH2F5	>BH2F10
GCCGTTTATC	AGCCATTAAG	GCCCATTAGA	ACCAATTATC	ATCAATTAAG
>BH2E2	>BH2E8	>BH2F1	>BH2F6	>BH2F11
TGCAATTAAA	GCTAATTGA	ACCAATTATC	AGCAATTATA	ACCAATTAAG
>BH2E4	>BH2E9	>BH2F2	>BH2F7	
GACTATTAAG	ACCTTTTAAG	GTCATTTAAG	ATCAATTAAC	
>BH2E5	>BH2E10	>BH2F3	>BH2F8	
GCCATTTAGT	ACCAATTAAT	ACCAATTACA	AGCAATTAAC	
>BH2E6	>BH2E11	>BH2F4	>BH2F9	
ACCTATTAAA	ATCATTTATC	TCCAATTATG	ATCAATTAAC	

Bsh				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RRRKARTVFSDPQLSGLEKRFEGQRYLSTPERVELATALGLSETQVKTW FQNRRMKHKKQLRRRDNANEP			
Selected sequences				
>BshE1	>BshE7	>BshE12	>BshF7	

ACCCATTAGG	CTCTTAACGA	GTCGATTAAG	GCTAATTAAT	
>BshE2	>BshE8	>BshF1	>BshF8	
GTCGATTAAA	GATCATTAAG	GCTCATTAGA	GCTAATTAAT	
>BshE3	>BshE9	>BshF3	>BshF9	
CGTCATTAAG	TTTTTAATTA	CTACTAATAT	GCTCGTTAAG	
>BshE4	>BshE10	>BshF4	>BshF10	
ACCAATTACA	AAGATAATGA	GCTCATTAGA	GCCAATTATA	
>BshE6	>BshE11	>BshF6	>BshF11	
TGCAATTAAT	GTCGATTAAA	GGCAATTAAC	ACCAATTACA	

Btn				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	NRKERTAFSKTQLKQLEAEFCYSNYLTRLRRYEIAVALELTERQVKVWF QNRRMKCKRIKLEEQQGSSAKT			
Selected sequences				
>BtnE1	>BtnE6	>BtnE11	>BtnF4	>BtnF9
TGCCATTAAA	GCTGTAATTA	GATAATTAA	AATATAATGA	TCCTTAATGA
>BtnE2	>BtnE7	>BtnE12	>BtnF5	>BtnF10
TCAGTAATGA	GGTCATTACC	CTCGTAATGA	GCTAATTATC	ACGTTAATGA
>BtnE3	>BtnE8	>BtnF1	>BtnF6	>BtnF11
CACTTGATTA	GGTCGTTAAG	TATCATTATA	CGTAATTAAT	GAGTTAATGA
>BtnE4	>BtnE9	>BtnF2	>BtnF7	
GTACATTAAT	AGCCTAATTA	CTTAATTAAC	AGTCATTAAG	
>BtnE5	>BtnE10	>BtnF3	>BtnF8	
CCCCATTAAG	AGCAATTATA	GGTCATTAAT	GACTTAATGA	

C15				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KRKKPRTSFTRIQVAELEKRFHKQKYLASAERAALARGLKMTDAQVKT WFQNRRTKWRRQTAEEREAERQ			
Selected sequences				
>C15A1	>C15A6	>C15A11	>C15B6	>C15B11
AATTTAAAGA	GTTTAACA	ATCATTTAAG	ATCTTATTAC	GGCAATTAAC
>C15A2	>C15A7	>C15A12	>C15B7	
GCCAATTAAC	TGTTTAACGA	AATTTAATTG	TCACGCCGAG	
>C15A3	>C15A8	>C15B1	>C15B8	
CCCCATTAAC	ACGTTAACGA	GCCAATTAAC	GCTAATTAAT	
>C15A4	>C15A9	>C15B2	>C15B9	
GTCGTTTAAG	ATCTTAATTG	ACCCATTATT	CCTATTTAAA	

>C15A5	>C15A10	>C15B4	>C15B10	
GCTCGTTAAG	TCGTTAATGA	GCTAATTAAA	TCACAATCAC	

Cad				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KDKYRVVYTDFQRLELEKEYCTSRYITIRRKSELAQTLSLSERQVKIWFQ NRRAKERKQNKKGSDPNVMGVG			
Selected sequences				
>CadE1	>CadE11	>CadF8	>Cad2E7	>Cad2F4
CCACAAATTA	AGCAATTAAG	CTCAATAAAA	ACTCTAATTG	ACCATAATTA
>CadE2	>CadE12	>CadF9	>Cad2E8	>Cad2F5
CTAATCAACA	CTAATAAAA	ATCATAAAAC	CCAATAAACT	GCAATAAAAA
>CadE3	>CadF1	>CadF10	>Cad2E9	>Cad2F6
GTAATAACTT	ACCGTAATTA	ATTCCGCTCT	GCAATCATTA	CTTTTTATTG
>CadE4	>CadF2	>CadF11	>Cad2E10	>Cad2F7
GAATTAATAG	CCAATAAATG	GTAATAAAGT	CCTTAAATTA	CCCATAAATT
>CadE5	>CadF3	>Cad2E1	>Cad2E11	>Cad2F9
GCTTAAATGA	GCCATTAAAG	CCAATAAAGG	AAAAGGATTC	GTTTTTATGA
>CadE6	>CadF4	>Cad2E2	>Cad2E12	>Cad2F10
ATGATTTATTT	AGTTTAATAA	GACATTATTA	AGGCACTACG	CCCATATAAT
>CadE8	>CadF5	>Cad2E3	>Cad2F1	>Cad2F11
CTTATAAAAT	CTATTTATTA	GCTAATAAAT	GTTCTAATTA	GCAATAAAAA
>CadE9	>CadF6	>Cad2E4	>Cad2F2	
ACAGTAATTA	TATTTTATTA	TTATTTATTA	CTCATAAACA	
>CadE10	>CadF7	>Cad2E5	>Cad2F3	
GAACACTACT	GCAATAAACA	TCGAGCATGT	GGATTTATAA	

Caup				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	LAARRKNATRESTATLKAWLSEHKKNPYPTKGEKIMLAIITKMTLTQVS TWFANARRRLKKENKMTWEPKNKTEDD			
Selected sequences				
>CaupA2	>CaupA7	>CaupA12	>CaupB5	>CaupB11
ATTATTAACA	TTCATTAACA	GGCTAAAACA	GATAATAACA	ATAAAAAACA
>CaupA3	>CaupA8	>CaupB1	>CaupB6	
GTCTTTTACA	GCAGTTAACA	GCTAATAACA	CTAAAAACA	
>CaupA4	>CaupA9	>CaupB2	>CaupB7	
AGCAAAAACA	GGAAAGAACA	ATTTGTGACA	TGAAAAACA	
>CaupA5	>CaupA10	>CaupB3	>CaupB9	

TGGTTGTACA	TTCTTTAACA	ACCGGAAACA	CGCAGCAACA	
>CaupA6	>CaupA11	>CaupB4	>CaupB10	
AACTATTACA	AAAAATGACA	TCGGTTAACA	AACTGTAACA	

Ct				
Promoter-Stringency	UV5-5mM			
Amino Acid Sequence	SKKQRVLFSEEQKEALRLAFALDPYPNVGTIEFLANELGLATRTITNWFHNHRMRLKQQVPHGPAGQDNPIPS			
Selected sequences				
>5CtE1	>5CtE6	>5CtE11	>5CtF4	>5CtF9
CTTGCTAAAC	TAGATTAAAC	TAGTTCAAAG	AGATTGACTA	TGGCTAAAAC
>5CtE2	>5CtE7	>5CtE12	>5CtF5	>5CtF10
TCGTCTGAAC	GCTCTTGAAC	ACGCTCGAGC	CATTCTGAAC	TGTCTTGAAC
>5CtE3	>5CtE8	>5CtF1	>5CtF6	>5CtF11
AAGGTTAAAC	GGCAAAGAGA	AGTCCTGAAC	AGCCTTGAAC	CGTCCTGAAC
>5CtE4	>5CtE9	>5CtF2	>5CtF7	
TACGTTAATC	GAAGTTAAAC	GGGATAAAC	GTACTTAAAC	
>5CtE5	>5CtE10	>5CtF3	>5CtF8	
CACATTGAAC	TATGCTAATC	CTAATTGAAC	GTGGTTGAAC	

Dfd				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PKRQRTAYTRHQILELEKEFHYNRYLTRRRRIEIAHTLVLSERQIKIWFQNRRMKWKKDNKLPNTKNVRKKT			
Selected sequences				
>dfd1	>dfd6	>dfd11	>dfd16	>dfd21
CTTCATTAAG	CCTAATTAAG	AGCTATTAAA	CTCATTACT	CGACTAATGA
>dfd2	>dfd7	>dfd12	>dfd17	>dfd22
GGTCATTAAT	GATAATTAAT	GCACTAATGA	CTTCATTAAG	TATCATTAAC
>dfd3	>dfd8	>dfd13	>dfd18	>dfd23
TATCATTAAA	CCTAATTAAG	TCGTAATGA	AGTCATTAGG	CCGTTAATGA
>dfd4	>dfd9	>dfd14	>dfd19	>dfd24
GGTCATTAAT	CCCCATTAAT	TGCTTAATGG	TACCTAATGA	CAATTAATGA
>dfd5	>dfd10	>dfd15	>dfd20	
GTCATTAACA	TTTTTAATGA	ATCGTAATTA	TGGATAATGA	

Dll	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	KMRKPRTIYSSLQLQQLNRRFQRTQYLALPERAELAASLGLTQTQVKIWFQNRRSKYKKMMKAAQGPGTNSG

Selected sequences				
>DllC1	>DllC6	>DllC11	>DllD4	>DllD9
AGACAATTAA	ACAAATTAGG	AGAGTAATTA	GGTTAATTAC	AGTAATTACA
>DllC2	>DllC7	>DllC12	>DllD5	>DllD10
GTCCATTATCA	TGCCATTAAA	GATCATTAGA	CGCAATTAGA	CTAATTACA
>DllC3	>DllC8	>DllD1	>DllD6	>DllD11
CGCTATTACA	CACAATTTGT	CGTTATTAAG	CGCAATTATA	CTACTAATTA
>DllC4	>DllC9	>DllD2	>DllD7	
AGAAATTAAC	ACGTAATTAT	CATAATTTTC	TCCAATTACG	
>DllC5	>DllC10	>DllD3	>DllD8	
TATAATTTTA	TAAGTAATTA	CTGATAGGCG	TGGATAATTA	

Dr				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	NRKPRTPFTTQQLLSLEKKFREKQYLSIAERAEFSSSLRLTETQVKIWFQNRRAKAKRLQEAEIEKIKMAALG			
Selected sequences				
>DrA2	>DrA7	>DrB2	>DrB7	>DrB12
ACCTCAATTA	TCACCAATTA	CCTCCAATTA	GGACCAATTA	AGTCCAATTA
>DrA3	>DrA8	>DrB3	>DrB8	
AAAGCAATTA	AGAGCAATTA	CGGCCAATTA	TCTCCAATTA	
>DrA4	>DrA9	>DrB4	>DrB9	
TGGCCAATTA	AGGGTAATTA	AAAACAATTA	TGAGTAATTA	
>DrA5	>DrA10	>DrB5	>DrB10	
TAACTAATTA	TGACTAATTA	GAACCAATTA	TCAGCAATTA	
>DrA6	>DrA11	>DrB6	>DrB11	
TCTCCAATTA	TCGCTAATTA	CGAGCAATTA	GCCCCAATTA	

E5				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PKRVRTAFSPTQLLKLEHAFEGNHYVVGAERKQLAQGLSLTETQVKVWFQNRRTKHKRMQQEGGDGSDTKSNK			
Selected sequences				
>E5C1	>E5C11	>E5D9	>E52C8	>E52D6
TTCAATTAGG	TTTATAATTA	AATCTAATTA	TCAATAATGA	TCAATTAAGA
>E5C2	>E5C12	>E5D10	>E52C9	>E52D7
TTAATTAAAT	TTAATTAGAA	GCACTAATGA	GAACCCACA	ACATTAATGA
>E5C3	>E5D1	>E5D11	>E52C10	>E52D8

CTAATTTACA	CCGGGAAGGT	TGACTAATGA	ATTCATTAGT	ACAATAATTA
>E5C4	>E5D2	>E52C1	>E52C11	>E52D9
CCATTAAGC	GGTATAATGA	TGACTAATAG	CGGGTAATTA	CATCTAATGA
>E5C5	>E5D3	>E52C2	>E52C12	>E52D10
ATAAATTAAC	TCAATAGTGC	CCCATTAACC	GGTTTAATGA	ACGGTAATTA
>E5C6	>E5D4	>E52C3	>E52D1	>E52D11
TTGTTAATAA	GTGTTAATTG	GGGCTAAATA	GATGTAATGA	AGTAATTAGC
>E5C7	>E5D5	>E52C4	>E52D2	
ACCGTAATTA	TAGCTAATGA	TTCATTACGT	TTTTTAATTA	
>E5C8	>E5D6	>E52C5	>E52D3	
GCTTAAGTA	AGTTGACGC	CACTATTAAG	CGCATAATTA	
>E5C9	>E5D7	>E52C6	>E52D4	
ATCAATTAGC	TGGCTAATTA	CATTAAATGA	CCGTTAATAG	
>E5C10	>E5D8	>E52C7	>E52D5	
CCTGTAATGA	ATATTAATTA	GACTTAATAA	AAACTAATTA	

Ems				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PKRIRTAFSPSQLLKLEHAFESNQYVVGAERKALAQNLNLSETQVKVWFQNRRTKHKRMQQEDEKGGEGGSQR			
Selected sequences				
>emsA1	>emsA6	>emsA11	>emsB5	>emsB10
TTAATTATA	GGTCATTACT	CCAATTATTG	TTTCTAATGA	CTAATTAGAG
>emsA2	>emsA7	>emsA12	>emsB6	>emsB11
CCATTTATGT	GTCCATTAAT	CCATAAATTA	TGCCTAATGA	CTAATTAGCG
>emsA3	>emsA8	>emsB1	>emsB7	
CATTTTATGA	TGTGATTAAC	TCTGGAGAGG	TCACTAATTA	
>emsA4	>emsA9	>emsB2	>emsB8	
GCCATGGACC	ACATAAATGA	GCCAATTATA	GGTCTAATGA	
>emsA5	>emsA10	>emsB3	>emsB9	
TTCACTAATA	ACTAATTAAA	CTCCATTAAA	CCAATTAGAG	

En				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	EKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKSTGSKNPLALQLMAQ			
Selected sequences				
>eng1	>eng7	>eng12	>eng17	>eng22
CTAATTAGCG	GTGCTAATTA	CAATTAAAA	CGACTAATTA	CCAATTAAAC

>eng2	>eng8	>eng13	>eng18	>eng23
TATTTAATTA	TCAATTAACC	CCAATTAAAA	CCAATTAAAA	TCAATTAAG
>eng3	>eng9	>eng14	>eng19	>eng24
CTCATTAGTG	ACGTTAATTA	TAACTAATTA	CTCAATTAAG	CGGCTAATTA
>eng4	>eng10	>eng15	>eng20	
AGGGTAATTA	TAGGTAATTA	CTCTTAATTG	GCGTTAATGA	
>eng5	>eng11	>eng16	>eng21	
GCAATTATCA	CGGCTAATTA	CCGATAATTG	GCTAATTAAG	

Eve				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	VRRYRTAFTRDQLGRLEKEFYKENYVSRPRRCELAAQLNLPESTIKVWFQNRRMKDKRQRIAVAWPYAAVYSD			
Selected sequences				
>Eve-G1	>Eve-G6	>Eve-G11	>Eve-H4	>Eve-H9
TCCGACATAA	CTTCTAACGA	ACACATTAAC	TGTTTAATGA	TTGCTAATGA
>Eve-G2	>Eve-G7	>Eve-G12	>Eve-H5	>Eve-H10
TTACTTAATT	CATCATTATA	CCTCATTATG	CGGCTAATTA	CCTCATTAAT
>Eve-G3	>Eve-G8	>Eve-H1	>Eve-H6	>Eve-H11
TCGATTATTA	GTCGTTAGTA	GTTAATTAAA	TTGCTAATTA	GGTCATTAAC
>Eve-G4	>Eve-G9	>Eve-H2	>Eve-H7	
CTTCTAATCA	TGGCTAATTG	TCCCATTAAC	ACACTAATTA	
>Eve-G5	>Eve-G10	>Eve-H3	>Eve-H8	
TAGTAAATTA	TCAATTAGAC	AGTCATTAAA	GTAATTAGTA	

Exd				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	ARRKRRNFSKQASEILNEYFYSHLSNPYPSEEAKEELARKCGITVSQVSNWFGNKRIRYKKNIGKAQEEANLYAAKKAAGAS			
Selected sequences				
>ExdE1	>ExdE6	>ExdE11	>ExdF4	>ExdF10
TGTTTGACAT	CCATTTGACA	GCGCATGAGA	GCATTTGACA	AGTCTGTGGT
>ExdE2	>ExdE7	>ExdE12	>ExdF5	
CATGATGATT	ACTTTGATGA	GCTTTTGACA	CGGTTTGACA	
>ExdE3	>ExdE8	>ExdF1	>ExdF6	
TCTTTGACAT	CGATTGATGA	GAGTTGACAT	TGTTTGATGA	
>ExdE4	>ExdE9	>ExdF2	>ExdF7	
CTTCATAC	CAAGTTGACA	ATTTTGATGG	AATTTTGACA	
>ExdE5	>ExdE10	>ExdF3	>ExdF9	

CCTTTTGACA	ATTTTGACGA	ACTTTGATGA	GTCTTTGACA	

Exex				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	TRRPRTAFTSQQLLELEKQFKQNKYLSRPKRFEVASGLMLSETQVKIWFQNRRMKWKRSKKAQQEAKERAKAN			
Selected sequences				
>ExexE1	>ExexE6	>ExexE11	>ExexF4	>ExexF9
CTTAATTAGA	ACAGTAATTA	TGGCTAATTA	TGTAATTAAT	AGCAATTAAA
>ExexE2	>ExexE7	>ExexE12	>ExexF5	>ExexF10
ACCAATTAAG	ACGTTAATTA	AGAGTAATTA	ACTTAATCAC	CATGTAATTA
>ExexE3	>ExexE8	>ExexF1	>ExexF6	>ExexF11
GGGTAATTAA	ATCTTATTAC	GGTAATTAGA	TCTAATTAA	CGCTAATTAA
>ExexE4	>ExexE9	>ExexF2	>ExexF7	
AGAGTAATTA	GCTAATTACT	TTCTAATTGA	ACTAATTAG	
>ExexE5	>ExexE10	>ExexF3	>ExexF8	
CTTAATTATC	TGTGTAATTA	AGTTAATTAC	TGGCTAATTA	

ftz				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	SKRTRQTYTRYQTLELEKEFHFNRYITRRRRIDIANALSLSERQIKIWFQNRRMKSKKDRTLDSSPEHCGAG			
Selected sequences				
>FtzG1	>FtzG6	>FtzG11	>FtzH5	>FtzH10
ATCATAATTG	TTACTAATGA	CATCATTAAC	GAGTTAATGA	ATATTAATTA
>FtzG2	>FtzG7	>FtzG12	>FtzH6	>FtzH11
CAGCCGCCC	TTATAAATGA	TGCTTAATTA	GCCCTAAGAT	TTGTTAATGA
>FtzG3	>FtzG8	>FtzH1	>FtzH7	
TCCCATTAAC	TTTTTAATTG	GGCTTAATGG	CTGTTAATTA	
>FtzG4	>FtzG9	>FtzH2	>FtzH8	
CTCTTAATTA	CGCCTAATGA	GAACCTACTT	GGGTTAATTA	
>FtzG5	>FtzG10	>FtzH3	>FtzH9	
ATGCTCCCGC	TAGTTAATTA	CCGTTAATTA	TTTTTAATGA	

Gsc				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KRRHRTIFTEEQLEQLEATFDKTHYPDVVLREQLALKVDLKEERVEVWFKNRRAKWRKQKREEQERLRKLQEE			
Selected sequences				
>GscA1	>GscA6	>GscA11	>GscB5	>GscB10

CCGATTACGA	TCAGATTATC	TAGGATTATG	TCGGATTAAG	ATGGATTAGT
>GscA2	>GscA7	>GscA12	>GscB6	>GscB11
CAAGTAATCC	TAGGATTACT	ACAATAATCC	ACGGATTAAA	GGATTAATG
>GscA3	>GscA8	>GscB1	>GscB7	
AAGATTAGTC	TGCGATTAAG	TCGTTAATCT	CCCCTAATCC	
>GscA4	>GscA9	>GscB3	>GscB8	
TAGGATTATT	ATCGTAATCC	GGGATTAACA	GGGATTAACA	
>GscA5	>GscA10	>GscB4	>GscB9	
AAGATTAGTA	TCGTTAATCT	CGAGATTAAG	TGGATTAGGA	

H2.0				
Promoter-Stringency	UV5-10mM			
Amino Acid Sequence	RSWSRAVFSNLQRKGLEIQFQQQKYITKPDRRKLAARLNLTDAQVKVWFQNRRMKWRHTRENLKSGQEKQPSA			
Selected sequences				
>H20C1	>H20C7	>H20C12	>H20D5	>H20D10
TTTTTATATA	CCTTTATGGG	GGGTAATTAG	CTGTAATAAA	CGTTTATTAA
>H20C2	>H20C8	>H20D1	>H20D6	>H20D11
CTCTTAATGA	CGTTGATTAA	CTATTATTAA	ACCTCATAAT	ATTTTATGAG
>H20C3	>H20C9	>H20D2	>H20D7	
ACATTATTGA	AGTCAATAAA	TGAATATTGA	AATTTATTAA	
>H20C4	>H20C10	>H20D3	>H20D8	
GGTTAATGAT	GATTAATTAT	GCTTAATTGA	ATCTAATTAA	
>H20C5	>H20C11	>H20D4	>H20D9	
TAGATATTAC	GGATAATTAA	AGATAATAAT	TGCTAATAAA	

Hbn				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KVRRSRTTFTTFQLHQLERAFEKTQYPDVFTREDLAMRLDLSEARVQVWFQNRRAKWRKREKFMNQDKAG			
Selected sequences				
>HbnA1	>HbnA8	>HbnB4	>HbnB10	
TTAATTATT	CTCTAATTGA	TACAATTAAT	GCTAATTAAC	
>HbnA4	>HbnA9	>HbnB5	>HbnB11	
AGTAATTACC	CCTTAATTAA	CTTAATTAAA	CCAATTAAT	
>HbnA5	>HbnA10	>HbnB6		
TTCAATTACA	ACAAATTAAG	CTTAATTATC		
>HbnA6	>HbnB2	>HbnB7		
GTCAATTATG	AGACTAATTA	ACCAATTAAA		

>HbnA7	>HbnB3	>HbnB9		
GATTAATTAA	GACAATTAAA	ATTAATTAAA		

Hgtx				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KKHTRPTFSGQQIFALEKTFEQTKYLAGPERAKLAYALGMSESQVKVWFQNRRTKWRKRHAAEMATAKRKQDD			
Selected sequences				
>HgtxC1	>HgtxC6	>HgtxD1	>HgtxD6	>HgtxD11
CCCTATTAAT	TCATTAAAAT	AGCAATTATC	GCTTTAATTA	AGTTTAATTA
>HgtxC2	>HgtxC7	>HgtxD2	>HgtxD7	
AGTAATTGAA	CATGTAATTA	CGTAATTAGA	CTCATTAAA	
>HgtxC3	>HgtxC8	>HgtxD3	>HgtxD8	
AGCTAATTA	ATCAATTAAC	CTTCATTATT	ACTTTAATTA	
>HgtxC4	>HgtxC9	>HgtxD4	>HgtxD9	
GCCAATTATC	CCTCATTATG	TTAATTAATT	GTTTTAATGA	
>HgtxC5	>HgtxC10	>HgtxD5	>HgtxD10	
TTTGTAATTA	CTTCATAC	TTTTTAATTA	TAGTTAATGA	

Hmx				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KKKTRTVFSRAQVFQLESTFDLKRYLSSSERAGLAASLRLTETQVKIWFQNRRNKWKRQLAAELEAANMANMA			
Selected sequences				
>HMXA2	>HMXA10	>HMXE5	>HMXE12	>HMXF9
ATGTTAATTG	TCCTGACAGT	TTATTAATCG	AGTTCTTAAG	GGGAATGAG
>HMXA3	>HMXA11	>HMXE6	>HMXF1	>HMXF10
AAGATAATTG	AGGAAATGAG	CACTTAATTA	TTGTCTTCGA	CGGAGAATTT
>HMXA4	>HMXA12	>HMXE7	>HMXF2	>HMXF11
ACGCTAATTG	TTAATTG	TAGTTAATTA	TCACGCCGAG	TCGTCATGCG
>HMXA5	>HMXE1	>HMXE8	>HMXF4	
AGCCATTTAA	ACGTTAATTA	CCACTAATTA	TGCAATTAAG	
>HMXA6	>HMXE2	>HMXE9	>HMXF6	
ATATTAATTG	GACTTAATCG	TGTTTAATTG	ATCAGTCCTT	
>HMXA7	>HMXE3	>HMXE10	>HMXF7	
CCCTTAATTG	CCTCTAATTG	TATTTAATTG	ATGAACTTGA	
>HMXA8	>HMXE4	>HMXE11	>HMXF8	
GTGTTAATTG	CACTTAATCG	ATCACTTGCA	GGCAATTAAG	

Hth				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	NQKKRGIFPKVATNILRAWLFQHLTHPYPSEDQKKQLAQDTGLTILQVN NWFINARRRIVQPMIDQ			
Selected sequences				
>hthC2	>hthC8	>hthD1	>hthD6	>hthD11
CTTGTGACAG	TGTCATC	CCTGTCACTG	AGCTGTCATT	TTTCGTGACA
>hthC3	>hthC9	>hthD2	>hthD7	
ATATGTCAAA	GACTGTCAGC	AGCTGTCAAA	TCACGCCGAG	
>hthC4	>hthC10	>hthD3	>hthD8	
ACGTCAAGG	AGACCTGACG	TATTTGACAT	ATACCCGTGC	
>hthC5	>hthC11	>hthD4	>hthD9	
CCTGTCACAG	GCGATGACAG	GAGGTGACAG	TGTCATTGTA	
>hthC6	>hthC12	>hthD5	>hthD10	
CTGTGACGT	TGTCAAACTC	GAAGAAACGT	ACGCTCGAGC	

Ind				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	SKRIRTAFTSTQLLELEREFSHNAYLSRLRRIEIANRLRLSEKQVKIWFQN RRVKQKKGGSESPTFNLSTNSN			
Selected sequences				
>IndC1	>IndC6	>IndC11	>IndD5	>IndD10
GGATTAATTA	GGTAATTAGA	GTAATTAATA	TTGCTAATTA	TCCCTAATTA
>IndC2	>IndC7	>IndC12	>IndD6	>IndD11
CGCTAATGA	ACGTTAATGA	CTCATTAACA	GTCCTAATTA	CTAATTAGCA
>IndC3	>IndC8	>IndD1	>IndD7	
TACCTAATGA	TGATTAATGA	CTAATTAAGG	TCACTAATGA	
>IndC4	>IndC9	>IndD3	>IndD8	
CACTTAATAG	GCGTTAATGA	GTAATTAGAA	GCAATTAATA	
>IndC5	>IndC10	>IndD4	>IndD9	
TAATTAGAG	GTAATTAGTA	ACACTAATGA	CCACTAATTA	

Inv				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	EDKRPRTAFSGTQLARLKHEFNENRYLTEKRRQQLSGELGLNEAQIKIWF QNKRAKLKKSSGTKNPLALQLMAQ			
Selected sequences				
>InvC1	>InvC7	>InvC12	>InvD5	>InvD10
ATAATTAACC	GGTAATTATA	CCTAATTAAA	GTAATTAGTA	TCACGCCGAG

>InvC2	>InvC8	>InvD1	>InvD6	>InvD11
ACTAATTAAT	TCTAATTAAA	CCAATTAAAT	GATGCTAAAC	GGTAATTAAC
>InvC3	>InvC9	>InvD2	>InvD7	
ATAATTAGCA	AGCCCTCGCA	TCAATTAGAG	CCAATTAGTT	
>InvC5	>InvC10	>InvD3	>InvD8	
CCACTAATTA	AGTCAGCATG	TCAATTAAAA	TCAATTAAAA	
>InvC6	>InvC11	>InvD4	>InvD9	
CTTCACTGAA	TAATTAGAG	CTAATTAGAA	ATTCCGCTCT	

Lab				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	NNSGRTNFTNKQLTELEKEFHFNRYLTRARRIEIANTLQLNETQVKIWFQNRRMKQKKRVKEGLIPADILTQH			
Selected sequences				
>LabE1	>LabE8	>LabF4	>LabF11	
TCATTAACGA	TAGTTAATTA	CTTCACTGAA	AGTCTAATGA	>LabH10
>LabE2	>LabE9	>LabF5	>LabH3	TATCGCCCAC
GCCTTAATTA	ATACTAATTA	CTGTTAATTA	GCTGTTATTT	>LabH11
>LabE3	>LabE10	>LabF6	>LabH5	AATGATCGTC
TGGCTAATTA	CGTTCTTTAA	AATGATCGTC	GTGCGCGCAG	>LabH12
>LabE4	>LabE11	>LabF7	>LabH6	TACATAATGA
GATAATTAAT	GCTTGATGCG	GTCCAGATTG	GTGTTAATTA	
>LabE5	>LabF1	>LabF8	>LabH7	
CATACCCAGA	AGTCATTAAG	CGTTAATT	GTCTTAATTA	
>LabE6	>LabF2	>LabF9	>LabH8	
AATTTAATTA	CTACCAGATT	GGTCATTAAT	TCACGCCGAG	
>LabE7	>LabF3	>LabF10	>LabH9	
GCTAATTAAT	GTAGCCAATG	CAGGCACCCA	CTACTAAATT	

Lag1				
Promoter-Stringency	UV5-5mM			
Amino Acid Sequence	IRSSRPKKAANVPILEKTYAKSTRLDKKKLVPLSKQTDMSEREIERWWRLRRAQDKPSTLVKFCENTWRC			
Selected sequences				
>5Lag1G1	>5Lag1G11	>5Lag1H7	>5Lag12C8	>5Lag12D4
GGACCTGAAG	AAGTTATTAG	CTACCAAGAT	GAGACGAGAT	CCACTAGATT
>5Lag1G4	>5Lag1G12	>5Lag12C1	>5Lag12C9	>5Lag12D5
GTAGAAGAGG	TCTCTCGCCT	GTGGTATGCT	CCACCATAAT	GCGCATGAGA
>5Lag1G5	>5Lag1H1	>5Lag12C2	>5Lag12C10	>5Lag12D7

TCCCCCATCC	TGACTATCAG	TCCCTCACTG	TAGTAATATT	CTACTAAAAC
>5Lag1G6	>5Lag1H2	>5Lag12C3	>5Lag12C11	>5Lag12D8
CCACCAACTT	CATCGCTATG	TCGTACGAAT	CTACCATGTT	CTACTAGTAT
>5Lag1G7	>5Lag1H3	>5Lag12C4	>5Lag12C12	>5Lag12D9
CTACCAAGAA	CCCCTAAATT	ACTACAAAGG	ACTCTGCTTA	CCACCAAAAT
>5Lag1G8	>5Lag1H4	>5Lag12C5	>5Lag12D1	>5Lag12D10
AGGTCAGATA	CTACCATAAA	CTACCAGAAG	CCCCTGGAAT	CTACCAACCT
>5Lag1G9	>5Lag1H5	>5Lag12C6	>5Lag12D2	>5Lag12D11
CCACTAATTC	CCACCAAATT	GGTATACGAC	TGTGATATAG	CCTCCCGCAC
>5Lag1G10	>5Lag1H6	>5Lag12C7	>5Lag12D3	
TGTACAACAG	CTACCAAGAT	CCACCAAATA	TACAAATAT	

Lbe				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KRKSRTAFTNHQIFELEKRFLYQKYLSPADRDEIAASLGLSNAQVITWFQ NRRAKQKRDIEELKKDFDSVK			
Selected sequences				
>IbeG2	>IbeG7	>IbeG12	>IbeH5	>IbeH10
GATGATTATG	GGTAATTACC	TCGATTAACTA	GCTAATTAAT	CCTCGTTAAA
>IbeG3	>IbeG8	>IbeH1	>IbeH6	>IbeH11
CATCTAAGTA	GCTATAAGTA	GCACTAACAA	ACCAATTAAC	CCTCGTTAAG
>IbeG4	>IbeG9	>IbeH2	>IbeH7	
AGGTTAACCA	TCTTGTTACA	CCTCGTTAGT	GCTAATTAAC	
>IbeG5	>IbeG10	>IbeH3	>IbeH8	
TCCTAATCAC	GCTGGTTAAC	CACATAATCA	TTGTTATC	
>IbeG6	>IbeG11	>IbeH4	>IbeH9	
CGTTAAATGA	CGCCTAATTA	GCTTTAACAA	ACCAATTAAC	

Lbl				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KRKSRTAFTNQQIFELEKRFLYQKYLSPADRDEIAGGLGLSNAQVITWFQ NRRAKLKRDMEELKKDVQCEKMS			
Selected sequences				
>LblC1	>LblC6	>LblC11	>LblD4	>LblD9
GATAATTATC	GTGCTAATTG	CTTGTTAAC	ACCAATTATC	CGACTAATGA
>LblC2	>LblC7	>LblC12	>LblD5	>LblD10
GGTAATTATA	GTTAATTAAA	CCTCGTTAAG	CGACTAATGA	CCTCGTTAAG
>LblC3	>LblC8	>LblD1	>LblD6	>LblD11
AGGATAATTG	AAACTAACGA	ACCAATTAAA	CTGCTAATCA	CGACTAATGA

>LblC4	>LblC9	>LblD2	>LblD7	
CCACTAATCA	CCTGATTAAG	GCTCGTTAAG	GCTAATTAAT	
>LblC5	>LblC10	>LblD3	>LblD8	
GCTAATTATT	CCAATTATC	ACCAATTAAG	GCTAATTAAT	

Lim1				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RRGPRTTIKAKQLEVLKTAFNQTPKPTRHIREQLAKETGLPMRVIQVWFQ NKRSKERRMKQITSMGRPPFFGG			
Selected sequences				
>LimC1	>LimC10	>LimD7	>Lim1C4a	>Lim1D4a
TAGATTACGT	GTTAATTAGC	GAACCAAAGG	TCGTTTTGAA	AATTAATTAT
>LimC2	>LimC11	>LimD8	>Lim1C5a	>Lim1D6a
CAACAGTGGG	GTTAATTGA	CTACTAATAT	TGTCCAAAT	TTTTAATTAA
>LimC3	>LimC12	>LimD9	>Lim1C6a	>Lim1D9a
AAGCAAGCCG	TTAATTAGT	AGTTCACGGG	CGGGGAACGT	ACTAATTAAA
>LimC4	>LimD1	>LimD10	>Lim1C9a	>Lim1D10a
AGCAGAATCG	TACTAATTA	GCTAATTAAT	GCCTAATTAA	TGCTAATTAA
>LimC5	>LimD2	>LimD11	>Lim1C10a	>Lim1D11a
TTTTGTGATA	ATTAATTAGA	AGTTAATTAA	GTTAATTACC	TTTAATTACA
>LimC7	>LimD3	>Lim1C1a	>Lim1C11a	
CCACCAGATT	TGGTAATTAA	CCCCCAATTT	CCGTGTTAGT	
>LimC8	>LimD4	>Lim1C2a	>Lim1C12a	
CCCAGTAAGG	TTATAATTAA	CTATAATTAG	ATTATACTGT	
>LimC9	>LimD6	>Lim1C3a	>Lim1D2a	
CAATCGTCTA	GTCCAGATTG	AAAAGGATTC	TACTAATTAA	

Lim3				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PNKRPRTTITAKQLETLKTAYNNSPKPARHVREQLSQDTGLDMRVVQV WFQNRRAKEKRLKKDAGRTRWSQY			
Selected sequences				
>Lim3C1	>Lim3C7	>Lim3D1	>Lim3D6	>Lim3D11
CCCCTGATTA	AAACAAATTA	TTTCATTAAA	GTCAATTACA	CTAATTAGAT
>Lim3C2	>Lim3C8	>Lim3D2	>Lim3D7	
ATTTATTAAG	TTTCATCAGA	ATTAACA	TCAATTAAGA	
>Lim3C3	>Lim3C10	>Lim3D3	>Lim3D8	
CGCTTAATCA	TTAATTTTTA	CTCATTAGAT	ACTAATTAAC	
>Lim3C4	>Lim3C11	>Lim3D4	>Lim3D9	

ATAATTATTC	TTTCATTTAG	CAATTACC	ATTAATCAAC	
>Lim3C5	>Lim3C12	>Lim3D5	>Lim3D10	
ATAATCATTA	TCTAATTTAT	CTCAATTAAG	TAATTAGCC	

Mirr				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	NGARRKNATRETTSTLKAWLNEHKKNPYPTKGEKIMLAIITKMTLTQVS TWFANARRRLKKENKMTWEPRNRVDDD			
Selected sequences				
>MirrE1	>MirrE10	>MirrF7	>Mirr2C7	>Mirr2D4
ACTTGTAACA	CCAAAAAACA	CGAATTTACA	AGATTTAACA	GTGTTAAACA
>MirrE2	>MirrE11	>MirrF8	>Mirr2C8	>Mirr2D5
TTCAGTAACA	TAAGAAAACA	GCTCAAAACA	GCTTGTAACA	CGCATAAACA
>MirrE3	>MirrE12	>MirrF9	>Mirr2C9	>Mirr2D6
AGTAAATACA	GCAAAAAACA	CTAGAAAACA	AGAGAAAACA	TAAGTTACA
>MirrE4	>MirrF1	>MirrF10	>Mirr2C10	>Mirr2D7
ACAGCAAACA	TCACGCCGAG	GCACAAAACA	ACATTTACA	CCTGAAAACA
>MirrE5	>MirrF2	>MirrF11	>Mirr2C11	>Mirr2D8
TGCTATAACA	GATACTTACA	CAAAATAACA	CCTGAAAACA	CTAGTTTACA
>MirrE6	>MirrF3	>Mirr2C2	>Mirr2C12	>Mirr2D9
GGTAATAACA	CAAGAAAACA	TGCAGAAACA	GTGATATACA	TTAGAAAACA
>MirrE7	>MirrF4	>Mirr2C3	>Mirr2D1	>Mirr2D10
ACTCGTGACA	CTGAAAAACA	CTAACAAACA	ATCTCTTACA	CTCAATAACA
>MirrE8	>MirrF5	>Mirr2C4	>Mirr2D2	>Mirr2D11
GGAAGTTACA	TTTGAAAACA	CTGTACTACA	GGAAATTACA	TTGAATAACA
>MirrE9	>MirrF6	>Mirr2C5	>Mirr2D3	
AGAACTTACA	ACGCCTGACA	ACTTCGAACA	GGATATAACA	

NK7.1				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RKKKARTTFTGRQIFELEKMFENKKYLSASERTEMAKLLMVTETQVKIW FQNRRTKWKKQDNVTNNEAAEHK			
Selected sequences				
>NK7E1	>NK7F1	>NK7F10	>NK72A9	>NK72B7
CATATAATGA	TAGCTAATTG	CGCTTAATTA	GACATAATAG	GTATTAATGA
>NK7E2	>NK7F2	>NK7F11	>NK72A10	>NK72B8
GAATTAAGTG	GCCTAGGGAG	TGACTAATTA	GAGATAATGA	CGTAATTAAG
>NK7E3	>NK7F3	>NK72A2	>NK72A11	>NK72B9
GAATTAAGTG	GAAGCCGCAA	ATCATTAAAC	GCTATCAATA	CCACTAATTA

>NK7E4	>NK7F4	>NK72A3	>NK72A12	>NK72B10
GGTATTTAAA	GCGCTAATGA	TACTTAAGTG	ACATTAAATG	GCCTATTAAA
>NK7E6	>NK7F5	>NK72A4	>NK72B2	>NK72B11
AAAATAATTA	CTTAGATTTT	CGCCGTTCAG	TGGTTAAATA	CCAATTAAGG
>NK7E7	>NK7F6	>NK72A5	>NK72B3	
GAGTAAATGA	CAATTAAAA	GTCCATTAAA	GCCGATGACT	
>NK7E8	>NK7F7	>NK72A6	>NK72B4	
AAACTAATTG	TCCTTAATAG	TGCTATTAAG	GAATTAAGTG	
>NK7E10	>NK7F8	>NK72A7	>NK72B5	
GCCACTTAAC	GGCAATTAAG	GACAATTATA	ACTATTAATA	
>NK7E12	>NK7F9	>NK72A8	>NK72B6	
TGTCCAAAT	ACCAATTAAG	TTAATTAAAA	AACCATCAAT	

oc				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	QRRERTTFTRAQLDVLEALFGKTRYPDIFMREEVALKINLPESRVQVWF KNRRAKCRQQLQQQQQSNSLSSSK			
Selected sequences				
>OcA1	>OcA6	>OcA11	>OcB5	>OcB10
CCACTAATC	TATATAATCC	GCAGATTAAC	CTGGATTAAG	TCGGATTAAG
>OcA2	>OcA7	>OcB1	>OcB6	>OcB11
AGCTTAAGCC	TTTGCTAATC	GTGGATTAAT	AGGGATTATA	CCGGATTAAC
>OcA3	>OcA8	>OcB2	>OcB7	
CGATAATCCC	CATTAATAAC	TTCATAATCC	GAGTTAATCC	
>OcA4	>OcA9	>OcB3	>OcB8	
GGGGCTTAAA	CGCGGATTAG	GAGGATTACG	CTGGATTAGT	
>OcA5	>OcA10	>OcB4	>OcB9	
GAGGATTATT	AGGATTAAGG	AGCGATTAAG	AGGATTAAT	

Oct				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RRKKRTSIETNIRVALEKSFLENQKPTSEEITMIADQLNMEKEVIRVWFCN RRQKEKRINPPS			
Selected sequences				
>oct1	>oct6	>oct13	>oct18	>oct23
GCGATAATGA	GAGATAATTT	CAGCTAATTA	GCTCATTAAC	GCTTTAATTT
>oct2	>oct7	>oct14	>oct19	>oct24
GGTTTAATGA	GATATAATAA	GCATTAATTT	CTCATAATCA	CTGATAATTA
>oct3	>oct8	>oct15	>oct20	

GTTTTAATAA	ACATTAATCA	TGGTAAATGA	AGTTTAATTG	
>oct4	>oct10	>oct16	>oct21	
CACATAATTT	TGTGATTAAA	TCTCATTAAA	GAGATAATTC	
>oct5	>oct11	>oct17	>oct22	
CTATAATTAG	TCTCATTAGG	GCCCTAATTA	GGTTTAATTT	

Odsh				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KKRRGRTNFNSWQLRELERVFQGSHYPDIFMREALATKLDLMEGRIAV WFQNRRAKWRKQEHTKKGPGRPAHNA			
Selected sequences				
>OdshC1	>OdshC7	>OdshC12	>OdshD5	>OdshD10
GTTGTAATTA	AGCTAATTA	GACAATTAGG	CCTAATTAAA	CTTAATTAAC
>OdshC2	>OdshC8	>OdshD1	>OdshD6	>OdshD11
GAAATTAAA	CCTAATTACA	TACAATTAAT	TCTAATTAAG	CTCAATTACC
>OdshC3	>OdshC9	>OdshD2	>OdshD7	
ATAATTAGCA	GCAATTAAGG	CCAATTAAAT	CCGCTAATTA	
>OdshC4	>OdshC10	>OdshD3	>OdshD8	
TACTAATTAA	AGTAATTACC	TGCTAATTAA	GTCAATTAGT	
>OdshC5	>OdshC11	>OdshD4	>OdshD9	
TAACTAATTG	GATTTAATTA	GACTTAATTA	TTCAATTACA	

onecut				
Promoter-Stringency	UV5-5mM			
Amino Acid Sequence	PKKPRLVFTDLQRRTLQAIFKETKRPSKEMQVTIARQLGLEPTTVGNFFM NARRRSMDKWRDDDSKSTMHVAH			
Selected sequences				
>5onectG1	>5onectG6	>5onectG12	>5onectH5	>5onectH10
GAGCTGATTA	GGCCAAGTCA	GTCAATCAGA	AGTCTTGATTT	CCCAATCAAA
>5onectG2	>5onectG7	>5onectH1	>5onectH6	>5onectH11
CGAAATCAAG	CGGTTTGCAA	AAGTTGATTG	GGAGAGCCTT	CTAATCAACG
>5onectG3	>5onectG8	>5onectH2	>5onectH7	
GCAATCAAGG	GCTATGATTA	ACAAATCAAT	GCAAATCAAC	
>5onectG4	>5onectG10	>5onectH3	>5onectH8	
TTAATCAATA	TTAATCAATA	CTGGGCTATA	GTTATGCGTC	
>5onectG5	>5onectG11	>5onectH4	>5onectH9	
TACCAACAA	CCAAATCAAT	CCAAATCAAG	CCACCATTTT	

Optix	
Promoter-Stringency	UV5m-10mM

Amino Acid Sequence	GEQKTHCFKERTRSLLREWYLQDPYPNPTKKRELAKATGLNPTQVGNWFKNRRQRDRAAAAKNRIQHSQNSSG			
Selected sequences				
>11617C2	>11617D1	>OptixE2	>Optix2A2	>Optix2B2
GAACCTACTT	GTAGTGCTAG	GTAGCCAATG	GCCCATGATA	TTTCTGCGTG
>11617C3	>11617D2	>OptixE3	>Optix2A3	>Optix2B3
GCGCATGAGA	CCGTCTAAAC	CGTTCTTTAA	ACATGTGATA	CATTGCGATA
>11617C4	>11617D3	>OptixE4	>Optix2A4	>Optix2B4
TGAAGTGATA	TAATGCACAC	CCCTAACATG	GGAGCTGATA	GTGATTGATA
>11617C5	>11617D5	>OptixE5	>Optix2A5	>Optix2B5
CAACGTATT	CTTTTCATCT	GAACCTACTT	CAATCTGATA	CCAAGTGATA
>11617C6	>11617D7	>OptixE6	>Optix2A6	>Optix2B6
CAATGTGATA	AGAAACTATG	ATCTTAATTAC	GTAGCTGATA	ATAAGTGATA
>11617C7	>11617D8	>OptixE7	>Optix2A7	>Optix2B7
ACCAGTGATA	ATAAATGATA	GTGCGTACTG	TACCCACGCC	CGTTATGATA
>11617C8	>11617D9	>OptixE8	>Optix2A9	>Optix2B8
GATTGCGATA	GTTAGTGATA	GGAAGTGATA	GAAGTGATAG	CTTTCTGATA
>11617C9	>11617D10	>OptixE9	>Optix2A10	>Optix2B9
CATCGCTATG	ATAAGTGATA	ATCTTATTAC	AACCGCGATA	ATTCAAACA
>11617C10	>11617D11	>OptixE10	>Optix2A11	
ACGTATTGGT	CGTAGTGATA	CTACTAAATT	GTGATTGATA	
>11617C11	>11617D12	>OptixE11	>Optix2A12	
CGTAGTGATA	AGAGATGATA	ATCAGTCCTT	TCCCTTGATA	
>11617C12	>OptixE1	>OptixE12	>Optix2B1	
ATCTTATTAC	TCTTCCATTA	CTACCAGATC	ATTCGCGATA	

Otp				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	QKRHRTRFTPAQLNELERCFSKTHYPDIFMREEIAMRIGLTESRVQVWFQNRRAKWKKRKKTTNVFRTPGA			
Selected sequences				
>OTPG1	>OTPG6	>OTPG12	>OTPH5	
TTCAATTATG	ACCATTAATT	CCTAATTAGG	TCAATTAAGG	
>OTPG2	>OTPG8	>OTPH1	>OTPH6	
CAAATTAGAA	TTCGCTAATTT	CCTTAATTAA	ATTAATTAAA	
>OTPG3	>OTPG9	>OTPH2	>OTPH7	
AGCAATTAAT	CTCATTAAAC	CTAATTACAG	ATTAATTAGA	
>OTPG4	>OTPG10	>OTPH3	>OTPH10	

GTTTAATCA	TCATTAAAG	GCATAATTA	TAACTAATTA	
>OTPG5	>OTPG11	>OTPH4	>OTPH11	
ATTGTAATTA	TTTAATTAAT	ATTCATTAAC	TTAATTAGAC	

pb				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIAASLDLTERQVKVWFQNRRMKHKRQTLSKTDDEDNKDS			
Selected sequences				
>Pb1	>pBG3	>pBG8	>pBH3	>pBH8
GGTCATTAGA	CTCATTAAA	TGACTAATGA	GGTAATTATA	AGGTTAATGA
>Pb2	>pBG4	>pBG9	>pBH4	>pBH9
GGTAATTAAC	GGTAATTATA	GCGTTAATGA	GATAATTATC	TCATTAATGA
>Pb3	>pBG5	>pBG11	>pBH5	>pBH10
TGTAATTAAA	CTGTTAATTA	GCTCATTAAG	CCAGCAAGAT	CCTCATTAGA
>Pb4	>pBG6	>pBG12	>pBH6	>pBH11
GGTCATTAAC	CGTAATTAAT	GCTAATTAAT	TTGCTAATGA	GGTAATTAGA
>pBG2	>pBG7	>pBH1	>pBH7	>pBH12
ATCCTAATTA	TACATAATGA	TGTAATTAAA	GCTAATTAAG	TTGCTAATTA

PhdP				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KQRRIRTTFTSNQLNELEKIFLETHYPDIYTREEIASKLHLTEARVQVWFQNRRAKFRKQERHAIYIMKDKS			
Selected sequences				
>PhdPA1	>PhdPA8	>PhdPB1	>PhdPB6	
ATTAATTTGT	TGCATAATTT	TTAAATTATC	CTTAATTATC	
>PhdPA2	>PhdPA9	>PhdPB2	>PhdPB7	
GGTTAATTAC	TTAATAATTG	ATCAATAAAA	TTTAATTAAT	
>PhdPA3	>PhdPA10	>PhdPB3	>PhdPB10	
CCTATTAGC	CCTAATTGG	TCCGCTAATTT	TACCTAATTA	
>PhdPA4	>PhdPA11	>PhdPB4	>PhdPB11	
GCTCATTAGG	GAAATAATTA	CGGAATTAAG	GCCATGGAT	
>PhdPA6	>PhdPA12	>PhdPB5		
GCTGCTAATT	CGATTAATTA	CTAATTATA		

Pph13	
Promoter-Stringency	UV5m-10mM
Amino Acid Sequence	KQRRYRTTFNTLQLQELERAFQRTHYPDVFFREELAVRIDLTEARVQVWFQNRRAKWRKQEKIGGLGGDYKEGA

Selected sequences				
>Pph13C1	>Pph13C6	>Pph13C11	>Pph13D4	>Pph13D9
CTAATAATTA	CATATAATTA	GTCCTAATTA	AGTAATTACA	ACTAATTAAA
>Pph13C2	>Pph13C7	>Pph13C12	>Pph13D5	>Pph13D11
GTGATAATTG	CATTTAATTA	GCCAATTACC	GTCAATTAAA	ATTAATTAAA
>Pph13C3	>Pph13C8	>Pph13D1	>Pph13D6	
TCTAATTTAA	CCTGATTAGT	AATAATTAGC	ACTAATTATA	
>Pph13C4	>Pph13C9	>Pph13D2	>Pph13D7	
GCCATAATTA	AACAATTATG	AACAATTAGC	GTTAATTAAT	
>Pph13C5	>Pph13C10	>Pph13D3	>Pph13D8	
CCGAATTAGT	TCCAATTACT	AGCCTAATTA	CAGTGTTAAT	

Ptx1				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	QRRQRTHFTSQQLQELEHTFSRNRYPDMSTREEIAMWTNLTEARVRVW FKNRRAKWRKRERNAMNAAVAAAD			
Selected sequences				
>PtxG1	>PtxG8	>PtxH1	>PtxH6	>PtxH11
CATCTAATCC	GGGGATTAAC	GCTAATCCTC	CGTTAATCCC	ACTTAATCCC
>PtxG2	>PtxG9	>PtxH2	>PtxH7	
TTTAATCCCT	TGGGATTAAC	CTCTTAATCC	GGTTAATCCC	
>PtxG4	>PtxG10	>PtxH3	>PtxH8	
AGAGGATTAG	CAATTAATCC	CTTACTGTTT	TCTTAATCCC	
>PtxG6	>PtxG11	>PtxH4	>PtxH9	
CAGGATTAGT	AGGCTAATCC	CGTTAATCTC	CGTTAATCCC	
>PtxG7	>PtxG12	>PtxH5	>PtxH10	
CTTAATCCTA	TCGTAATCCC	CCTTAATCCC	CGTTAATCCC	

Repo				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KKKTRTTFTAYQLEELERAFERAPYPDVFAREELAIKLNLSESRVQVWFQ NRRAKWRKHEPPRKTGYIKTST			
Selected sequences				
>RepoE2	>RepoE9	> Repo2A5	> Repo2B1	> Repo2B8
CCTTTAATTA	CTAATTAATG	TAATTACAT	GTCAATTAAA	GTTAATTAAA
>RepoE3	>RepoE10	> Repo2A6	> Repo2B2	> Repo2B9
AGGGTAATTA	TTGGGTACAA	TCTAATTAAA	CCTAATTAAA	CCAATTAAAA
>RepoE4	>RepoE11	> Repo2A7	> Repo2B3	> Repo2B10

AGCTAATTA	TGCGCATCGA	TCAATAAATA	GGGCTAATTA	TAGTTAATTA
>RepoE5	>RepoE12	> Repo2A8	> Repo2B4	> Repo2B11
CTGTTAATTA	TCCTGACAGT	CGCGTAATTA	CCACTAATTA	TATTTAATTA
>RepoE6	> Repo2A2	> Repo2A9	> Repo2B5	
GCAAACCCCT	AGCAATTTAA	TAGCTAATTA	CACTTAATTA	
>RepoE7	> Repo2A3	> Repo2A11	> Repo2B6	
TCAATTAAGA	CCAATTACTA	GATTTAATTG	AGTTTAATTA	
>RepoE8	> Repo2A4	> Repo2A12	> Repo2B7	
CTTTTAATTA	GCTATAATTA	CCATTAATTA	TCATTAATTA	

Ro				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	QRRQRTTFSTEQTLRLEVEFHRNEYISRSRRFELAETLRLTETQIKIWFQN RRAKDKRIEKAQIDQHYRN			
Selected sequences				
>RoG1	>RoG6	>RoG11	>RoH4	>RoH9
GGTAATTACC	CTAATAATTA	TAGTTAATTA	TCGCTAATTA	GCTAATTAAT
>RoG2	>RoG7	>RoG12	>RoH5	>RoH10
AGCAATTAGC	CTTTAATTA	GTGTTAATTA	GCTAATTAAT	GCGGTAATTA
>RoG3	>RoG8	>RoH1	>RoH6	>RoH11
TTACTAATGA	TTACTAATGA	GTCAATTAAA	GCTAATTAAT	GGCAATTAAG
>RoG4	>RoG9	>RoH2	>RoH7	
TGTAATTAAA	CCACTAATTA	ACTAATTAAA	GCTAATTAAT	
>RoG5	>RoG10	>RoH3	>RoH8	
GTCAATTAGG	GTGCTAATTA	TCGTAATGA	GTCTTAATTA	

Rx				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	HRRNRTTFTTYQLHELERAFEKSHYPDVYSREELAMKVNLPEVRVQVW FQNRRAKWRRQEKSESLRLGLTHFT			
Selected sequences				
>RxF2	>RxF9	> Rx2C5	> Rx2C12	> Rx2D8
GGTAATTAGC	TGCCAATATA	TACTTAATTA	GTCCAGATTG	GCTAATTAAA
>RxF3	>RxF11	> Rx2C6	> Rx2D1	> Rx2D9
CTACTAATTA	TTGGGTACAA	CCAATTAGTG	AGCAATTAAA	ATTAATTAGC
>RxF4	>RxF12	> Rx2C7	> Rx2D2	> Rx2D10
CTACTAATTA	GCTAATTAAT	CCTAATTATG	GCTAATTAGA	GCTAATTAAC
>RxF5	>Rx2C1	> Rx2C8	> Rx2D3	> Rx2D11
CTAATTAACC	TAATTATGA	GTGTTAATTA	AATAATTAGA	CCCAATTAAT

>RxF6	> Rx2C2	> Rx2C9	> Rx2D4	
CTACTAACAT	ACTAATTAGA	GTCAATTAAT	ACCAATTAAG	
>RxF7	> Rx2C3	> Rx2C10	> Rx2D5	
GCCAATTAAC	CTTAATTAAC	ATCAATTATG	GCTAATTAAG	
>RxF8	> Rx2C4	> Rx2C11	> Rx2D6	
GTCAATTAAA	TATCAACCCC	GTCAATTAGC	GTCAATTAAA	

Scr				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	TKRQRTSYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQN RRMKWKKEHKMASMNIVPYHMG			
Selected sequences				
>ScrE04	>ScrF01	>ScrF11	>ScrG7	>ScrH3
CCTTAATGA	GGAAAGTGGA	TCATTAATGA	CTGTTAATTA	ACGTTAATGA
>ScrE05	>ScrF02	>ScrF12	>ScrG8	>ScrH4
TACATAATGA	CGATAATGA	TCGTTAATGA	CACTAATTA	CTATTAATGA
>ScrE06	>ScrF04	>ScrG1	>ScrG9	>ScrH7
TTGGGTACAA	GACTTAATGA	GCAATTAAAG	CCGTTAATTA	TTGGGTACAA
>ScrE08	>ScrF06	>ScrG2	>ScrG10	>ScrH9
CTACTAATTA	ACGCTAATGA	GAATTAATGA	GTCCAGATTG	GTACTAATGA
>ScrE09	>ScrF07	>ScrG3	>ScrG12	>ScrH10
ATAGGTCCGT	TACACACAGC	GGTTTAATGA	GCGCTAATGA	TATTTAATGA
>ScrE11	>ScrF09	>ScrG5	>ScrH1	>ScrH11
AGTGCTTCAC	GGAAATACGC	CCACTAATTA	TGTTAATGA	TCACTAATGA
>ScrE12	>ScrF10	>ScrG6	>ScrH2	
TCCTTAATGA	CTACTAACTT	GCAATTAACG	GACATAATGA	

Six4 (-)				
Promoter-Stringency	UV5-10mM			
Amino Acid Sequence	DGEETVYCFKEKSRNALKDCYLTNRYPTPDEKKTLAKKTGLTLTQVSN WFKNRRQRDRTPQQR			
Selected sequences				
>5Six4E1	>5Six4E6	>5Six4E11	>5Six4F4	>5Six4F10
GTATCAAATC	GGTATCAAGA	GGTATCATAC	GGTGTCACAG	GTCTCAAATA
>5Six4E2	>5Six4E7	>5Six4E12	>5Six4F5	>5Six4F11
GTCTCAAAGC	CATTGTACTA	GGTGTCAGAC	GGTATCATTA	GTATCACAAA
>5Six4E3	>5Six4E8	>5Six4F1	>5Six4F6	
GTATCAAATG	GTATCATCTT	GGTGTCATGT	GGTGTCATCA	
>5Six4E4	>5Six4E9	>5Six4F2	>5Six4F7	

GGTGTCAGAT	TTCTCAAAAG	GTATCACTTA	GGTATCAAAA	
>5Six4E5	>5Six4E10	>5Six4F3	>5Six4F9	
GTCTCATTTA	GGTGTCAACT	GGTGTCAATT	GTGAGCATGT	

Slou				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PRRARTAFTYEQLVSLENKFKTTRYLSVCERLNLALSLSLTETQVKIWFQ NRRTKWKKQNPGMDVNSPT			
Selected sequences				
>SlouE1	>SlouE6	>SlouE11	>SlouF4	>SlouF9
AATGTCTCAT	TACAATTAGC	GGTCTAATGA	GTTAATTAGT	ACCAATTAAA
>SlouE2	>SlouE7	>SlouE12	>SlouF5	>SlouF10
AGCCATTAAG	GTTAATTATT	TTTCTAATGA	CCCAATTAAG	CGTAATTAAA
>SlouE3	>SlouE8	>SlouF1	>SlouF6	>SlouF11
CTCATTAAAA	ACCGATTAAA	GGGCTAATTA	GACAATTATA	GCCAATTATC
>SlouE4	>SlouE9	>SlouF2	>SlouF7	
CGAGTAATGA	GATTATTAAA	CCTCTAATTG	ATCAATTAAG	
>SlouE5	>SlouE10	>SlouF3	>SlouF8	
CGTCAATTAC	CAATTAATTG	CTATAATTAG	CTACTAAAGT	

So				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	DGEETSYCFKEKSRSVLRDWYSHNPYPSPREKRDLAEATGLTTTQVSNW FKNRRQRDRAAEHKDGSTDKQHL			
Selected sequences				
>SoE1	>SoE8	>SoF5	>So2E1	>So2E8
GTATCACAGT	GTATCAGATC	ATATCCCTCA	GATCGTGATA	ATACGTCAGT
>SoE2	>SoE9	>SoF6	>So2E2	>So2E9
AACGATGATA	TCTAGTGATA	TCGGAATGAA	AATATTGATA	CATATCACAA
>SoE3	>SoE11	>SoF7	>So2E3	>So2E10
AGTATCAATA	GTATCAAAAC	GTATCATTTA	CATATCACTG	AAACATGATA
>SoE4	>SoE12	>SoF8	>So2E4	>So2E11
GGTCATCACT	CAGGATGATA	CGCAGCACAT	TGGCATGATA	TGTGATGATA
>SoE5	>SoF2	>SoF9	>So2E5	>So2E12
AGTATCATTC	GATGATGATA	GCGCATGAGA	GGTATCACAC	CACAATGATA
>SoE6	>SoF3	>SoF10	>So2E6	
AGAAATGATA	TCACGCCGAG	ATCAGTCCTT	AAGGATGATA	
>SoE7	>SoF4	>SoF11	>So2E7	
GTATCACAAT	GTATCATTTT	GTATCACTTA	GGCAGTGATA	

Tin				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KRKPRVLFSQAQVLELECRFRLKKYLTGAEREIIAQKLNLSATQVKIWFQ NRRYKSKRGDIDCEGIAKHLKLK			
Selected sequences				
>TinA2	>TinA7	>TinB1	>TinB7	
GGCTCAAGTA	CGCACTTGAC	CCCACTTGAG	CTTCGCAGTA	
>TinA3	>TinA8	>TinB3	>TinB8	
GGTACTTGAC	AGTACTTAAG	AACACTTAAG	GTCCAGATTG	
>TinA4	>TinA9	>TinB4	>TinB9	
CCACTTGACG	TCCACTTAAG	CCACTTGAA	TCCACTTCAA	
>TinA5	>TinA10	>TinB5	>TinB10	
ATCTCAAGTG	CCCACTTAAA	GCCACTTGAG	CTTCGCAGTA	
>TinA6	>TinA11	>TinB6	>TinB11	
CCACTTGAAT	CCACTTGTGG	TGCACTTGAG	GTCCAGATTG	

Tup				
Promoter-Stringency	UV5m-5mM			
Amino Acid Sequence	PTRVRTVLNEKQLHTLRTCYNANPRPDALMKEQLVEMTSLSPRVIRVWF QNKRCKDKKKTIQMKLQMQQEKEG			
Selected sequences				
>TupE1	>TupE6	>TupE11	>TupF5	>TupF10
ATCAATTACC	CACCATTATA	ACCTTAATGG	GATAGATC	CTATCATAAC
>TupE2	>TupE7	>TupE12	>TupF6	>TupF11
CTAATTAGCG	TTTAATTATC	ACCTTAATGG	ATCCATTAAG	CGTCTCACGA
>TupE3	>TupE8	>TupF1	>TupF7	
ACACTAATGT	ATGATACCAT	TGCATGCATC	GCGATAAGTG	
>TupE4	>TupE9	>TupF2	>TupF8	
TTATTAATGG	ACCACTTAAC	TGCAATTAAG	ATCAATTAAG	
>TupE5	>TupE10	>TupF4	>TupF9	
TTCTATTAAG	TTCAATTATG	TTCCATTTAG	TGACCTCAGT	

Ubx				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RRRGRQTYTRYQTLELEKEFHTNHYLTRRRRIEMAHALCLTERQIKIWF QNRRMKLKKEIQAIKELNEQEKQA			
Selected sequences				
>UbxC02	>UbxC07	>UbxD02	>UbxD08	
TAATTAATTA	TGCAATAAAA	GCGTTAATTA	TCCTTAATGA	

>UbxC03	>UbxC08	>UbxD04	>UbxD09	
AATTTTATTA	GGCAATTAAG	GCCTTAATTA	GCCTTAATTA	
>UbxC04	>UbxC11	>UbxD05	>UbxD10	
GCTTTAATTA	GTATTAATGA	GACAATTAAA	CTTTTTATGA	
>UbxC05	>UbxC12	>UbxD06	>UbxD11	
GCTTTAATTA	AATTTAATGG	CCGTTAATTA	GGTAATTAAC	
>UbxC06	>UbxD01	>UbxD07	>UbxD12	
CTATTAATTA	TCTTAATGA	GCCCATTAAA	TGCAATTAAA	

Unc4				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KRRRSRTNFNSWQLEELERAFSASHYPDIFMREALAMRLDLKESRVAV WFQNRRAKVRKREHTKKGPGRPAH			
Selected sequences				
>Unc4C1	>Unc4C6	>Unc4C11	>Unc4D4	>Unc4D9
TCTAATTAGA	TTCTATTAAG	AGCAATTAAC	GTCAATTAAG	TCCTGACAGT
>Unc4C2	>Unc4C7	>Unc4C12	>Unc4D5	>Unc4D10
TCTGTAATTA	TCAATTAAAA	TTCAATTAGG	GTCAATTAAC	GTCAATTAAC
>Unc4C3	>Unc4C8	>Unc4D1	>Unc4D6	>Unc4D11
GCGAATTAAG	GTCTTAATTG	GCACTAATTA	TCCAATTAAG	ACTAATTAAG
>Unc4C4	>Unc4C9	>Unc4D2	>Unc4D7	
GCTAATTATG	TTCAATTAGG	GTACAACCAA	ACCAATTAAA	
>Unc4C5	>Unc4C10	>Unc4D3	>Unc4D8	
CTTAATTATC	ATCAATTACA	ACCAATTATT	ACCAATTAAC	

Unpg				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KSRRRRTAFTSEQLLELEREFHAKKYLSLTERSQIATSLKLSEVQVKIWFQ NRRAKWKRVKAGLTSHGLGRNGT			
Selected sequences				
>UnpgC1	>UnpgC6	>UnpgC11	>UnpgD5	>UnpgD11
GGTCATTAGT	CTAATTACGC	ACTAATTATC	ACCAATTAAT	CTAATTAATA
>UnpgC2	>UnpgC7	>UnpgD1	>UnpgD6	
TACGTAATTA	AGCAATTAAG	CGCCTAATTA	CTAATTTGGA	
>UnpgC3	>UnpgC8	>UnpgD2	>UnpgD8	
CGCTAATTAG	CTCATTAAA	ATAATTAGCG	GCTAATTAAT	
>UnpgC4	>UnpgC9	>UnpgD3	>UnpgD9	
AAGATAATTG	ATCAATTAAA	TCAATTAAG	TGTAATTATC	
>UnpgC5	>UnpgC10	>UnpgD4	>UnpgD10	

CGTAATTAG	GCAATTAAGG	GCCAATTAAG	CTAATTAAGG	

Vis				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	LRKRRGNLPKSSVKILKRWLYEHRYNAYPSDAEKFTLSQEANLTVLQVCNWFINARRRILPEMIRREGNDPLHFT			
Selected sequences				
>VisG1b	>VisG8b	>VisG3	>VisG10	>VisH5
TAAATGACAC	CTGTCATCGA	ATTATGACAC	TACTGTCAAA	AGTCCCGCTG
>VisG2b	>VisG9b	>VisG4	>VisG11	>VisH6
GTACTTGACA	CGATGTTGTA	GGCGTTGACA	AGCTGTCACT	GAACGCACTT
>VisG3b	>VisG10b	>VisG5	>VisG12	>VisH8
GCACCCCCAC	TCCTGACAGT	TAAGTGACAA	GCTGTCATTA	CGATGTTGTA
>VisG4b	>VisG11b	>VisG6	>VisH1	>VisH9
GAGGTGACAT	GGTCTCTGCG	GCTTGTCATC	TATCCTTTAG	TCGGGATGTG
>VisG5b	>VisG12b	>VisG7	>VisH2	>VisH10
CGCTGTCATT	ATCAGTCCTT	TGTTTTGACA	AAGTTGACAT	TCCTGACAGT
>VisG6b	>VisG1	>VisG8	>VisH3	>VisH11
TGGTGTCAAC	CAAATGACA	GGAAGTGACA	AGGGTTAAGA	AAAGTGAACT
>VisG7b	>VisG2	>VisG9	>VisH4	
CAAGTTGACA	GCTCTGACAG	CGTGTCAAAC	TCACGCCGAG	

Vnd				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KRKRRVLFTKAQTYELERRFRQQRYLSAPEREHLASLIRLTPTQVKIWFQNHRYKTKRAQNEKGYEGHPGLLH			
Selected sequences				
>VndE2	>VndE9	>VndF2	>VndF7	>VndF12
GTCTCAAGTG	ATCTCAAGTG	TTTTCAAGTG	CACTCAAGTG	CATTCAAGTG
>VndE3	>VndE10	>VndF3	>VndF8	
CTCTCAAGTA	TTTTTAAGTA	GTACCTGGAT	TTTTCAAGTG	
>VndE5	>VndE11	>VndF4	>VndF9	
TATTGAAGTA	GTCTCAAGTG	CTCTTAAGTA	TTTTCAAGTA	
>VndE7	>VndE12	>VndF5	>VndF10	
TGGTCAAGTA	TATTGAAGTA	TTATCAAGAG	GTTTCAAGTG	
>VndE8	>VndF1	>VndF6	>VndF11	
ATCTCAAGTA	TGTTCAAGAG	GTCTCAAGTA	CGTCCTCAAC	

Zen	
Promoter-Stringency	UV5m-10mM

Amino Acid Sequence	LKRSRTAFTSVQLVELENEFKSNMYLYRTRRIEIAQRLSLCERQVKIWFQ NRRMKFKKDIQGHREPKSNAKLA			
Selected sequences				
>ZenC1	>ZenC7	>ZenC12	>ZenD6	
GCCTTAATTA	CATGCAATTT	TCACGCCGAG	TCCTTAATGA	
>ZenC3	>ZenC8	>ZenD2	>ZenD7	
GACTAATTA	TACCTAATGA	TCCCTAATGA	TTTTTAATTA	
>ZenC4	>ZenC9	>ZenD3	>ZenD9	
GGGCTAATTA	CGTTTAATGA	GGCTTAATGA	TACATAATGA	
>ZenC5	>ZenC10	>ZenD4	>ZenD11	
ATAATAATGA	TCCCTAATGA	TACCTAATGA	TGCTAATGA	
>ZenC6	>ZenC11	>ZenD5		
CGTCTCGCGA	GCGCTAATGA	GTGCTAATTA		

Zen2				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	SKRSRTAFSSLQLIELEREFHLNKYLARTRRIEISQRLALTERQVKIWFQN RRMKLKKSTNRKGAIGALTTS			
Selected sequences				
>Zen2A1	>Zen2A10	>Zen2B8	>Zen2E6	>Zen2F3
GTACAAGAGG	TTAGTAATTA	CTAGCTTACG	TGCTAAATTA	TTACTAATGA
>Zen2A2	>Zen2A11	>Zen2B9	>Zen2E7	>Zen2F4
ATTATAATGA	ATCCGAAGTG	GCCTTAATTA	GCACTAACGA	GCCACACGCG
>Zen2A3	>Zen2B1	>Zen2B10	>Zen2E8	>Zen2F5
GCGATCGGTG	GCATAACATC	GTTACGCGTG	GTATTGCAAG	TTGCTAATTA
>Zen2A4	>Zen2B2	>Zen2B11	>Zen2E9	>Zen2F6
ACCGTAATTA	ACTCTGTGAG	GAGTTAATGA	GCAATTAAGG	CAGCTAATTA
>Zen2A5	>Zen2B3	>Zen2E1	>Zen2E10	>Zen2F7
CCCCAAAGTG	ATACTAATTA	ATAATTAAGT	GCCTTAATTA	TGTTAATGA
>Zen2A6	>Zen2B4	>Zen2E2	>Zen2E11	>Zen2F8
CTACTAATTG	GTCATTAGTA	CGTTTAAATG	GCAATTAAAA	CACCCGTGTG
>Zen2A7	>Zen2B5	>Zen2E3	>Zen2E12	>Zen2F9
TGAGTAATGA	GTACTGTGTG	TATTTAATTA	AACACCTGTG	CGAATCAGCG
>Zen2A8	>Zen2B6	>Zen2E4	>Zen2F1	>Zen2F10
CTGATAATGA	TGGCATTACG	TCAATTATGA	TCCGTAATGA	AGTTCCTGTG
>Zen2A9	>Zen2B7	>Zen2E5	>Zen2F2	>Zen2F11
CCCGTAATTA	CTTGATCGTG	GTCATTAAAA	TTCGCCGACG	GTGTTAATTA

CG11085

Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	KPRRRRTAFTHAQLAYLERKFRCQKYLSVADRSDVAETLNLSETQVKT WYQNRRTKWKRQNQLRLEQLRHQA			
Selected sequences				
>CG11085G1	>CG11085G6	>CG11085G11	>CG11085H4	>CG11085H9
TCCCATTAAG	TCCCATTAAG	AGCAATTAGG	ACCAATTAAC	ATCAGTCCTT
>CG11085G2	>CG11085G7	>CG11085G12	>CG11085H5	>CG11085H10
CCACTAATTA	TGTAATTAAA	TCTGTTGAGT	ATGTTGGTAT	ATGTTGGTAT
>CG11085G3	>CG11085G8	>CG11085H1	>CG11085H6	>CG11085H11
CCCAATTAAA	GGCTATTAGA	AGTTAAGTTC	AACATTTAAT	GTCAATTAAT
>CG11085G4	>CG11085G9	>CG11085H2	>CG11085H7	
CGCAATTAAA	GCCGATTAAG	ATCAATTAGC	GTCCAGATTG	
>CG11085G5	>CG11085G10	>CG11085H3	>CG11085H8	
CCCTATTAAA	ACTCCCAGTG	TGTGACCATC	GTCCAAGCGC	

CG11294				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	QRRNRTTFTPQQLQELEALFQKTHYPDVFLREEVALRISLSEARVQVWF QNRRAKWRKQARLQLLQDAWRMRC			
Selected sequences				
>CG11294A1	>CG11294A11	>CG11294B4	>CG11294B9	
CTTAATTATC	TAACTAATTA	GCTAATTAAT	GCTAATTAAT	
>CG11294A3	>CG11294A12	>CG11294B5	>CG11294B10	
TTTAATTAGC	TAATTAGA	ATCAATTAAG	GCTAATTAAT	
>CG11294A6	>CG11294B1	>CG11294B6	>CG11294B11	
CCCCTATTT	CTAATTAATA	CCTAATTAAA	GCCATACCAC	
>CG11294A9	>CG11294B2	>CG11294B7		
GTCAATTAGC	CTAATTAGGT	ATTAATTATG		
>CG11294A10	>CG11294B3	>CG11294B8		
GTAGCCAATG	TGCTAATTAA	GCTAATTAAC		

CG11617				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	SRATKRLFTPDIKRMLKDWLIRRRENPYPSREEKKQLAAETGLTYTQICN WFANWRRKLKNSEREKAKKSWGHLIK			
Selected sequences				
>LagG2	>LagG9	>LagH2	>LagH8	
TTTTTTTACA	CGAATTTACA	CTGTTTTACA	TAAGTTGACA	
>LagG3	>LagG10	>LagH3	>LagH10	

CGATTTAACA	TCATTTAACA	AGATTTAACA	ATCAGTCCTT	
>LagG5	>LagG11	>LagH4	>LagH11	
AATTTTAACA	CATTTTGACA	TTTTTTTACA	GATTTTAACA	
>LagG6	>LagG12	>LagH5		
GATTTTAACA	ATGTCAAAAA	GTTTAACA		
>LagG8	>LagH1	>LagH6		
ATGTTTAACA	GAATTTAACA	AAAAATAACA		

CG12361				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RGMMRRAVFSDSQRKGLEKRFQQQKYISKPDRKKLAERLGLKDSQVKIWFQNRRMKWRNSKERELLASGGSRD			
Selected sequences				
>CG12361A1	>CG12361A6	>CG12361A11	>CG12361B4	>CG12361B11
ATGTTTATGA	TCCTTAATGA	GTGATAAAAA	CCAATAAAAT	ATCAGTCCTT
>CG12361A2	>CG12361A7	>CG12361A12	>CG12361B5	
TTATAAATTA	CCCATAAACC	GGTATTATTA	GTAATTAAAA	
>CG12361A3	>CG12361A8	>CG12361B1	>CG12361B6	
CAGCCTCATCA	GTGATAAAAT	GGTATTATTA	CAAGCATGTT	
>CG12361A4	>CG12361A9	>CG12361B2	>CG12361B7	
TAATTGATGA	TGTCATTAAA	GTAATCAAAG	TCACGCCGAG	
>CG12361A5	>CG12361A10	>CG12361B3	>CG12361B10	
TCATTTAATA	GGCAATTAAG	CTAATAAACG	CTACTAGATC	

CG13424				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RKKRPRTAFSAAQIKALETEFERGKYLSVAKRTALAKQLQLTETQIKIWFQNRRTKWKRKYTSDVETLASHYYA			
Selected sequences				
>CG13424E1	>CG13424E11	>CG13424F9	>CG13424C7	>CG13424D4
TCAATTAGGC	CTAATTAATA	AGTTGACCAC	AGCTATTAAG	GGCTGAGTC
>CG13424E2	>CG13424E12	>CG13424F10	>CG13424C8	>CG13424D5
GCTAATTAAA	CTGTTAATGA	AGGGTCCCGC	CAGTTAATAG	ACTGCAAAAG
>CG13424E3	>CG13424F2	>CG13424F11	>CG13424C9	>CG13424D7
TGTTTAATTA	TAACTAATGG	GGTATACGAC	GATACTTCAA	ACTGCAAAAG
>CG13424E4	>CG13424F3	>CG13424F12	>CG13424C10	>CG13424D8
GACAATTATA	TGCAATTACC	CCACTATATT	ACAAAGTTCA	AGTGCTTCAC
>CG13424E5	>CG13424F4	>CG13424C1	>CG13424C11	>CG13424D9
TTATCTCCAA	GCCTATTAAA	CTAATTAACC	CAAGACATTG	ACTAGCCGTG

>CG13424E7	>CG13424F5	>CG13424C2	>CG13424C12	>CG13424D10
CATGCTATGA	TGTCTAATGA	GCCTAAGAGA	CCAATTAGGA	CGAAATCCA
>CG13424E8	>CG13424F6	>CG13424C3	>CG13424D1	>CG13424D11
GAGCAGCATC	TGTTATAATG	GCCCCATTCT	GCAATTAGTA	AGGGTTAAGA
>CG13424E9	>CG13424F7	>CG13424C5	>CG13424D2	
CCCATTAATG	CCAATTATGC	CGATTAATTA	ATAATTAGGA	
>CG13424E10	>CG13424F8	>CG13424C6	>CG13424D3	
TCAGCCG	TCAATTAAAC	CTACTAATTG	TCACGCCGAG	

CG15696				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	GRLPRIPFTPQQLQALENAYKESNYLSAEDANKLADSLELTNTRVKIWFQNRRARERREKREKDESCDSTFSS			
Selected sequences				
>CG15696G1	>CG15696G10	>CG15696H7	>156962G4	>156962G12
CCCTTAATTA	ATTAATAATA	CCTAATTAAT	CCTAATAATA	CCAATTTAA
>CG15696G2	>CG15696G11	>CG15696H8	>156962G5	
CCTCATTAAA	CCCAATTGAG	ATCAATTAAC	CTTAATTACG	
>CG15696G3	>CG15696G12	>CG15696H9	>156962G6	
ACCAATTGAC	GTTAATTAGC	GTTAATTAAC	CCGCTAATTA	
>CG15696G5	>CG15696H1	>CG15696H10	>156962G7	
TGTTTTATTA	ACAAAGTTCA	CCTAATTAAG	TTCAATTACG	
>CG15696G6	>CG15696H2	>CG15696H11	>156962G8	
CCAATTAGAG	CTCAATCAAC	TCCAATTAAC	CGCAATAAAA	
>CG15696G7	>CG15696H3	>156962G1	>156962G9	
CCAATTACAA	GCCATCAAG	AGTAATTACA	TCAATCAAT	
>CG15696G8	>CG15696H4	>156962G2	>156962G10	
CTCAATCTAG	ACCAATCAAG	TGGGTAATTG	GCTAATTAAT	
>CG15696G9	>CG15696H5	>156962G3	>156962G11	
TCTAATTAGA	TTCAATTAAA	CATAATCAAA	ACCAATAAAA	

CG18599				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	NKRVRTIFTPEQLECLEAEFERQQYMVGPERLYLAHTLKLTEAQVKVWFQNRRIKWRKHHLELTQQRLALIRQ			
Selected sequences				
>CG18599C1	>CG18599C7	>CG18599D2	>CG18599F1	>CG18599F7
ATCAATTAAA	AATTAATTA	CTCATTATAA	CCGTTAATTA	GCTTAATGA
>CG18599C2	>CG18599C8	>CG18599D3	>CG18599F2	>CG18599F9

TTGCTAATGA	CCCCTAATTG	CCGCGTCCG	TTAATAATGA	AATTCTGTCA
>CG18599C3	>CG18599C9	>CG18599D5	>CG18599F3	>CG18599F11
ATGCTAATTA	TGATTAATGA	GCCACGTGT	CCGCTAATTA	GTTATAATGA
>CG18599C4	>CG18599C10	>CG18599D9	>CG18599F4	>CG18599F12
TAATTAATTA	TATCTAATTA	CCCGTAATTA	GCGTTAATTA	AGCCTAATTA
>CG18599C5	>CG18599C11	>CG18599D10	>CG18599F5	
CCTTTAATTA	CCACTAATGA	TTAATCACTC	CACTAATTA	
>CG18599C6	>CG18599C12	>CG18599D11	>CG18599F6	
TTAATTAAAA	CCGCTAATTA	TACATAATGA	CTAATTAGTA	

CG32105				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PKRPRTILTSQQRKQFKASFDQSPKPCRKVREALAKDTGLSVRVVQVWFQNQRAKMKKIQRKAKQNGGSGGGS			
Selected sequences				
>CG32105G2	>CG32105G7	>CG32105G12	>CG32105H5	>CG32105H11
GCTAATTAAA	ACCAATTAGA	CTTAATTAAA	GCTAATTAAT	TTTAATTAAG
>CG32105G3	>CG32105G8	>CG32105H1	>CG32105H6	
GTTAATTAGT	TTAATATAAC	CTTAATTAAA	CTACCAACTT	
>CG32105G4	>CG32105G9	>CG32105H2	>CG32105H7	
GCAATTAGAC	TGTCATTAAA	GTAATTAATA	GCTAATTAAT	
>CG32105G5	>CG32105G10	>CG32105H3	>CG32105H9	
ACTAATTATT	TCAATTATAG	ATTAATTAGA	CTACTAAATT	
>CG32105G6	>CG32105G11	>CG32105H4	>CG32105H10	
GCAATTAATG	CTTCACTGAA	CTAATTTGGA	GCTAATTAAT	

CG32532				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RRRHRTTFTQEQLAELEAAFAKSHYPDIYCREELARTTKLNEARIQVWFQNRRAKYRKQEKQLQKALAPS			
Selected sequences				
>CG32432G1	>CG32432G6	>CG32432G11	>CG32432H4	>CG32432H9
TGTAATTAGC	TCTAATTATT	GTCAATTAAT	ACCAATTAAT	ATCAATTAAG
>CG32432G2	>CG32432G7	>CG32432G12	>CG32432H5	>CG32432H10
CCAATTATCA	GCTTAATGA	ACTAATTAAG	TTTAATTATA	GTTAATTAAC
>CG32432G3	>CG32432G8	>CG32432H1	>CG32432H6	>CG32432H11
TTAATTATCT	GGTAATTACC	CCAATTAACG	GTTAATTAGC	TACATAATGA
>CG32432G4	>CG32432G9	>CG32432H2	>CG32432H7	
GCAAATTAAA	GCAATTAGTA	ACCAATTAGG	TCAATTAGGA	

>CG32432G5	>CG32432G10	>CG32432H3	>CG32432H8	
TAGGTAATTA	CCAATTAGAT	CTTAATTAAA	GCCAATTAAT	

CG33980				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	RRHSRTIFTSYQLEKLEEAFKEAHYPDVYAREMLSLKTELPEDRIQVWFQNRRAKWRKTEKVWGGSTIMAEYG			
Selected sequences				
>CG33980A1	>CG33980A6	>CG33980A11	>CG33980B5	
TTTAATTACA	GGCAATTAAC	GAACACTACT	TAAACTCTCC	
>CG33980A2	>CG33980A7	>CG33980A12	>CG33980B6	
TCTAATTATG	GCCAATTAAC	GCCAATTAAC	ACTAATTAAA	
>CG33980A3	>CG33980A8	>CG33980B2	>CG33980B7	
CTCGTCAAG	GTTAATTAAA	TCTAATTAGA	GCTAATTAGA	
>CG33980A4	>CG33980A9	>CG33980B3	>CG33980B8	
GCTAATTAAC	AGGTCAGATA	GTTAATTAC	ATCAGTCCTT	
>CG33980A5	>CG33980A10	>CG33980B4	>CG33980B10	
GCTAATTAAC	CGACTGGAGA	CCTAATTATA	GCTAATTAAT	

CG34031				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	DRKPRQAYSASQLERLENEFNLDKYLSVSKRVELSKSLSLTEVQVKTWFQNRRTKWKKQLTSRLKIAHRHG			
Selected sequences				
>CG34031E1	>CG34031E10	>CG34031F8	>CG34031A6	>CG34031B5
AACAATTACA	AGGGTTAGA	CAGTGTTAAT	GACAATTAGA	CCTATTTAAA
>CG34031E2	>CG34031E11	>CG34031F9	>CG34031A7	>CG34031B6
GCAATTAAAG	AATGATCGTC	TATTCTAAGA	GCAATTAAAC	GTCTATTAAA
>CG34031E3	>CG34031E12	>CG34031F10	>CG34031A8	>CG34031B7
GCTAATTAAC	ACGTAACGA	AAGAGTCCTA	AACTATAAAA	TCCCTAACGA
>CG34031E4	>CG34031F1	>CG34031F11	>CG34031A9	>CG34031B8
ATTTTTATTG	GTCCAGATTG	TTGGGTACAA	GAACCTACTT	ACTGCAAAAG
>CG34031E5	>CG34031F2	>CG34031A1	>CG34031A10	>CG34031B9
AACATTTAAT	AGGAATTGGA	CTAATTAATA	GTGTTAATTG	CTTCACTGAA
>CG34031E6	>CG34031F3	>CG34031A2	>CG34031A11	>CG34031B10
ATTTTAATAG	AGCAATTAAC	ATCTATTAAG	AACAATTAAG	ATCAGTCCTT
>CG34031E7	>CG34031F4	>CG34031A3	>CG34031A12	>CG34031B11
GCTTTTATAG	GACCATTAAG	CCAATTAAAC	ACCTATTAAA	GCGCTAATGA
>CG34031E8	>CG34031F6	>CG34031A4	>CG34031B2	

ACTATTAAAG	GAGTTGTAGA	GCAATTAAAC	TGCAATTAAG	
>CG34031E9	>CG34031F7	>CG34031A5	>CG34031B3	
AGGGTACAA	GAGTTGTAGA	TGCAATTAAA	ATCAATTAAC	

CG4136				
Promoter-Stringency	UV5-10mM			
Amino Acid Sequence	KRRHGRTIFTSSQLEELEKAFKEAHYPDVSARELLSMKTGLAEDRIQVWYQNRRAKWRKTEKCWGHSTKMAEYG			
Selected sequences				
>5CG4136G1	>5CG4136G6	>5CG4136G11	>5CG4136H4	>5CG4136H9
AGCTAATTA	GCCTAATTG	GCTCATTAGG	GGATAATTAA	AACAATTACA
>5CG4136G2	>5CG4136G7	>5CG4136G12	>5CG4136H5	>5CG4136H10
GTAAATTAAC	GAATTAATGA	ACTCATTAAC	GCTTAATTG	TTTAATTAAA
>5CG4136G3	>5CG4136G8	>5CG4136H1	>5CG4136H6	>5CG4136H11
GTCTATTAAG	CCGATAATTA	CTACTAATTG	GCTCTAATTA	CTTAATTATC
>5CG4136G4	>5CG4136G9	>5CG4136H2	>5CG4136H7	
TATAATTAGC	TCCTATTAAA	CAACTAATTA	GGTTTAATTA	
>5CG4136G5	>5CG4136G10	>5CG4136H3	>5CG4136H8	
ACTAATTAAG	AGTAATTAGG	CTTAATTACG	CCGCTGTATG	

CG4328				
Promoter-Stringency	UV5-10mM			
Amino Acid Sequence	PKRPRTILNTQQRRAFKASFEVSPKPCRKVRENLAKDTGLSLRIVQVWFQNQRAKVKKIQKKAKQEPPSKGAS			
Selected sequences				
>5CG4328E1	>5CG4328E7	>5CG4328F1	>5CG4328F9	>5CG43282A5
GCAATAAGCA	ATCAATAAAT	CACAATTAGA	TTAATTATCG	CTTAATAAAT
>5CG4328E2	>5CG4328E8	>5CG4328F2	>5CG4328F10	>5CG43282A6
GAATATATTA	GCTATAATTA	CCAATTATAC	TCAATAATGA	CGTATATGAA
>5CG4328E3	>5CG4328E9	>5CG4328F4	>5CG4328F11	>5CG43282A7
TTTCATAATGT	CGTCATTATT	GCCAATTACA	TTTAATTATT	GCAATAAATT
>5CG4328E4	>5CG4328E10	>5CG4328F5	>5CG43282A1	>5CG43282A8
ATCGTTATTG	CCATTAATA	GTTAATTATT	ACTCATAATT	TCAATTATCA
>5CG4328E5	>5CG4328E11	>5CG4328F6	>5CG43282A2	>5CG43282A10
GCAATATCTA	CGTAATATAA	TGCAATATAT	ATCAATAATA	TTAATATGAG
>5CG4328E6	>5CG4328E12	>5CG4328F7	>5CG43282A3	>5CG43282A11
CCTCAATTATG	CCTAATTAAG	AGCAATATAG	ATATTTATTA	GCAATTAAAG

CG7056	
Promoter-Stringency	UV5m-10mM

Amino Acid Sequence	RKGGQIRFTSQQTKNLEARFASSKYLSPEERRHLALQLKLTDRQVKTWF QNRRAKWRRANLSKRSASAQGPI			
Selected sequences				
>CG7065A1	>CG7065A9	>CG7065B5	>CG7065B11	>CG7056G6
ATTTTGATAA	TACTGAAGTA	GTTTCGATTA	CATTAATAAC	TATTTAATTA
>CG7065A4	>CG7065A10	>CG7065B6	>CG7056G1	>CG7056G7
CCTTTAATAA	TATTTAATGA	CATTGTACTA	ATATTAATTA	GGTAATTAGA
>CG7065A5	>CG7065A11	>CG7065B7	>CG7056G2	>CG7056G8
ACTCTAATTA	GAACCTACTT	TAACCAATAA	CTTTGAAGTA	CTTTTAATAA
>CG7065A6	>CG7065B1	>CG7065B8	>CG7056G3	>CG7056G10
TGCTCAACAA	AAGTCAATTA	CATTGTACTA	ACCCTAATTG	ATCTTATTAC
>CG7065A7	>CG7065B3	>CG7065B9	>CG7056G4	>CG7056G11
AAATGAATTG	AGCACAACAA	GTTTTAATAA	AATCAATTAC	AGACCAATTG
>CG7065A8	>CG7065B4	>CG7065B10	>CG7056G5	>CG7056G12
TGGTTAATTA	TTATCTCCAA	TTCTGACCAT	ATTTGAATTG	GGACGAAGTA

CG9876				
Promoter-Stringency	UV5m-10mM			
Amino Acid Sequence	PRRNRTTFSSAQLTALEKVFERTHYPDAFVREELATKVHLSEARVQVWF QNRRAKFRRNERSVGSRTLLDTA			
Selected sequences				
>CG9876A1	>CG9876A6	>CG9876B2	>CG9876B7	>CG9876B12
ACAATTAGCA	AGTTATTAAG	CCAATTAGTG	TTTAATTAAT	TGCAATTATC
>CG9876A2	>CG9876A7	>CG9876B3	>CG9876B8	
ACTCATTACG	CCAATTAGTG	CAATAATTAG	TCTAATTATC	
>CG9876A3	>CG9876A8	>CG9876B4	>CG9876B9	
TGACTAATTA	CTATTAATTA	ACTAATTAGA	GCTAATTAAC	
>CG9876A4	>CG9876A9	>CG9876B5	>CG9876B10	
CTCGTAATTG	AGACCAAGCA	TTTAATTAAG	ACTAATTAAC	
>CG9876A5	>CG9876A10	>CG9876B6	>CG9876B11	
GTTAATTGA	ACTTATTAAC	TTTAATTACA	ATTAATTATG	

Table A.3

Master alignment of selected binding sites. Sequences that were identified for each factor to contain an overrepresented motif based on CONSENSUS analysis were aligned at the common Ade at binding site position 3 that is shared by all Asn51 containing homeodomains.

```
>Vis:VisG1b:-:5:1
aaaTGACAc-
>Vis:VisG2b:-:6:2
actTGACA--
>Vis:VisG4b:-:5:4
aggTGACAt-
>Vis:VisG5b:+:4:5
-aaTGACAgc
>Vis:VisG6b:+:4:6
-gtTGACAcc
>Vis:VisG7b:-:6:7
agtTGACA--
>Vis:VisG8b:+:2:8
cgaTGACAg-
>Vis:VisG10b:-:4:10
tccTGACAgt
>Vis:VisG1:-:5:13
aaaTGACA--
>Vis:VisG2:-:5:14
ctcTGACAg-
>Vis:VisG3:-:5:15
ttaTGACAc-
>Vis:VisG4:-:6:16
cgtTGACA--
>Vis:VisG5:-:5:17
aagTGACAa-
>Vis:VisG6:+:4:18
-gaTGACAag
>Vis:VisG7:-:6:19
tttTGACA--
>Vis:VisG8:-:6:20
aagTGACA--
>Vis:VisG9:+:3:21
gttTGACAcg
>Vis:VisG10:+:4:22
-ttTGACAgt
>Vis:VisG11:+:4:23
-agTGACAgc
>Vis:VisG12:+:3:24
taaTGACAgc
>Vis:VisH2:-:5:26
agtTGACAt-
>Vis:VisH10:-:4:33
tccTGACAgt
>Achi:AchiG1:-:1:1
-CTTGACA--
>Achi:AchiG2:+:3:2
aTGTGACA--
>Achi:AchiG3:-:4:3
-CATGACAgc
>Achi:AchiG4:+:4:4
aTCTGACA--
>Achi:AchiG5:+:3:5
gCATGACAg-
>Achi:AchiG6:+:3:6
gTTTGACAt-
>Achi:AchiG8:-:4:8
-TATGACAgg
>Achi:AchiG9:-:3:9
tTTTGACAac
>Achi:AchiG10:+:4:10
aTCTGACA--
>Achi:AchiH1:-:2:11
aTTTGATAt-
>Achi:AchiH2:-:4:12
-TCTGACAgc
>Achi:AchiH3:+:3:13
aAATGACA--
>Achi:AchiH4:-:4:14
-ACTGACAga
>Achi:AchiH6:-:3:16
tTTTGACAgc
>Achi:AchiH11:-:3:19
cTTTGACAgc
>Achi:achi2F1:+:3:20
gAATGACAa-
>Achi:achi2F2:-:4:21
-TTTGACAac
>Achi:achi2F3:-:2:22
gATTGACAg-
>Achi:achi2F4:-:4:23
-TTTGACAgc
>Achi:achi2F6:+:4:25
aGATGACA--
>Achi:achi2F8:-:3:26
-TTTGACAgg
>Achi:achi2F11:-
:4:29
-TTTGACAgc
>Achi:achi2F12:-
:3:30
gTTTGACAgg
>Hth:hthC2:+:5:1
ttgTGACAg-
>Hth:hthC3:-:4:2
-ttTGACAta
>Hth:hthC4:-:2:3
cctTGACGt-
>Hth:hthC5:-:3:4
ctgTGACAgg
>Hth:hthC6:+:4:5
ctgTGACGt-
>Hth:hthC8:-:1:6
-gaTGACA--
>Hth:hthC9:-:4:7
-gcTGACAgt
>Hth:hthC10:+:6:8
accTGACG--
>Hth:hthC11:+:5:9
cgaTGACAg-
>Hth:hthC12:-:1:10
gttTGACA--
>Hth:hthD1:-:3:11
cagTGACAgg
>Hth:hthD2:-:4:12
-ttTGACAgc
>Hth:hthD3:+:5:13
attTGACAt-
>Hth:hthD4:+:5:14
aggTGACAg-
>Hth:hthD6:-:4:16
-aaTGACAgc
>Hth:hthD9:-:1:19
caaTGACA--
>Hth:hthD11:+:6:21
tcgTGACA--
>Six4:5Six4E1:+:1:1
ATTTGATAC-
>Six4:5Six4E2:+:1:2
CTTTGAGAC-
>Six4:5Six4E3:+:1:3
ATTTGATAC-
>Six4:5Six4E4:+:2:4
ATCTGACACc
>Six4:5Six4E5:+:1:5
AAATGAGAC-
>Six4:5Six4E6:+:2:6
TCTTGATACc
>Six4:5Six4E8:+:1:8
AGATGATAC-
>Six4:5Six4E9:+:1:9
TTTTGAGAA-
>Six4:5Six4E10:+:2:1
0
AGTTGACACc
>Six4:5Six4E11:+:2:1
1
GTATGATACc
>Six4:5Six4E12:+:2:1
2
GTCTGACACc
>Six4:5Six4F1:+:2:13
ACATGACACc
>Six4:5Six4F2:+:1:14
AAGTGATAC-
>Six4:5Six4F3:+:2:15
AATTGACACc
>Six4:5Six4F4:+:2:16
CTGTGACACc
>Six4:5Six4F5:+:2:17
TAATGATACc
>Six4:5Six4F6:+:2:18
TGATGACACc
>Six4:5Six4F7:+:2:19
TTTTGATACc
>Six4:5Six4F10:+:1:2
1
ATTTGAGAC-
>Six4:5Six4F11:+:1:2
2
TTGTGATAC-
>Optix:11617C4:+:1:3
AAGTGATA--
>Optix:11617C6:+:1:5
ATGTGATA--
>Optix:11617C7:+:1:6
CAGTGATA--
>Optix:11617C8:+:1:7
TTGCGATA--
>Optix:11617C9:-:1:8
TAGCGATG--
>Optix:11617C11:+:1:
10
TAGTGATA--
>Optix:11617D8:+:1:1
7
AAATGATA--
>Optix:11617D9:+:1:1
8
TAGTGATA--
>Optix:11617D10:+:1:
19
AAGTGATA--
>Optix:11617D11:+:1:
20
TAGTGATA--
>Optix:11617D12:+:1:
21
AGATGATA--
>Optix:OptixE8:+:1:2
9
AAGTGATA--
>Optix:Optix2A2:+:1:
34
CCATGATA--
>Optix:Optix2A3:+:1:
35
ATGTGATA--
>Optix:Optix2A4:+:1:
36
AGCTGATA--
>Optix:Optix2A5:+:1:
37
ATCTGATA--
>Optix:Optix2A6:+:1:
38
AGCTGATA--
>Optix:Optix2A10:+:1
:41
CCGCGATA--
>Optix:Optix2A11:+:1
:42
GATTGATA--
>Optix:Optix2A12:+:1
:43
CCTTGATA--
>Optix:Optix2B1:+:1:
44
TCGCGATA--
>Optix:Optix2B3:+:1:
46
TTGCGATA--
>Optix:Optix2B4:+:1:
47
GATTGATA--
>Optix:Optix2B5:+:1:
48
AAGTGATA--
>Optix:Optix2B6:+:1:
49
AAGTGATA--
>Optix:Optix2B7:+:1:
50
TTATGATA--
```

```
>Optix:Optix2B8:+:1:
51
TTCTGATA--
>So:SoE1:-:2:1
cTGTGATAc-
>So:SoE2:+:4:2
cGATGATA--
>So:SoE3:-:3:3
tATTGATAct
>So:SoE5:-:3:5
gAATGATAct
>So:SoE6:+:4:6
aAATGATA--
>So:SoE7:-:2:7
tTGTGATAc-
>So:SoE8:-:2:8
aTCTGATAc-
>So:SoE9:+:4:9
tAGTGATA--
>So:SoE11:-:2:10
tTTTGATAc-
>So:SoE12:+:4:11
gGATGATA--
>So:SoF2:+:4:12
tGATGATA--
>So:SoF4:-:2:14
aAATGATAc-
>So:SoF5:-:2:15
gAGGGATAt-
>So:SoF7:-:2:17
aAATGATAc-
>So:SoF9:+:4:19
gCATGAGA--
>So:SoF11:-:2:21
aAGTGATAc-
>So:So2E1:+:4:22
tCGTGATA--
>So:So2E2:+:4:23
tATTGATA--
>So:So2E3:-:3:24
cAGTGATAtg
>So:So2E4:+:4:25
gCATGATA--
>So:So2E5:-:3:26
gTGTGATAcc
>So:So2E6:+:4:27
gGATGATA--
>So:So2E7:+:4:28
cAGTGATA--
>So:So2E9:-:3:30
tTGTGATAtg
>So:So2E10:+:4:31
aCATGATA--
>So:So2E11:+:4:32
tGATGATA--
>So:So2E12:+:4:33
cAATGATA--
>Exd:ExdE1:+:1:1
GTTTGACAt-
>Exd:ExdE3:+:1:3
CTTTGACAt-
>Exd:ExdE5:+:2:5
TTTTGACA--
>Exd:ExdE6:+:2:6
ATTTGACA--
>Exd:ExdE7:+:1:7
CTTTGATGa-
>Exd:ExdE8:+:1:8
GATTGATGa-
>Exd:ExdE9:+:2:9
AGTTGACA--
>Exd:ExdE10:+:1:10
TTTTGACGa-
>Exd:ExdE12:+:2:12
TTTTGACA--
>Exd:ExdF1:+:1:13
AGTTGACAt-
>Exd:ExdF2:+:1:14
TTTTGATGg-
>Exd:ExdF3:+:1:15
CTTTGATGa-
>Exd:ExdF4:+:2:16
ATTTGACA--
>Exd:ExdF5:+:2:17
GTTTGACA--
>Exd:ExdF6:+:1:18
GTTTGATGa-
>Exd:ExdF7:+:2:19
TTTTGACA--
>Exd:ExdF9:+:2:20
CTTTGACA--
>onecut:5onectG1:+:4
:1
agCTGATTA-
>onecut:5onectG2:-
:3:2
-cTTGATTTc
>onecut:5onectG3:-
:2:3
ccTTGATTGc
>onecut:5onectG4:-
:2:4
taTTGATTAa
>onecut:5onectG8:+:4
:8
ctATGATTA-
>onecut:5onectG10:-
:2:9
taTTGATTAa
>onecut:5onectG11:-
:3:10
-aTTGATTTg
>onecut:5onectG12:-
:3:11
-tCTGATTGa
>onecut:5onectH1:+:4
:12
agTTGATTG-
>onecut:5onectH2:-
:3:13
-aTTGATTTg
>onecut:5onectH4:-
:3:15
-cTTGATTTg
>onecut:5onectH5:+:5
:16
tcTTGATTT-
>onecut:5onectH7:-
:3:18
-gTTGATTTg
>onecut:5onectH10:-
:3:21
-tTTGATTGg
>onecut:5onectH11:-
:2:22
cgTTGATTAg
>Vnd:VndE2:+:1:1
TCTCAAGTG-
>Vnd:VndE3:+:1:2
TCTCAAGTA-
>Vnd:VndE5:+:1:3
ATTGAAGTA-
>Vnd:VndE7:+:1:4
GGTCAAGTA-
>Vnd:VndE8:+:1:5
TCTCAAGTA-
>Vnd:VndE9:+:1:6
TCTCAAGTG-
>Vnd:VndE10:+:1:7
TTTTAAGTA-
>Vnd:VndE11:+:1:8
TCTCAAGTG-
>Vnd:VndE12:+:1:9
ATTGAAGTA-
>Vnd:VndF1:+:1:10
GTTCAAGAG-
>Vnd:VndF2:+:1:11
TTTCAAGTG-
>Vnd:VndF4:+:1:13
TCTTAAGTA-
>Vnd:VndF5:+:1:14
TATCAAGAG-
>Vnd:VndF6:+:1:15
TCTCAAGTA-
>Vnd:VndF7:+:1:16
ACTCAAGTG-
>Vnd:VndF8:+:1:17
TTTCAAGTG-
>Vnd:VndF9:+:1:18
TTTCAAGTA-
>Vnd:VndF10:+:1:19
TTTCAAGTG-
>Vnd:VndF12:+:1:21
ATTCAAGTG-
>Tin:TinA2:-:3:1
gCTCAAGTA-
>Tin:TinA3:+:3:2
-GTCAAGTAc
>Tin:TinA4:+:2:3
cGTCAAGTGg
>Tin:TinA5:-:3:4
tCTCAAGTG-
>Tin:TinA6:+:2:5
aTTCAAGTGg
>Tin:TinA7:+:3:6
-GTCAAGTGc
>Tin:TinA8:+:3:7
-CTTAAGTAc
>Tin:TinA9:+:3:8
-CTTAAGTGg
>Tin:TinA10:+:3:9
-TTTAAGTGg
>Tin:TinA11:+:2:10
cCACAAGTGg
>Tin:TinB1:+:3:11
-CTCAAGTGg
>Tin:TinB3:+:3:12
-CTTAAGTGt
>Tin:TinB4:+:2:13
-TTCAAGTGg
>Tin:TinB5:+:3:14
-CTCAAGTGg
>Tin:TinB6:+:3:15
-CTCAAGTGc
>Tin:TinB9:+:3:18
-TTGAAGTGg
>Bap:BapA2:-:1:1
GCTTAAGTGG
>Bap:BapA3:-:1:2
TCTTAAGTGG
>Bap:BapA12:-:1:10
AGTTAAGTGG
>Bap:BapB1:-:1:11
GTTTAAGTGG
>Bap:BapB5:-:1:14
TGTTAAGTGG
>Bap:BapB6:-:1:15
TCTTAAGTGG
>Bap:BapB7:-:1:16
ACTTAAGTAC
>Bap:BapB9:-:1:18
TGTTAAGTGG
>Bap:BapB11:-:1:20
CGTTAAGTGG
>Bap:Bap2C1:+:1:21
CTTTAAGTGT
>Bap:Bap2C2:-:1:22
ACTTAAGTAC
>Bap:Bap2C3:-:1:23
TTTTAAGTGT
>Bap:Bap2C4:-:1:24
CTGTAAGTGT
>Bap:Bap2C6:-:1:26
GCTTAAGTGC
>Bap:Bap2C7:-:1:27
ACTTAAGAAC
>Bap:Bap2C8:-:1:28
TCTTAAGTAC
>Bap:Bap2C9:-:1:29
ATTTAAGTGA
>Bap:Bap2C10:-:1:30
TATTAAGTAC
>Bap:Bap2C12:-:1:31
TTTTAAGTGA
>Bap:Bap2D2:-:1:32
CTTTAAGTAC
>Bap:Bap2D3:-:1:33
GGTTAAGTGG
>Bap:Bap2D10:-:1:38
ATTTAAGTGA
>Bap:Bap2D11:-:1:39
ACTTAAGTAC
>CG7056:CG7065A1:+:3
:1
tTTTGATAA-
>CG7056:CG7065A4:+:3
:2
cTTTAATAA-
>CG7056:CG7065A5:+:3
:3
cTCTAATTA-
>CG7056:CG7065A6:+:3
:4
gCTCAACAA-
>CG7056:CG7065A7:+:3
:5
aATGAATTG-
>CG7056:CG7065A8:+:3
:6
gGTTAATTA-
>CG7056:CG7065A9:+:3
:7
aCTGAAGTA-
>CG7056:CG7065A10:+:
3:8
aTTTAATGA-
>CG7056:CG7065B1:+:3
:10
aGTCAATTA-
>CG7056:CG7065B5:+:3
:13
tTTCGATTA-
>CG7056:CG7065B6:+:3
:14
```

```
aTTGTACTA-
>CG7056:CG7065B7:+:3
:15
aACCAATAA-
>CG7056:CG7065B8:+:3
:16
aTTGTACTA-
>CG7056:CG7065B9:+:3
:17
tTTTAATAA-
>CG7056:CG7065B11:+:
2:19
cATTAATAAc
>Ubx:UbxC02:+:3:1
aATTAATTA-
>Ubx:UbxC03:+:3:2
aTTTTATTA-
>Ubx:UbxC04:+:3:3
cTTTAATTA-
>Ubx:UbxC05:+:3:4
cTTTAATTA-
>Ubx:UbxC06:+:3:5
tATTAATTA-
>Ubx:UbxC07:-:3:6
-TTTTATTGc
>Ubx:UbxC08:-:3:7
-CTTAATTGc
>Ubx:UbxC11:+:3:8
tATTAATGA-
>Ubx:UbxC12:+:3:9
aTTTAATGG-
>Ubx:UbxD01:+:2:10
tCTTAATGA-
>Ubx:UbxD02:+:3:11
cGTTAATTA-
>Ubx:UbxD04:+:3:12
cCTTAATTA-
>Ubx:UbxD05:-:3:13
-TTTAATTGt
>Ubx:UbxD06:+:3:14
cGTTAATTA-
>Ubx:UbxD07:-:3:15
-TTTAATGGg
>Ubx:UbxD08:+:3:16
cCTTAATGA-
>Ubx:UbxD09:+:3:17
cCTTAATTA-
>Ubx:UbxD10:+:3:18
tTTTTATGA-
>Ubx:UbxD11:-:3:19
-GTTAATTAc
>Ubx:UbxD12:-:3:20
-TTTAATTGc
>C15:C15A1:-:3:1
aTTTAAAGA-
>C15:C15A2:+:3:2
-GTTAATTGg
>C15:C15A3:+:3:3
-GTTAATGGg
>C15:C15A4:+:3:4
-CTTAAACGa
>C15:C15A5:+:3:5
-CTTAACGAg
>C15:C15A6:-:1:6
-GTTTAACA-
>C15:C15A7:+:1:7
cGTTAAACA-
>C15:C15A8:-:3:8
cGTTAACGA-
>C15:C15A9:-:3:9
tCTTAATTG-
>C15:C15A10:-:3:10
cGTTAATGA-
>C15:C15A11:+:3:11
-CTTAAATGa
>C15:C15A12:-:3:12
aTTTAATTG-
>C15:C15B1:+:3:13
-GTTAATTGg
>C15:C15B2:+:3:14
-AATAATGGg
>C15:C15B4:+:3:15
-TTTAATTAg
>C15:C15B6:-:2:16
aTCTTATTAc
>C15:C15B8:+:3:18
-ATTAATTAg
>C15:C15B9:+:3:19
-TTTAAATAg
>C15:C15B11:+:3:21
-GTTAATTGc
>Mirr:MirrE1:+:1:1
TTGTAACA--
>Mirr:MirrE2:+:1:2
CAGTAACA--
>Mirr:MirrE3:+:1:3
TAAATACA--
>Mirr:MirrE4:+:1:4
AGCAAACA--
>Mirr:MirrE5:+:1:5
CTATAACA--
>Mirr:MirrE6:+:1:6
TAATAACA--
>Mirr:MirrE7:+:1:7
TCGTGACA--
>Mirr:MirrE8:+:1:8
AAGTTACA--
>Mirr:MirrE9:+:1:9
AACTTACA--
>Mirr:MirrE10:+:1:10
AAAAAACA--
>Mirr:MirrE11:+:1:11
AGAAAACA--
>Mirr:MirrE12:+:1:12
AAAAAACA--
>Mirr:MirrF2:+:1:14
TACTTACA--
>Mirr:MirrF3:+:1:15
AGAAAACA--
>Mirr:MirrF4:+:1:16
GAAAAACA--
>Mirr:MirrF5:+:1:17
TGAAAACA--
>Mirr:MirrF6:+:1:18
GCCTGACA--
>Mirr:MirrF7:+:1:19
AATTTACA--
>Mirr:MirrF8:+:1:20
TCAAAACA--
>Mirr:MirrF9:+:1:21
AGAAAACA--
>Mirr:MirrF10:+:1:22
ACAAAACA--
>Mirr:MirrF11:+:1:23
AAATAACA--
>Mirr:Mirr2C2:+:1:24
CAGAAACA--
>Mirr:Mirr2C3:+:1:25
AACAAACA--
>Mirr:Mirr2C4:+:1:26
GTACTACA--
>Mirr:Mirr2C5:+:1:27
TTCGAACA--
>Mirr:Mirr2C7:+:1:28
ATTTAACA--
>Mirr:Mirr2C8:+:1:29
TTGTAACA--
>Mirr:Mirr2C9:+:1:30
AGAAAACA--
>Mirr:Mirr2C11:+:1:3
2
TGAAAACA--
>Mirr:Mirr2C12:+:1:3
3
GATATACA--
>Mirr:Mirr2D1:+:1:34
CTCTTACA--
>Mirr:Mirr2D2:+:1:35
AAATTACA--
>Mirr:Mirr2D3:+:1:36
ATATAACA--
>Mirr:Mirr2D4:+:1:37
GTTAAACA--
>Mirr:Mirr2D5:+:1:38
CATAAACA--
>Mirr:Mirr2D7:+:1:40
TGAAAACA--
>Mirr:Mirr2D8:+:1:41
AGTTTACA--
>Mirr:Mirr2D9:+:1:42
AGAAAACA--
>Mirr:Mirr2D10:+:1:4
3
CAATAACA--
>Mirr:Mirr2D11:+:1:4
4
GAATAACA--
>Ara:AraG1:+:1:1
GTATTACA--
>Ara:AraG2:-:1:2
TAAGTACA--
>Ara:AraG4:+:1:3
AGTTTACA--
>Ara:AraG5:+:1:4
TAATTACA--
>Ara:AraG7:+:1:6
GGAAAACA--
>Ara:AraG8:+:1:7
GTATTACA--
>Ara:AraG11:+:1:9
GATATACA--
>Ara:AraG12:+:1:10
TTAGAACA--
>Ara:AraH3:+:1:11
GAATAACA--
>Ara:AraH5:+:1:12
ACATAACA--
>Ara:AraH6:+:1:13
TGTAAACA--
>Ara:AraH7:-:1:14
GCTGAACA--
>Ara:AraH8:+:1:15
CAAAAACA--
>Ara:AraH9:+:1:16
ACATAACA--
>Ara:AraH10:+:1:17
CAAAAACA--
>Ara:AraH11:+:1:18
CTTTTACA--
>Ara:Ara2G1:+:1:19
CCCAAACA--
>Ara:Ara2G2:+:1:20
AATTAACA--
>Ara:Ara2G3:+:1:21
AAGTTACA--
>Ara:Ara2G4:+:1:22
AGAGAACA--
>Ara:Ara2G5:+:1:23
TATAAACA--
>Ara:Ara2G6:+:1:24
GAAAAACA--
>Ara:Ara2G7:+:1:25
TAATAACA--
>Ara:Ara2G8:+:1:26
GTAGAACA--
>Ara:Ara2G10:+:1:27
TGAAAACA--
>Ara:Ara2G12:+:1:29
AATTAACA--
>Ara:Ara2H1:+:1:30
AAAAAACA--
>Ara:Ara2H2:+:1:31
CCAAAACA--
>Ara:Ara2H4:+:1:33
AAATTACA--
>Ara:Ara2H6:+:1:34
GAAAAACA--
>Ara:Ara2H8:+:1:35
ACATTACA--
>Ara:Ara2H9:+:1:36
ACAAAACA--
>Ara:Ara2H10:+:1:37
AGAAAACA--
>Ara:Ara2H11:+:1:38
GAATTACA--
>Caup:CaupA2:+:1:1
TATTAACA--
>Caup:CaupA3:+:1:2
CTTTTACA--
>Caup:CaupA4:+:1:3
CAAAAACA--
>Caup:CaupA5:+:1:4
GTTGTACA--
>Caup:CaupA6:+:1:5
CTATTACA--
>Caup:CaupA7:+:1:6
CATTAACA--
>Caup:CaupA8:+:1:7
AGTTAACA--
>Caup:CaupA9:+:1:8
AAAGAACA--
>Caup:CaupA10:+:1:9
CTTTAACA--
>Caup:CaupA11:+:1:10
AAATGACA--
>Caup:CaupA12:+:1:11
CTAAAACA--
>Caup:CaupB1:+:1:12
TAATAACA--
>Caup:CaupB2:+:1:13
TTGTGACA--
>Caup:CaupB3:+:1:14
CGGAAACA--
>Caup:CaupB4:+:1:15
GGTTAACA--
>Caup:CaupB5:+:1:16
TAATAACA--
>Caup:CaupB9:+:1:19
CAGCAACA--
>Caup:CaupB10:+:1:20
CTGTAACA--
>Caup:CaupB11:+:1:21
AAAAAACA--
>CG11617:LagG2:+:3:1
```

```
TTTTTACA--
>CG11617:LagG3:+:3:2
ATTTAACA--
>CG11617:LagG5:+:3:3
TTTTAACA--
>CG11617:LagG6:+:3:4
TTTTAACA--
>CG11617:LagG8:+:3:5
GTTTAACA--
>CG11617:LagG9:+:3:6
AATTTACA--
>CG11617:LagG10:+:3:
7
ATTTAACA--
>CG11617:LagG11:+:3:
8
TTTTGACA--
>CG11617:LagG12:-
:2:9
TTTTGACAt-
>CG11617:LagH1:+:3:1
0
ATTTAACA--
>CG11617:LagH2:+:3:1
1
GTTTTACA--
>CG11617:LagH3:+:3:1
2
ATTTAACA--
>CG11617:LagH4:+:3:1
3
TTTTTACA--
>CG11617:LagH5:+:1:1
4
GTTTAACA--
>CG11617:LagH6:+:3:1
5
AAATAACA--
>CG11617:LagH8:+:3:1
6
AGTTGACA--
>CG11617:LagH11:+:3:
18
TTTTAACA--
>Ct:5CtE1:+:1:1
TGCTAAAC--
>Ct:5CtE2:+:1:2
GTCTGAAC--
>Ct:5CtE3:+:1:3
GGTTAAAC--
>Ct:5CtE4:+:1:4
CGTTAATC--
>Ct:5CtE5:+:1:5
CATTGAAC--
>Ct:5CtE6:+:1:6
GATTAAAC--
>Ct:5CtE7:+:1:7
TCTTGAAC--
>Ct:5CtE9:+:1:9
AGTTAAAC--
>Ct:5CtE10:+:1:10
TGCTAATC--
>Ct:5CtE11:+:1:11
GTTCAAAG--
>Ct:5CtE12:+:1:12
GCTCGAGC--
>Ct:5CtF1:+:1:13
TCCTGAAC--
>Ct:5CtF3:+:1:15
AATTGAAC--
>Ct:5CtF5:+:1:17
TTCTGAAC--
>Ct:5CtF6:+:1:18
CCTTGAAC--
>Ct:5CtF7:+:1:19
ACTTAAAC--
>Ct:5CtF8:+:1:20
GGTTGAAC--
>Ct:5CtF9:+:1:21
GCTAAAAC--
>Ct:5CtF10:+:1:22
TCTTGAAC--
>Ct:5CtF11:+:1:23
TCCTGAAC--
>CG15696:CG15696G1:-
:3:1
cCTTAATTA-
>CG15696:CG15696G2:+
:3:2
-TTTAATGAg
>CG15696:CG15696G3:+
:3:3
-GTCAATTGg
>CG15696:CG15696G5:-
:3:4
gTTTTATTA-
>CG15696:CG15696G6:+
:2:5
cTCTAATTGg
>CG15696:CG15696G7:+
:2:6
tTGTAATTGg
>CG15696:CG15696G8:+
:3:7
-CTAGATTGa
>CG15696:CG15696G9:-
:1:8
-TCTAATTAg
>CG15696:CG15696G10:
+:3:9
-TATTATTAa
>CG15696:CG15696G11:
+:3:10
-CTCAATTGg
>CG15696:CG15696G12:
-:1:11
-GTTAATTAg
>CG15696:CG15696H2:+
:3:13
-GTTGATTGa
>CG15696:CG15696H3:+
:2:14
-CTTGATGGc
>CG15696:CG15696H4:+
:3:15
-CTTGATTGg
>CG15696:CG15696H5:+
:3:16
-TTTAATTGa
>CG15696:CG15696H7:+
:3:17
-ATTAATTAg
>CG15696:CG15696H8:+
:3:18
-GTTAATTGa
>CG15696:CG15696H9:+
:3:19
-GTTAATTAa
>CG15696:CG15696H10:
+:3:20
-CTTAATTAg
>CG15696:CG15696H11:
+:3:21
-GTTAATTGg
>CG15696:156962G1:+:
3:22
-TGTAATTAc
>CG15696:156962G2:-
:3:23
gGGTAATTG-
>CG15696:156962G3:+:
3:24
-TTTGATTAt
>CG15696:156962G4:+:
3:25
-TATTATTAg
>CG15696:156962G5:-
:1:26
-CTTAATTAc
>CG15696:156962G6:-
:3:27
cGCTAATTA-
>CG15696:156962G7:-
:1:28
-TTCAATTAc
>CG15696:156962G8:+:
3:29
-TTTTATTGc
>CG15696:156962G9:+:
2:30
-ATTGATTGa
>CG15696:156962G10:+
:3:31
-ATTAATTAg
>CG15696:156962G11:+
:3:32
-TTTTATTGg
>CG15696:156962G12:+
:2:33
-TTAAATTGg
>CG4328:5CG4328E1:+:
2:1
tGCTTATTGc
>CG4328:5CG4328E2:-
:3:2
aATATATTA-
>CG4328:5CG4328E3:+:
3:3
aCATTATGAa
>CG4328:5CG4328E4:-
:3:4
tCGTTATTG-
>CG4328:5CG4328E5:+:
2:5
tAGATATTGc
>CG4328:5CG4328E6:+:
4:6:mod
cctCAATTAT
>CG4328:5CG4328E7:+:
3:7
-ATTTATTGa
>CG4328:5CG4328E8:-
:3:8
cTATAATTA-
>CG4328:5CG4328E9:+:
3:9
-AATAATGAc
>CG4328:5CG4328E10:+
:1:10
tATTAATGG-
>CG4328:5CG4328E11:+
:3:11
-TTATATTAc
>CG4328:5CG4328E12:+
:3:12
-CTTAATTAg
>CG4328:5CG4328F1:+:
3:13
-TCTAATTGt
>CG4328:5CG4328F2:+:
2:14
gTATAATTGg
>CG4328:5CG4328F4:+:
3:15
-TGTAATTGg
>CG4328:5CG4328F5:+:
3:16
-AATAATTAa
>CG4328:5CG4328F6:+:
3:17
-ATATATTGc
>CG4328:5CG4328F7:+:
3:18
-CTATATTGc
>CG4328:5CG4328F9:+:
2:19
cGATAATTAa
>CG4328:5CG4328F10:+
:2:20
tCATTATTGa
>CG4328:5CG4328F11:+
:3:21
-AATAATTAa
>CG4328:5CG43282A1:+
:3:22
-AATTATGAg
>CG4328:5CG43282A2:+
:3:23
-TATTATTGa
>CG4328:5CG43282A3:-
:3:24
tATTTATTA-
>CG4328:5CG43282A5:+
:3:25
-ATTTATTAa
>CG4328:5CG43282A6:-
:2:26
cGTATATGAa
>CG4328:5CG43282A7:+
:2:27
aATTTATTGc
>CG4328:5CG43282A8:+
:2:28
tGATAATTGa
>CG4328:5CG43282A10:
-:2:29
tTAATATGAg
>CG4328:5CG43282A11:
+:2:30
cTTTAATTGc
>CG12361:CG12361A1:+
:3:1
tGTTTATGA-
>CG12361:CG12361A2:+
:3:2
tATAAATTA-
>CG12361:CG12361A4:+
:3:4
aATTGATGA-
>CG12361:CG12361A5:-
:1:5
aTTAAATGA-
```

```
>CG12361:CG12361A6:+
:3:6
cCTTAATGA-
>CG12361:CG12361A7:-
:2:7
gGTTTATGGg
>CG12361:CG12361A8:-
:2:8
aTTTTATCAc
>CG12361:CG12361A9:-
:3:9
-TTTAATGAc
>CG12361:CG12361A10:
-:3:10
-CTTAATTGc
>CG12361:CG12361A11:
-:2:11
tTTTTATCAc
>CG12361:CG12361A12:
+:3:12
gTATTATTA-
>CG12361:CG12361B1:+
:3:13
gTATTATTA-
>CG12361:CG12361B2:-
:2:14
cTTTGATTAc
>CG12361:CG12361B3:-
:2:15
cGTTTATTAg
>CG12361:CG12361B4:-
:2:16
aTTTTATTGg
>CG12361:CG12361B5:-
:2:17
tTTTAATTAc
>Cad:CadE1:-:3:1
cACAAATTA-
>Cad:CadE2:+:2:2
tGTTGATTAg
>Cad:CadE3:+:2:3
aAGTTATTAc
>Cad:CadE4:-:3:4
aATTAATAG-
>Cad:CadE5:-:3:5
cTTAAATGA-
>Cad:CadE6:-:4:6
gATTTATTT-
>Cad:CadE8:+:2:7
aTTTTATAAg
>Cad:CadE9:-:3:8
cAGTAATTA-
>Cad:CadE11:+:3:10
-CTTAATTGc
>Cad:CadE12:+:2:11
-TTTTATTAg
>Cad:CadF1:-:3:12
cCGTAATTA-
>Cad:CadF2:+:2:13
cATTTATTGg
>Cad:CadF3:+:2:14
cTTTAATGGc
>Cad:CadF4:-:3:15
gTTTAATAA-
>Cad:CadF5:-:3:16
tATTTATTA-
>Cad:CadF6:-:3:17
aTTTTATTA-
>Cad:CadF7:+:2:18
tGTTTATTGc
>Cad:CadF8:+:3:19
-TTTTATTGa
>Cad:CadF9:+:2:20
gTTTTATGAt
>Cad:CadF11:+:2:22
aCTTTATTAc
>Cad:Cad2E1:+:2:23
cCTTTATTGg
>Cad:Cad2E2:-:3:24
aCATTATTA-
>Cad:Cad2E3:+:3:25
-ATTTATTAg
>Cad:Cad2E4:-:3:26
tATTTATTA-
>Cad:Cad2E7:-:3:28
cTCTAATTG-
>Cad:Cad2E8:+:2:29
aGTTTATTGg
>Cad:Cad2E9:+:2:30
tAATGATTGc
>Cad:Cad2E10:-:3:31
cTTAAATTA-
>Cad:Cad2F1:-:3:34
tTCTAATTA-
>Cad:Cad2F2:+:2:35
tGTTTATGAg
>Cad:Cad2F3:-:3:36
gATTTATAA-
>Cad:Cad2F4:-:3:37
cCATAATTA-
>Cad:Cad2F5:+:2:38
tTTTTATTGc
>Cad:Cad2F6:-:3:39
tTTTTATTG-
>Cad:Cad2F7:+:2:40
aATTTATGGg
>Cad:Cad2F9:-:3:41
tTTTTATGA-
>Cad:Cad2F10:+:2:42
aTTATATGGg
>Cad:Cad2F11:+:2:43
tTTTTATTGc
>H2:H20C1:-:4:1:mod
tttTTATATA
>H2:H20C2:+:4:2
tcTTAATGA-
>H2:H20C3:+:3:3
acATTATTGa
>H2:H20C4:+:3:4
ggTTAATGAt
>H2:H20C5:+:3:5
taGATATTAc
>H2:H20C7:+:3:6
ccTTTATGGg
>H2:H20C8:+:3:7
cgTTGATTAa
>H2:H20C9:-:4:8:mod
agtCAATAAA
>H2:H20C10:+:3:9
gaTTAATTAt
>H2:H20C11:-
:4:10:mod
ggaTAATTAA
>H2:H20C12:+:3:11
ggGTAATTAg
>H2:H20D1:+:3:12
ctATTATTAa
>H2:H20D2:+:3:13
tgAATATTGa
>H2:H20D3:+:3:14
gcTTAATTGa
>H2:H20D4:-:4:15:mod
agaTAATAAT
>H2:H20D5:-:4:16:mod
ctgTAATAAA
>H2:H20D6:-:4:17
accTCATAAT
>H2:H20D7:+:3:18
aaTTTATTAa
>H2:H20D8:-:4:19:mod
atcTAATTAA
>H2:H20D9:-:4:20:mod
tgcTAATAAA
>H2:H20D10:+:3:21
cgTTTATTAa
>H2:H20D11:+:3:22
atTTTATGAg
>H2:5H202H2:+:3:23
caTTTATTGg
>H2:5H202H3:+:3:24
tgTTTATGAa
>H2:5H202H4:-
:4:25:mod
cgtTAATAAA
>H2:5H202H5:-:3:26
-cTTTTTGGc
>H2:5H202H6:+:4:27
atTTTATTA-
>H2:5H202H7:+:3:28
ctTTAATGAg
>H2:5H202H8:-:1:29
ttATTATGG-
>H2:5H202H9:-
:4:30:mod
catTAATAAA
>H2:5H202H10:+:3:31
agGTAATTAc
>H2:5H202H11:+:3:32
gtTTAATGAc
>AbdB:abdb2:+:2:1
GGTTTATAG-
>AbdB:abdb3:+:1:2
GTTTTATTGt
>AbdB:abdb4:+:2:3
TTTTTATGG-
>AbdB:abdb5:+:2:4
GATTAATGG-
>AbdB:abdb6:+:2:5
GCTTTATGT-
>AbdB:abdb7:+:1:6
GGTTTACAAc
>AbdB:abdb8:+:2:7
GTTTAATGT-
>AbdB:abdb10:+:2:8
GATTTATGT-
>AbdB:abdb11:+:2:9
ATATTATGA-
>AbdB:abdb12:+:2:10
CATTTATTA-
>AbdB:abdb13:+:1:11
TTTTTATAAc
>AbdB:abdb14:+:2:12
TATTAATTA-
>AbdB:abdb15:+:1:13
GTTTTATGA-
>AbdB:abdb17:+:2:14
GTTTTATGG-
>AbdB:abdb18:+:2:15
CTTTAACGA-
>AbdB:abdb19:+:2:16
CTTTTATTA-
>AbdB:abdb20:+:2:17
CTTTTACGA-
>AbdB:abdb21:+:2:18
GATTTATTA-
>AbdB:abdb22:+:2:19
GATTTATTA-
>AbdB:abdb23:+:2:20
CTTTAATTA-
>AbdB:abdb24:+:2:21
GTTTTATGA-
>Lim3:Lim3C1:-:3:1
cCCTGATTA-
>Lim3:Lim3C2:-:1:2
-ATTTATTAa
>Lim3:Lim3C3:-:3:3
gCTTAATCA-
>Lim3:Lim3C4:+:2:4
gAATAATTAt
>Lim3:Lim3C5:+:2:5
tAATGATTAt
>Lim3:Lim3C7:-:3:6
aACAAATTA-
>Lim3:Lim3C8:+:3:7
-TCTGATGAa
>Lim3:Lim3C10:+:2:8
tAAAAATTAa
>Lim3:Lim3C11:+:3:9
-CTAAATGAa
>Lim3:Lim3C12:+:3:10
-ATAAATTAg
>Lim3:Lim3D1:+:3:11
-TTTAATGAa
>Lim3:Lim3D3:+:2:13
aTCTAATGAg
>Lim3:Lim3D4:+:1:14
-GGTAATTG-
>Lim3:Lim3D5:+:3:15
-CTTAATTGa
>Lim3:Lim3D6:+:3:16
-TGTAATTGa
>Lim3:Lim3D7:+:2:17
tCTTAATTGa
>Lim3:Lim3D8:-:1:18
-ACTAATTAa
>Lim3:Lim3D9:-:1:19
-ATTAATCAa
>Lim3:Lim3D10:+:1:20
gGCTAATTA-
>Lim3:Lim3D11:+:2:21
aTCTAATTAg
>Awh:AwhC1:-:2:1
atTTGATTAc
>Awh:AwhC2:-:2:2
ggCTGATTGg
>Awh:AwhC3:-:3:3
-tTTAATGAa
>Awh:AwhC4:+:1:4
--CTGATTAc
>Awh:AwhC5:+:4:5
atTTGATTA-
>Awh:AwhC7:-:2:7
aaTTAAGTAg
>Awh:AwhC8:+:1:8
--CTAATTAc
>Awh:AwhC9:+:1:9
--ATAATTAt
>Awh:AwhC10:-:2:10
taTTAATGAa
>Awh:AwhC12:+:4:12
acTTGATTA-
>Awh:AwhD1:-:2:13
cgCTAATGAg
>Awh:AwhD2:+:4:14
```

```
gtCTAATTA-
>Awh:AwhD3:-:2:15
acCTAATTAc
>Awh:AwhD5:+:4:16
caCTAATTA-
>Awh:AwhD6:-:2:17
ctTTAATTAc
>Awh:AwhD7:+:4:18
ttTTAATTA-
>Awh:AwhD8:-:1:19
cgCTAATTG-
>Awh:AwhD9:+:4:20
agTTAATTA-
>Awh:AwhD10:+:4:21
taCTAATGA-
>Awh:AwhD11:+:4:22
acCTAATGA-
>Awh:Awh2A1:+:4:23
ccTTAATTT-
>Awh:Awh2A2:-:2:24
atTTGATTAg
>Awh:Awh2A3:+:1:25
--TTAATTAt
>Awh:Awh2A4:+:4:26
ttTTGATTA-
>Awh:Awh2A8:+:4:27
ccTAAATGA-
>Awh:Awh2A9:-:3:28
-tTTAATTAg
>Awh:Awh2A10:-:2:29
tgCTAATTGg
>Awh:Awh2A11:+:2:30
-cATAATTA-
>Awh:Awh2A12:+:1:31
--TTAATTAg
>Awh:Awh2B1:+:4:32
agTTAATTA-
>Awh:Awh2B2:+:4:33
agATAATTA-
>Awh:Awh2B3:-:2:34
-cTTAATTAt
>Awh:Awh2B4:+:4:35
aaCTAATTA-
>Awh:Awh2B5:+:4:36
caCTAATTT-
>Awh:Awh2B6:+:4:37
tcCTAATTA-
>Awh:Awh2B7:-:3:38
-tTTAATTAg
>Awh:Awh2B8:+:4:39
taTTAATTA-
>Awh:Awh2B9:+:4:40
ccCTAATGA-
>Awh:Awh2B10:+:4:41
agTTAATGA-
>Awh:Awh2B11:+:4:42
ctCTAATTG-
>Dll:DllC1:-:3:1
--tTAATTGT
>Dll:DllC2:-:5:2:mod
TGATAATgga
>Dll:DllC3:-:2:3
-tgTAATAGC
>Dll:DllC4:-:2:4
-gtTAATTTC
>Dll:DllC5:+:3:5
-taTAATTTT
>Dll:DllC6:-:2:6
-ccTAATTTG
>Dll:DllC7:-:2:7
-ttTAATGGC
>Dll:DllC8:-:2:8
-acAAATTGT
>Dll:DllC9:-:3:9
--aTAATTAC
>Dll:DllC10:-:4:10
---TAATTAC
>Dll:DllC11:-:4:11
---TAATTAC
>Dll:DllC12:-:2:12
-tcTAATGAT
>Dll:DllD1:-:2:13
-ctTAATAAC
>Dll:DllD2:+:3:14
-caTAATTTT
>Dll:DllD3:+:2:15
--cTGATAGG
>Dll:DllD4:+:4:16
ggtTAATTAC
>Dll:DllD5:-:2:17
-tcTAATTGC
>Dll:DllD6:-:2:18
-taTAATTGC
>Dll:DllD7:-:2:19
-cgTAATTGG
>Dll:DllD8:-:4:20
---TAATTAT
>Dll:DllD9:+:3:21
-agTAATTAC
>Dll:DllD10:+:2:22
--cTAATTAC
>Dll:DllD11:-:4:23
---TAATTAG
>CG4136:5CG4136G1:+:
3:1
agCTAATTA-
>CG4136:5CG4136G2:-
:3:2
-gTTAATTTa
>CG4136:5CG4136G3:-
:3:3
-cTTAATAGa
>CG4136:5CG4136G4:-
:3:4
-gCTAATTAt
>CG4136:5CG4136G5:-
:3:5
-cTTAATTAg
>CG4136:5CG4136G6:+:
3:6
gcCTAATTG-
>CG4136:5CG4136G7:+:
4:7
aaTTAATGA-
>CG4136:5CG4136G8:+:
4:8
cgATAATTA-
>CG4136:5CG4136G9:-
:3:9
-tTTAATAGg
>CG4136:5CG4136G10:-
:3:10
-cCTAATTAc
>CG4136:5CG4136G11:-
:3:11
-cCTAATGAg
>CG4136:5CG4136G12:-
:3:12
-gTTAATGAg
>CG4136:5CG4136H1:+:
4:13
taCTAATTG-
>CG4136:5CG4136H2:+:
4:14
aaCTAATTA-
>CG4136:5CG4136H3:+:
2:15
-cTTAATTAc
>CG4136:5CG4136H4:-
:4:16
--TTAATTAt
>CG4136:5CG4136H5:+:
3:17
gcTTAATTG-
>CG4136:5CG4136H6:+:
4:18
ctCTAATTA-
>CG4136:5CG4136H7:+:
4:19
gtTTAATTA-
>CG4136:5CG4136H9:-
:3:21
-tGTAATTGt
>CG4136:5CG4136H10:-
:3:22
-tTTAATTAa
>CG4136:5CG4136H11:+
:2:23
-cTTAATTAt
>Al:AlE5:+:3:2
--tTAATTAA
>Al:AlE2:-:3:1
-gcTAATTAA
>Al:AlE6:-:4:3
tgcTAATTAA
>Al:AlE7:-:4:4
gcaTAATTAA
>Al:AlE8:-:4:5
gacTAATTAA
>Al:AlE9:+:3:6
--tTAATTAA
>Al:AlE10:-:4:7
ttcTAATTAA
>Al:AlE11:-:3:8
-taTAATTAA
>Al:AlE12:-:4:9
tgcTAATTAA
>Al:AlF2:-:4:11
gacTAATTAA
>Al:AlF3:-:4:12
cgcTAATTGA
>Al:AlF4:-:4:13
tcaTAATTAA
>Al:AlF5:-:4:14
cgcTAATTGG
>Al:AlF6:-:4:15
accTAATTAA
>Al:AlF7:-:1:16
---TAATTAA
>Al:AlF8:+:3:17
--gTAATTAG
>Al:AlF9:-:4:18
gtcTAATTAA
>Al:AlF10:-:4:19
ggcTAATTAA
>Al:AlF11:-:4:20
cccTAATTGA
>Al:AlF12:-:3:21
-gtTAATTAA
>CG11294:CG11294A1:+
:2:1
-cTTAATTAt
>CG11294:CG11294A3:+
:2:2
-tTTAATTAg
>CG11294:CG11294A9:+
:2:4
-gTCAATTAg
>CG11294:CG11294A11:
+:4:6
aaCTAATTA-
>CG11294:CG11294A12:
-:1:7
-tCTAATTA-
>CG11294:CG11294B1:-
:2:8
taTTAATTAg
>CG11294:CG11294B2:-
:2:9
acCTAATTAg
>CG11294:CG11294B3:-
:4:10
--TTAATTAg
>CG11294:CG11294B4:-
:3:11
-aTTAATTAg
>CG11294:CG11294B5:+
:2:12
-aTCAATTAa
>CG11294:CG11294B6:-
:3:13
-tTTAATTAg
>CG11294:CG11294B7:+
:2:14
-aTTAATTAt
>CG11294:CG11294B8:-
:3:15
-gTTAATTAg
>CG11294:CG11294B9:-
:3:16
-aTTAATTAg
>CG11294:CG11294B10:
-:3:17
-aTTAATTAg
>Lim1:LimC10:+:2:9
-gTTAATTAg
>Lim1:LimC11:+:2:10
-gTTAATTGa
>Lim1:LimC12:+:1:11
--TTAATTAg
>Lim1:LimD1:+:3:12
taCTAATTA-
>Lim1:LimD2:+:2:13
-aTTAATTAg
>Lim1:LimD3:-:4:14
--TTAATTAc
>Lim1:LimD4:-:4:15
--TTAATTAt
>Lim1:LimD10:-:3:20
-aTTAATTAg
>Lim1:LimD11:-:4:21
--TTAATTAa
>Lim1:Lim1C2a:-:4:23
--CTAATTAt
>Lim1:Lim1C9a:-:4:28
--TTAATTAg
>Lim1:Lim1C10a:+:2:2
9
-gTTAATTAc
>Lim1:Lim1D2a:-:4:32
--TTAATTAg
>Lim1:Lim1D4a:+:3:33
aaTTAATTAt
```

```
>Lim1:Lim1D6a:-:4:34
--TTAATTAa
>Lim1:Lim1D9a:-:3:35
-tTTAATTAg
>Lim1:Lim1D10a:-:4:36
--TTAATTAg
>Lim1:Lim1D11a:+:2:37
-tTTAATTAc
>Hbn:HbnA1:+:1:1:mod
-aaTAATTAA
>Hbn:HbnA4:+:2:2
-aGTAATTAc
>Hbn:HbnA5:+:2:3:mod
-tgTAATTGA
>Hbn:HbnA6:+:2:4:mod
-caTAATTGA
>Hbn:HbnA7:+:3:5
gaTTAATTAa
>Hbn:HbnA8:-:4:6:mod
ctcTAATTGA
>Hbn:HbnA9:-:4:7:mod
cctTAATTAA
>Hbn:HbnA10:-:3:8
-cTTAATTTg
>Hbn:HbnB2:+:4:9
gaCTAATTA-
>Hbn:HbnB3:-:3:10
-tTTAATTGt
>Hbn:HbnB4:-:3:11
-aTTAATTGt
>Hbn:HbnB5:-:3:12
-tTTAATTAa
>Hbn:HbnB6:+:2:13
-cTTAATTAt
>Hbn:HbnB7:-:3:14
-tTTAATTGg
>Hbn:HbnB9:-:3:15
-tTTAATTAa
>Hbn:HbnB10:-:3:16
-gTTAATTAg
>Hbn:HbnB11:-:2:17
-aTTAATTGg
>Repo:RepoE2:-:3:1
cTTTAATTA-
>Repo:RepoE3:-:3:2
gGGTAATTA-
>Repo:RepoE4:-:2:3
aGCTAATTA-
>Repo:RepoE5:-:3:4
tGTTAATTA-
>Repo:RepoE7:+:2:6
tCTTAATTGa
>Repo:RepoE8:-:3:7
tTTTAATTA-
>Repo:RepoE9:+:2:8
cATTAATTAg
>Repo:Repo2A2:+:3:12
-TTAAATTGc
>Repo:Repo2A3:+:2:13
tAGTAATTGg
>Repo:Repo2A4:-:3:14
cTATAATTA-
>Repo:Repo2A5:+:1:15
aTGTAATTA-
>Repo:Repo2A6:+:3:16
-TTTAATTAg
>Repo:Repo2A7:+:2:17
tATTTATTGa
>Repo:Repo2A8:-:3:18
gCGTAATTA-
>Repo:Repo2A9:-:3:19
aGCTAATTA-
>Repo:Repo2A11:-:3:20
aTTTAATTG-
>Repo:Repo2A12:-:3:21
cATTAATTA-
>Repo:Repo2B1:+:3:22
-TTTAATTGa
>Repo:Repo2B2:+:3:23
-TTTAATTAg
>Repo:Repo2B3:-:3:24
gGCTAATTA-
>Repo:Repo2B4:-:3:25
cACTAATTA-
>Repo:Repo2B5:-:3:26
aCTTAATTA-
>Repo:Repo2B6:-:3:27
gTTTAATTA-
>Repo:Repo2B7:-:3:28
cATTAATTA-
>Repo:Repo2B8:+:3:29
-TTTAATTAa
>Repo:Repo2B9:+:2:30
tTTTAATTGg
>Repo:Repo2B10:-:3:31
aGTTAATTA-
>Repo:Repo2B11:-:3:32
aTTTAATTA-
>CG32105:CG32105G2:+:2:1
-TTTAATTAG
>CG32105:CG32105G3:+:2:2
-ACTAATTAA
>CG32105:CG32105G4:+:1:3
gTCTAATTGC
>CG32105:CG32105G5:+:2:4
-AATAATTAG
>CG32105:CG32105G6:+:1:5
cATTAATTGC
>CG32105:CG32105G7:+:2:6
-TCTAATTGG
>CG32105:CG32105G8:+:1:7
gTTATATTAA
>CG32105:CG32105G9:+:2:8
-TTTAATGAC
>CG32105:CG32105G10:+:1:9
cTATAATTGA
>CG32105:CG32105G12:+:2:11
-TTTAATTAA
>CG32105:CG32105H1:+:2:12
-TTTAATTAA
>CG32105:CG32105H2:+:1:13
tATTAATTAC
>CG32105:CG32105H3:-:1:14
-ATTAATTAG
>CG32105:CG32105H4:+:1:15
tCCAAATTAG
>CG32105:CG32105H5:+:2:16
-ATTAATTAG
>CG32105:CG32105H7:+:2:18
-ATTAATTAG
>CG32105:CG32105H9:+:1:19
aATTTAGTAG
>CG32105:CG32105H10:+:2:20
-ATTAATTAG
>CG32105:CG32105H11:-:1:21
-TTTAATTAA
>CG33980:CG33980A1:-:1:1
-TTTAATTAC
>CG33980:CG33980A2:+:1:2
-CATAATTAG
>CG33980:CG33980A4:+:1:4
-GTTAATTAG
>CG33980:CG33980A5:+:1:5
-GTTAATTAG
>CG33980:CG33980A6:+:1:6
-GTTAATTGC
>CG33980:CG33980A7:+:1:7
-GTTAATTGG
>CG33980:CG33980A8:+:1:8
-TTTAATTAA
>CG33980:CG33980A12:+:1:12
-GTTAATTGG
>CG33980:CG33980B2:-:1:13
-TCTAATTAG
>CG33980:CG33980B4:+:1:15
-TATAATTAG
>CG33980:CG33980B6:+:1:17
-TTTAATTAG
>CG33980:CG33980B7:+:1:18
-TCTAATTAG
>CG33980:CG33980B10:+:1:20
-ATTAATTAG
>Exex:ExexE1:-:3:1
-tCTAATTAa
>Exex:ExexE2:+:2:2:mod
-ctTAATTGG
>Exex:ExexE3:+:3:3
ggGTAATTAa
>Exex:ExexE4:+:4:4
gaGTAATTA-
>Exex:ExexE5:+:2:5
-cTTAATTAt
>Exex:ExexE6:+:4:6
caGTAATTA-
>Exex:ExexE7:+:4:7
cgTTAATTA-
>Exex:ExexE8:-:4:8
--GTAATAAg
>Exex:ExexE9:-:3:9
-aGTAATTAg
>Exex:ExexE10:+:4:10
gtGTAATTA-
>Exex:ExexE11:+:4:11
ggCTAATTA-
>Exex:ExexE12:+:4:12
gaGTAATTA-
>Exex:ExexF1:+:2:13
-gGTAATTAg
>Exex:ExexF2:-:4:14:mod
ttcTAATTGA
>Exex:ExexF3:-:4:15:mod
agtTAATTAC
>Exex:ExexF4:+:2:16
-tGTAATTAa
>Exex:ExexF5:-:4:17:mod
actTAATCAC
>Exex:ExexF6:+:2:18
-tCTAATTAa
>Exex:ExexF7:+:2:19
-aCTAATTAg
>Exex:ExexF8:+:4:20
ggCTAATTA-
>Exex:ExexF9:+:2:21
-ttTAATTGC
>Exex:ExexF10:+:4:22
atGTAATTA-
>Exex:ExexF11:+:3:23
cgCTAATTAa
>Rx:RxF2:+:3:1
-GCTAATTAc
>Rx:RxF3:-:3:2
tACTAATTA-
>Rx:RxF4:-:3:3
tACTAATTA-
>Rx:RxF5:+:2:4
gGTTAATTAg
>Rx:RxF7:-:1:6:mod
-gtTAATTGG
>Rx:RxF8:-:1:7:mod
-ttTAATTGA
>Rx:RxF12:-:1:10
-GCTAATTAa
>Rx:Rx2C1:+:1:11
tCATAATTA-
>Rx:Rx2C2:-:1:12
-ACTAATTAg
>Rx:Rx2C3:+:3:13
-GTTAATTAa
>Rx:Rx2C5:-:3:15
aCTTAATTA-
>Rx:Rx2C6:+:2:16
cACTAATTGg
>Rx:Rx2C7:-:1:17
-CCTAATTAt
>Rx:Rx2C8:-:3:18
tGTTAATTA-
>Rx:Rx2C9:-:1:19:mod
-atTAATTGA
>Rx:Rx2C10:-:1:20:mod
-caTAATTGA
```

```
>Rx:Rx2C11:-
:1:21:mod
-gcTAATTGA
>Rx:Rx2D1:-:1:23:mod
-ttTAATTGC
>Rx:Rx2D2:-:1:24
-GCTAATTAg
>Rx:Rx2D3:-:1:25
-AATAATTAg
>Rx:Rx2D4:-:1:26:mod
-ctTAATTGG
>Rx:Rx2D5:-:1:27
-GCTAATTAa
>Rx:Rx2D6:-:1:28:mod
-ttTAATTGA
>Rx:Rx2D8:-:1:29
-GCTAATTAa
>Rx:Rx2D9:+:3:30
-GCTAATTAa
>Rx:Rx2D10:-:1:31
-GCTAATTAa
>Rx:Rx2D11:-
:1:32:mod
-atTAATTGG
>Ro:RoG1:+:3:1
-GGTAATTAc
>Ro:RoG2:-:1:2:mod
-gcTAATTGC
>Ro:RoG3:-:3:3
tACTAATGA-
>Ro:RoG4:+:3:4
-TTTAATTAc
>Ro:RoG5:-:1:5:mod
-ccTAATTGA
>Ro:RoG6:-:3:6
tAATAATTA-
>Ro:RoG7:-:2:7
cTTTAATTA-
>Ro:RoG8:-:3:8
tACTAATGA-
>Ro:RoG9:-:3:9
cACTAATTA-
>Ro:RoG10:-:3:10
tGCTAATTA-
>Ro:RoG11:-:3:11
aGTTAATTA-
>Ro:RoG12:-:3:12
tGTTAATTA-
>Ro:RoH1:-:1:13:mod
-ttTAATTGA
>Ro:RoH2:-:1:14
-ACTAATTAa
>Ro:RoH3:-:2:15
tCGTAATGA-
>Ro:RoH4:-:3:16
cGCTAATTA-
>Ro:RoH5:-:1:17
-GCTAATTAa
>Ro:RoH6:-:1:18
-GCTAATTAa
>Ro:RoH7:-:1:19
-GCTAATTAa
>Ro:RoH8:-:3:20
tCTTAATTA-
>Ro:RoH9:-:1:21
-GCTAATTAa
>Ro:RoH10:-:3:22
cGGTAATTA-
>Ro:RoH11:-:1:23:mod
-ctTAATTGC
>Pph13:Pph13C1:+:3:1
tAATAATTA-
>Pph13:Pph13C2:+:3:2
tGATAATTG-
>Pph13:Pph13C3:+:1:3
-TCTAATTTa
>Pph13:Pph13C4:+:3:4
cCATAATTA-
>Pph13:Pph13C5:-:3:5
-ACTAATTCg
>Pph13:Pph13C6:+:3:6
aTATAATTA-
>Pph13:Pph13C7:+:3:7
aTTTAATTA-
>Pph13:Pph13C8:-:3:8
-ACTAATCAg
>Pph13:Pph13C9:+:1:9
:mod
-caTAATTGT
>Pph13:Pph13C10:+:1:
10:mod
-agTAATTGG
>Pph13:Pph13C11:+:3:
11
tCCTAATTA-
>Pph13:Pph13C12:+:1:
12:mod
-ggTAATTGG
>Pph13:Pph13D1:+:1:1
3
-AATAATTAg
>Pph13:Pph13D2:+:1:1
4:mod
-gcTAATTGT
>Pph13:Pph13D3:+:3:1
5
gCCTAATTA-
>Pph13:Pph13D4:+:1:1
6
-AGTAATTAc
>Pph13:Pph13D5:+:1:1
7:mod
-ttTAATTGA
>Pph13:Pph13D6:+:1:1
8
-ACTAATTAt
>Pph13:Pph13D7:-
:3:19
-ATTAATTAa
>Pph13:Pph13D9:+:1:2
1
-ACTAATTAa
>Pph13:Pph13D11:+:1:
22
-ATTAATTAa
>Inv:InvC1:+:2:1
gGTTAATTAt
>Inv:InvC2:-:1:2
-ACTAATTAa
>Inv:InvC3:+:2:3
tGCTAATTAt
>Inv:InvC5:-:3:4
cACTAATTA-
>Inv:InvC7:+:3:6
-TATAATTAc
>Inv:InvC8:-:1:7
-TCTAATTAa
>Inv:InvC11:+:1:10
cTCTAATTA-
>Inv:InvC12:+:3:11
-TTTAATTAg
>Inv:InvD1:+:2:12
aTTTAATTGg
>Inv:InvD2:+:2:13
cTCTAATTGa
>Inv:InvD3:+:2:14
tTTTAATTGa
>Inv:InvD4:+:2:15
tTCTAATTAg
>Inv:InvD5:+:2:16
tACTAATTAc
>Inv:InvD7:+:2:18
aACTAATTGg
>Inv:InvD8:+:2:19
tTTTAATTGa
>Inv:InvD11:+:3:22
-GTTAATTAc
>CG9876:CG9876A1:-
:2:1
tGCTAATTGt
>CG9876:CG9876A2:+:1
:2:mod
-cgTAATGAG
>CG9876:CG9876A3:+:3
:3
gACTAATTA-
>CG9876:CG9876A4:+:3
:4
tCGTAATTG-
>CG9876:CG9876A5:+:1
:5
-GTTAATTGa
>CG9876:CG9876A6:+:1
:6
-AGTTATTAa
>CG9876:CG9876A7:-
:2:7
cACTAATTGg
>CG9876:CG9876A8:+:3
:8
tATTAATTA-
>CG9876:CG9876A10:+:
1:10
-ACTTATTAa
>CG9876:CG9876B2:-
:2:11
cACTAATTGg
>CG9876:CG9876B3:+:2
:12
cAATAATTAg
>CG9876:CG9876B4:+:1
:13
-ACTAATTAg
>CG9876:CG9876B5:+:1
:14
-TTTAATTAa
>CG9876:CG9876B6:+:1
:15
-TTTAATTAc
>CG9876:CG9876B7:-
:3:16
-ATTAATTAa
>CG9876:CG9876B8:+:1
:17
-TCTAATTAt
>CG9876:CG9876B9:+:1
:18
-GCTAATTAa
>CG9876:CG9876B10:+:
1:19
-ACTAATTAa
>CG9876:CG9876B11:+:
1:20
-ATTAATTAt
>CG9876:CG9876B12:-
:3:21
-GATAATTGc
>En:eng1:-:2:1
cGCTAATTAg
>En:eng2:+:3:2
aTTTAATTA-
>En:eng3:-:2:3
cACTAATGAg
>En:eng4:+:3:4
gGGTAATTA-
>En:eng5:-:2:5
tGATAATTGc
>En:eng7:+:3:6
tGCTAATTA-
>En:eng8:-:2:7
gGTTAATTGa
>En:eng9:+:3:8
cGTTAATTA-
>En:eng10:+:3:9
aGGTAATTA-
>En:eng11:+:3:10
gGCTAATTA-
>En:eng12:-:1:11
tTTTAATTG-
>En:eng13:-:2:12
tTTTAATTGg
>En:eng14:+:3:13
aACTAATTA-
>En:eng15:+:3:14
tCTTAATTG-
>En:eng16:+:3:15
cGATAATTG-
>En:eng17:+:3:16
gACTAATTA-
>En:eng18:-:2:17
tTTTAATTGg
>En:eng19:-:3:18
-CTTAATTGa
>En:eng20:+:3:19
cGTTAATGA-
>En:eng21:+:1:20
-GCTAATTAa
>En:eng22:-:2:21
gTTTAATTGg
>En:eng23:-:2:22
-CTTAATTGa
>En:eng24:+:3:23
gGCTAATTA-
>CG32532:CG32432G1:+
:3:1
-gCTAATTAc
>CG32532:CG32432G2:+
:2:2
tgATAATTGg
>CG32532:CG32432G3:-
:1:3
--TTAATTAt
>CG32532:CG32432G4:+
:3:4
-tTTAATTTg
>CG32532:CG32432G5:-
:4:5
agGTAATTA-
>CG32532:CG32432G6:-
:2:6
-tCTAATTAt
>CG32532:CG32432G7:-
:3:7
gcTTAATGA-
```

```
>CG32532:CG32432G8:+
:3:8
-gGTAATTAc
>CG32532:CG32432G9:+
:2:9
taCTAATTGc
>CG32532:CG32432G10:
+:2:10
atCTAATTGg
>CG32532:CG32432G11:
+:3:11
-aTTAATTGa
>CG32532:CG32432G12:
+:3:12
-cTTAATTAg
>CG32532:CG32432H1:+
:2:13
cgTTAATTGg
>CG32532:CG32432H2:+
:3:14
-cCTAATTGg
>CG32532:CG32432H3:+
:3:15
-tTTAATTAa
>CG32532:CG32432H4:+
:3:16
-aTTAATTGg
>CG32532:CG32432H5:-
:2:17
-tTTAATTAt
>CG32532:CG32432H6:-
:2:18
-gTTAATTAg
>CG32532:CG32432H7:+
:2:19
tcCTAATTGa
>CG32532:CG32432H8:+
:3:20
-aTTAATTGg
>CG32532:CG32432H9:+
:3:21
-cTTAATTGa
>CG32532:CG32432H10:
+:3:22
-gTTAATTAa
>CG32532:CG32432H11:
-:4:23
acATAATGA-
>Unpg:UnpgC1:-:3:1
-ACTAATGAc
>Unpg:UnpgC2:+:3:2
aCGTAATTA-
>Unpg:UnpgC3:+:2:3
cGCTAATTAg
>Unpg:UnpgC4:+:3:4
aGATAATTG-
>Unpg:UnpgC5:+:1:5
-CGTAATTAg
>Unpg:UnpgC6:-:2:6
gCGTAATTAg
>Unpg:UnpgC7:-:3:7
-CTTAATTGc
>Unpg:UnpgC8:-:2:8
-TTTAATGAg
>Unpg:UnpgC9:-:3:9
-TTTAATTGa
>Unpg:UnpgC10:-:2:10
cCTTAATTGc
>Unpg:UnpgC11:+:1:11
-ACTAATTAt
>Unpg:UnpgD1:+:3:12
gCCTAATTA-
>Unpg:UnpgD2:-:2:13
cGCTAATTAt
>Unpg:UnpgD3:-:2:14
-CTTAATTGa
>Unpg:UnpgD4:-:3:15
-CTTAATTGg
>Unpg:UnpgD5:-:3:16
-ATTAATTGg
>Unpg:UnpgD6:-:2:17
tCCAAATTAg
>Unpg:UnpgD8:-:3:18
-ATTAATTAg
>Unpg:UnpgD9:+:1:19
-TGTAATTAt
>Unpg:UnpgD10:-:2:20
cCTTAATTAg
>Unpg:UnpgD11:-:2:21
tATTAATTAg
>PhdP:PhdPA1:+:2:1
-aTTAATTtg
>PhdP:PhdPA2:+:3:2
ggTTAATTac
>PhdP:PhdPA3:-:3:3
-gCTAATAgg
>PhdP:PhdPA4:+:2:4:m
od
-cctAATGAG
>PhdP:PhdPA8:+:4:6
gcATAATTt-
>PhdP:PhdPA9:+:4:7
taATAATTg-
>PhdP:PhdPA10:+:2:8
-cCTAATTgg
>PhdP:PhdPA11:+:4:9
aaATAATTa-
>PhdP:PhdPA12:+:4:10
gaTTAATTa-
>PhdP:PhdPB1:-:4:11
-gATAATTta
>PhdP:PhdPB2:-:4:12
-tTTTATTga
>PhdP:PhdPB3:+:5:13
cgCTAATTt-
>PhdP:PhdPB4:-:4:14
-cTTAATTcc
>PhdP:PhdPB5:+:1:15
--CTAATTat
>PhdP:PhdPB6:+:2:16
-cTTAATTat
>PhdP:PhdPB7:+:2:17
-tTTAATTaa
>PhdP:PhdPB10:+:4:18
acCTAATTa-
>CG7056:CG7056G1:+:3
:20
tATTAATTA-
>CG7056:CG7056G2:+:3
:21
tTTGAAGTA-
>CG7056:CG7056G3:+:3
:22
cCCTAATTG-
>CG7056:CG7056G4:+:2
:23
aATCAATTAc
>CG7056:CG7056G5:+:3
:24
tTTGAATTG-
>CG7056:CG7056G6:+:3
:25
aTTTAATTA-
>CG7056:CG7056G7:-
:3:26
-TCTAATTAc
>CG7056:CG7056G8:+:3
:27
tTTTAATAA-
>CG7056:CG7056G10:+:
2:28
aTCTTATTAc
>CG7056:CG7056G11:+:
3:29
gACCAATTG-
>CG7056:CG7056G12:+:
3:30
gACGAAGTA-
>Oc:OcA2:+:4:2
gcTTAAGCC-
>Oc:OcA3:+:3:3
cgATAATCCc
>Oc:OcA4:-:3:4
-tTTAAGCCc
>Oc:OcA5:-:3:5
-aATAATCCt
>Oc:OcA6:+:4:6
atATAATCC-
>Oc:OcA9:-:4:9
--CTAATCCg
>Oc:OcA10:-:2:10
ccTTAATCCt
>Oc:OcA11:-:3:11
-gTTAATCTg
>Oc:OcB1:-:3:12
-aTTAATCCa
>Oc:OcB2:+:4:13
tcATAATCC-
>Oc:OcB3:-:3:14
-cGTAATCCt
>Oc:OcB4:-:3:15
-cTTAATCGc
>Oc:OcB5:-:3:16
-cTTAATCCa
>Oc:OcB6:-:3:17
-tATAATCCc
>Oc:OcB7:+:4:18
agTTAATCC-
>Oc:OcB8:-:3:19
-aCTAATCCa
>Oc:OcB9:-:2:20
-aTTAATCCt
>Oc:OcB10:-:3:21
-cTTAATCCg
>Oc:OcB11:-:3:22
-gTTAATCCg
>Bcd:bcd1:-:2:1
tGTTAATCCg
>Bcd:bcd2:+:3:2
-TCTAATCCa
>Bcd:bcd3:-:2:3
cGTTAATCTc
>Bcd:bcd4:-:3:4
gTTTAATCC-
>Bcd:bcd5:-:3:5
cTATAATCC-
>Bcd:bcd6:-:2:6
tCTTAATCCc
>Bcd:bcd7:-:2:7
gCTTAATCCg
>Bcd:bcd8:-:3:8
gGTTAATCC-
>Bcd:bcd9:-:3:9
aGATAATCC-
>Bcd:bcd10:-:2:10
aGCTTATCC-
>Bcd:bcd11:-:3:11
gGGTAATCC-
>Bcd:bcd13:-:2:12
tGTTAATCC-
>Bcd:bcd14:+:3:13
-TATAATCCc
>Bcd:bcd15:-:2:14
gCGTAATCCa
>Bcd:bcd16:+:1:15
gCTTAAGCC-
>Bcd:bcd17:-:1:16
-GGTTATCCg
>Bcd:bcd18:-:2:17
tGTTAATCCc
>Bcd:bcd20:-:3:18
gCTTAATCC-
>Bcd:bcd21:-:3:19
tACTAATCC-
>Bcd:bcd22:-:3:20
tCCTAATCC-
>Bcd:bcd23:-:2:21
gGTTAATCCg
>Bcd:bcd24:+:3:22
-TCTAATCCa
>Ptx1:PtxG1:+:4:1
atCTAATCC-
>Ptx1:PtxG2:+:2:2
-tTTAATCCc
>Ptx1:PtxG4:-:4:3
--CTAATCCt
>Ptx1:PtxG6:-:3:4
-aCTAATCCt
>Ptx1:PtxG7:+:2:5
-cTTAATCCt
>Ptx1:PtxG8:-:3:6
-gTTAATCCc
>Ptx1:PtxG9:-:3:7
-gTTAATCCc
>Ptx1:PtxG10:+:4:8
aaTTAATCC-
>Ptx1:PtxG11:+:4:9
ggCTAATCC-
>Ptx1:PtxG12:+:3:10
tcGTAATCCc
>Ptx1:PtxH1:+:2:11
-gCTAATCCt
>Ptx1:PtxH2:+:4:12
tcTTAATCC-
>Ptx1:PtxH4:+:3:14
cgTTAATCTc
>Ptx1:PtxH5:+:3:15
ccTTAATCCc
>Ptx1:PtxH6:+:3:16
cgTTAATCCc
>Ptx1:PtxH7:+:3:17
ggTTAATCCc
>Ptx1:PtxH8:+:3:18
tcTTAATCCc
>Ptx1:PtxH9:+:3:19
cgTTAATCCc
>Ptx1:PtxH10:+:3:20
cgTTAATCCc
>Ptx1:PtxH11:+:3:21
acTTAATCCc
>Gsc:GscA1:+:2:1
tCGTAATCGg
>Gsc:GscA2:-:3:2
aAGTAATCC-
```

```
>Gsc:GscA3:+:2:3
gACTAATCTt
>Gsc:GscA4:+:3:4
-AATAATCCt
>Gsc:GscA5:+:2:5
tACTAATCTt
>Gsc:GscA6:+:3:6
-GATAATCTg
>Gsc:GscA7:+:3:7
-AGTAATCCt
>Gsc:GscA8:+:3:8
-CTTAATCGc
>Gsc:GscA9:-:3:9
tCGTAATCC-
>Gsc:GscA10:-:3:10
cGTTAATCT-
>Gsc:GscA11:+:3:11
-CATAATCCt
>Gsc:GscA12:-:3:12
cAATAATCC-
>Gsc:GscB1:-:3:13
cGTTAATCT-
>Gsc:GscB3:+:2:14
tGTTAATCCc
>Gsc:GscB4:+:3:15
-CTTAATCTc
>Gsc:GscB5:+:3:16
-CTTAATCCg
>Gsc:GscB6:+:3:17
-TTTAATCCg
>Gsc:GscB7:-:3:18
cCCTAATCC-
>Gsc:GscB8:+:2:19
tGTTAATCCc
>Gsc:GscB9:+:2:20
tCCTAATCCa
>Gsc:GscB10:+:3:21
-ACTAATCCa
>Gsc:GscB11:+:1:22
cATTAATCC-
>Oct:oct1:+:3:1
cGATAATGA-
>Oct:oct2:+:3:2
gTTTAATGA-
>Oct:oct3:+:3:3
tTTTAATAA-
>Oct:oct4:+:3:4
aCATAATTT-
>Oct:oct5:+:2:5
cTATAATTAg
>Oct:oct6:+:3:6
aGATAATTT-
>Oct:oct7:+:3:7
aTATAATAA-
>Oct:oct8:+:3:8
cATTAATCA-
>Oct:oct10:-:3:9
-TTTAATCAc
>Oct:oct11:-:3:10
-CCTAATGAg
>Oct:oct13:+:3:11
aGCTAATTA-
>Oct:oct14:+:3:12
cATTAATTT-
>Oct:oct15:+:3:13
gGTAAATGA-
>Oct:oct16:-:3:14
-TTTAATGAg
>Oct:oct17:+:3:15
cCCTAATTA-
>Oct:oct18:-:3:16
-GTTAATGAg
>Oct:oct19:+:3:17
tCATAATCA-
>Oct:oct20:+:3:18
gTTTAATTG-
>Oct:oct21:+:3:19
aGATAATTC-
>Oct:oct22:+:3:20
gTTTAATTT-
>Oct:oct23:+:3:21
cTTTAATTT-
>Oct:oct24:+:3:22
tGATAATTA-
>Bsh:BshE1:+:1:1
-CCTAATGGG
>Bsh:BshE2:+:1:2
-TTTAATCGA
>Bsh:BshE3:+:1:3
-CTTAATGAC
>Bsh:BshE4:+:1:4
-TGTAATTGG
>Bsh:BshE6:+:1:5
-ATTAATTGC
>Bsh:BshE8:+:1:7
-CTTAATGAT
>Bsh:BshE11:+:1:10
-TTTAATCGA
>Bsh:BshE12:+:1:11
-CTTAATCGA
>Bsh:BshF1:+:1:12
-TCTAATGAG
>Bsh:BshF4:+:1:14
-TCTAATGAG
>Bsh:BshF6:+:1:15
-GTTAATTGC
>Bsh:BshF7:+:1:16
-ATTAATTAG
>Bsh:BshF8:+:1:17
-ATTAATTAG
>Bsh:BshF9:+:1:18
-CTTAACGAG
>Bsh:BshF10:+:1:19
-TATAATTGG
>Bsh:BshF11:+:1:20
-TGTAATTGG
>Tup:TupE1:-:3:1
-GGTAATTGa
>Tup:TupE2:-:2:2
cGCTAATTAg
>Tup:TupE3:-:2:3:mod
CACTAATGt-
>Tup:TupE4:+:3:4
tATTAATGG-
>Tup:TupE5:-:3:5
-CTTAATAGa
>Tup:TupE6:-:3:6
-TATAATGGt
>Tup:TupE7:-:3:7
-GATAATTAa
>Tup:TupE9:-:3:9
-GTTAAGTGg
>Tup:TupE10:-:3:10
-CATAATTGa
>Tup:TupE11:+:3:11
cCTTAATGG-
>Tup:TupE12:+:3:12
cCTTAATGG-
>Tup:TupF2:-:3:14
-CTTAATTGc
>Tup:TupF4:-:3:15
-CTAAATGGa
>Tup:TupF6:-:3:17
-CTTAATGGa
>Tup:TupF7:+:3:18
cGATAAGTG-
>Tup:TupF8:-:3:19
-CTTAATTGa
>NK7:NK7E1:+:3:1
aTATAATGA-
>NK7:NK7E2:+:3:2
aATTAAGTG-
>NK7:NK7E3:+:3:3
aATTAAGTG-
>NK7:NK7E4:-:3:4
-TTTAAATAc
>NK7:NK7E6:+:3:5
aAATAATTA-
>NK7:NK7E7:+:2:6
gAGTAAATGa
>NK7:NK7E8:+:3:7
aACTAATTG-
>NK7:NK7E10:-:3:8
-GTTAAGTGg
>NK7:NK7F1:+:3:10
aGCTAATTG-
>NK7:NK7F4:+:3:13
cGCTAATGA-
>NK7:NK7F6:-:1:15
tTTTAATTG-
>NK7:NK7F7:+:3:16
cCTTAATAG-
>NK7:NK7F8:-:3:17
-CTTAATTGc
>NK7:NK7F9:-:3:18
-CTTAATTGg
>NK7:NK7F10:+:3:19
gCTTAATTA-
>NK7:NK7F11:+:3:20
gACTAATTA-
>NK7:NK72A2:-:2:21
gTTTAATGAt
>NK7:NK72A3:+:3:22
aCTTAAGTG-
>NK7:NK72A5:-:3:24
-TTTAATGGa
>NK7:NK72A6:-:3:25
-CTTAATAGc
>NK7:NK72A7:-:3:26
-TATAATTGt
>NK7:NK72A8:-:2:27
tTTTAATTAa
>NK7:NK72A9:+:3:28
aCATAATAG-
>NK7:NK72A10:+:3:29
aGATAATGA-
>NK7:NK72A11:-:2:30
tATTGATAGc
>NK7:NK72A12:+:3:31
cATTAAATG-
>NK7:NK72B2:+:3:32
gGTTAAATA-
>NK7:NK72B4:+:3:34
aATTAAGTG-
>NK7:NK72B5:-:2:35
tATTAATAGt
>NK7:NK72B6:-:3:36
-ATTGATGGt
>NK7:NK72B7:+:3:37
tATTAATGA-
>NK7:NK72B8:-:3:38
-CTTAATTAc
>NK7:NK72B9:+:3:39
cACTAATTA-
>NK7:NK72B10:-:3:40
-TTTAATAGg
>NK7:NK72B11:-:2:41
cCTTAATTGg
>CG13424:CG13424E1:+
:2:1
gCCTAATTGa
>CG13424:CG13424E2:+
:3:2
-TTTAATTAg
>CG13424:CG13424E3:-
:3:3
gTTTAATTA-
>CG13424:CG13424E4:+
:3:4
-TATAATTGt
>CG13424:CG13424E9:+
:2:8
cATTAATGGg
>CG13424:CG13424E11:
+:2:10
tATTAATTAg
>CG13424:CG13424E12:
-:3:11
tGTTAATGA-
>CG13424:CG13424F2:-
:3:12
aACTAATGG-
>CG13424:CG13424F3:+
:3:13
-GGTAATTGc
>CG13424:CG13424F4:+
:3:14
-TTTAATAGg
>CG13424:CG13424F5:-
:3:15
gTCTAATGA-
>CG13424:CG13424F7:+
:2:17
gCATAATTGg
>CG13424:CG13424F8:+
:2:18
gTTTAATTGa
>CG13424:CG13424C1:+
:2:23
gGTTAATTAg
>CG13424:CG13424C5:-
:3:26
gATTAATTA-
>CG13424:CG13424C6:-
:3:27
tACTAATTG-
>CG13424:CG13424C7:+
:3:28
-CTTAATAGc
>CG13424:CG13424C8:-
:3:29
aGTTAATAG-
>CG13424:CG13424C12:
+:2:33
tCCTAATTGg
>CG13424:CG13424D1:+
:2:34
tACTAATTGc
>CG13424:CG13424D2:+
:2:35
tCCTAATTAt
>Hgtx:HgtxC1:+:3:1
-ATTAATAGg
>Hgtx:HgtxC2:-:1:2
```

```
-AGTAATTGa
>Hgtx:HgtxC3:-:2:3
aGCTAATTA-
>Hgtx:HgtxC4:+:3:4
-GATAATTGg
>Hgtx:HgtxC5:-:3:5
tTGTAATTA-
>Hgtx:HgtxC6:+:1:6
tTTTAATGA-
>Hgtx:HgtxC7:-:3:7
aTGTAATTA-
>Hgtx:HgtxC8:+:3:8
-GTTAATTGa
>Hgtx:HgtxC9:+:3:9
-CATAATGAg
>Hgtx:HgtxD1:+:3:11
-GATAATTGc
>Hgtx:HgtxD2:+:3:12
-TCTAATTAc
>Hgtx:HgtxD3:+:3:13
-AATAATGAa
>Hgtx:HgtxD4:+:2:14
aATTAATTAa
>Hgtx:HgtxD5:-:3:15
tTTTAATTA-
>Hgtx:HgtxD6:-:3:16
cTTTAATTA-
>Hgtx:HgtxD7:+:2:17
-TTTAATGAg
>Hgtx:HgtxD8:-:3:18
cTTTAATTA-
>Hgtx:HgtxD9:-:3:19
tTTTAATGA-
>Hgtx:HgtxD10:-:3:20
aGTTAATGA-
>Hgtx:HgtxD11:-:3:21
gTTTAATTA-
>Ems:emsA1:+:2:1
-TATAATTAa
>Ems:emsA2:+:1:2
cATAAATGG-
>Ems:emsA3:-:3:3
aTTTTATGA-
>Ems:emsA5:+:2:5
tATTAGTGAa
>Ems:emsA6:+:3:6
-AGTAATGAc
>Ems:emsA7:+:3:7
-ATTAATGGa
>Ems:emsA8:+:3:8
-GTTAATCAc
>Ems:emsA9:-:3:9
cATAAATGA-
>Ems:emsA10:+:3:10
-TTTAATTAg
>Ems:emsA11:+:2:11
cAATAATTGg
>Ems:emsA12:-:3:12
cATAAATTA-
>Ems:emsB2:+:3:14
-TATAATTGg
>Ems:emsB3:+:3:15
-TTTAATGGa
>Ems:emsB5:-:3:16
tTCTAATGA-
>Ems:emsB6:-:3:17
gCCTAATGA-
>Ems:emsB7:-:3:18
cACTAATTA-
>Ems:emsB8:-:3:19
gTCTAATGA-
>Ems:emsB9:+:2:20
cTCTAATTGg
>Ems:emsB10:+:2:21
cTCTAATTAg
>Ems:emsB11:+:2:22
cGCTAATTAg
>Otp:OTPG1:+:1:1
-TTCAATTAt
>Otp:OTPG2:-:2:2
tTCTAATTTg
>Otp:OTPG3:-:3:3
-ATTAATTGc
>Otp:OTPG4:+:2:4
gTTTAATCA-
>Otp:OTPG5:+:3:5
tTGTAATTA-
>Otp:OTPG6:-:2:6
aATTAATGGt
>Otp:OTPG8:+:4:7
cGCTAATTT-
>Otp:OTPG9:-:2:8
gTTTAATGAg
>Otp:OTPG10:-:1:9
cTTTAATGA-
>Otp:OTPG11:+:1:10
-TTTAATTAa
>Otp:OTPG12:+:1:11
-CCTAATTAg
>Otp:OTPH1:+:2:12
cCTTAATTAa
>Otp:OTPH2:-:2:13
cTGTAATTAg
>Otp:OTPH3:+:2:14
gCATAATTA-
>Otp:OTPH4:-:3:15
-GTTAATGAa
>Otp:OTPH5:-:2:16
cCTTAATTGa
>Otp:OTPH6:-:3:17
-TTTAATTAa
>Otp:OTPH7:-:3:18
-TCTAATTAa
>Otp:OTPH10:+:3:19
aACTAATTA-
>Otp:OTPH11:-:2:20
gTCTAATTAa
>Ftz:FtzG1:+:3:1
tCATAATTG-
>Ftz:FtzG3:-:3:3
-GTTAATGGg
>Ftz:FtzG4:+:3:4
tCTTAATTA-
>Ftz:FtzG6:+:3:6
tACTAATGA-
>Ftz:FtzG7:+:3:7
tATAAATGA-
>Ftz:FtzG8:+:3:8
tTTTAATTG-
>Ftz:FtzG9:+:3:9
gCCTAATGA-
>Ftz:FtzG10:+:3:10
aGTTAATTA-
>Ftz:FtzG11:-:3:11
-GTTAATGAt
>Ftz:FtzG12:+:3:12
gCTTAATTA-
>Ftz:FtzH1:+:3:13
gCTTAATGG-
>Ftz:FtzH3:+:3:15
cGTTAATTA-
>Ftz:FtzH5:+:3:16
aGTTAATGA-
>Ftz:FtzH7:+:3:18
tGTTAATTA-
>Ftz:FtzH8:+:3:19
gGTTAATTA-
>Ftz:FtzH9:+:3:20
tTTTAATGA-
>Ftz:FtzH10:+:3:21
tATTAATTA-
>Ftz:FtzH11:+:3:22
tGTTAATGA-
>Antp:AntpA02:-:3:1
cCTTAATTA-
>Antp:AntpA03:-:3:2
aTTTAATTA-
>Antp:AntpA04:-:3:3
gCTTAATGA-
>Antp:AntpA05:-:3:4
tGTTAATGA-
>Antp:AntpA07:-:3:6
gTTTAATGA-
>Antp:AntpA08:-:3:7
gCTTAATGA-
>Antp:AntpA10:-:3:8
tACTAATTA-
>Antp:AntpA11:-:3:9
cCTTAATGG-
>Antp:AntpB01:-:3:11
tCATAATTA-
>Antp:AntpB02:-:3:12
tGTTAATTA-
>Antp:AntpB04:-:3:13
aTTTAATTA-
>Antp:AntpB06:-:3:14
gTTTAATGA-
>Antp:AntpB07:-:3:15
tTTTAATGA-
>Antp:AntpB09:-:3:17
gTTTAATTA-
>Antp:AntpB11:-:3:18
tTTTAATGA-
>Antp:AntpB12:-:3:19
cTTTAATGA-
>Zen2:Zen2A2:+:2:2
TTATAATGA-
>Zen2:Zen2A4:+:2:4
CCGTAATTA-
>Zen2:Zen2A6:+:2:6
TACTAATTG-
>Zen2:Zen2A7:+:2:7
GAGTAATGA-
>Zen2:Zen2A8:+:2:8
TGATAATGA-
>Zen2:Zen2A9:+:2:9
CCGTAATTA-
>Zen2:Zen2A10:+:2:10
TAGTAATTA-
>Zen2:Zen2B3:+:2:14
TACTAATTA-
>Zen2:Zen2B4:-:2:15
TACTAATGAc
>Zen2:Zen2B9:+:2:20
CCTTAATTA-
>Zen2:Zen2B11:+:2:22
AGTTAATGA-
>Zen2:Zen2E1:-:2:23
ACTTAATTAt
>Zen2:Zen2E3:+:2:25
ATTTAATTA-
>Zen2:Zen2E4:-:2:26
TCATAATTGa
>Zen2:Zen2E5:-:2:27
TTTTAATGAc
>Zen2:Zen2E6:+:2:28
GCTAAATTA-
>Zen2:Zen2E7:+:2:29
CACTAACGA-
>Zen2:Zen2E9:-:2:31
CCTTAATTGc
>Zen2:Zen2E10:+:2:32
CCTTAATTA-
>Zen2:Zen2E11:-:2:33
TTTTAATTGc
>Zen2:Zen2F1:+:2:35
CCGTAATGA-
>Zen2:Zen2F3:+:2:37
TACTAATGA-
>Zen2:Zen2F5:+:2:39
TGCTAATTA-
>Zen2:Zen2F6:+:2:40
AGCTAATTA-
>Zen2:Zen2F7:+:1:41
TGTTAATGA-
>Zen2:Zen2F11:+:2:45
TGTTAATTA-
>Slou:SlouE2:+:3:2
-CTTAATGGc
>Slou:SlouE3:+:2:3
tTTTAATGAg
>Slou:SlouE4:-:3:4
gAGTAATGA-
>Slou:SlouE5:-:2:5
cGTCAATTAc
>Slou:SlouE6:+:3:6
-GCTAATTGt
>Slou:SlouE7:-:1:7
-GTTAATTAt
>Slou:SlouE8:+:3:8
-TTTAATCGg
>Slou:SlouE9:+:3:9
-TTTAATAAt
>Slou:SlouE10:+:1:10
aATTAATTG-
>Slou:SlouE11:-:3:11
gTCTAATGA-
>Slou:SlouE12:-:3:12
tTCTAATGA-
>Slou:SlouF1:-:3:13
gGCTAATTA-
>Slou:SlouF2:-:3:14
cTCTAATTG-
>Slou:SlouF3:-:2:15
cTATAATTAg
>Slou:SlouF4:-:1:16
-GTTAATTAg
>Slou:SlouF5:+:3:17
-CTTAATTGg
>Slou:SlouF6:+:3:18
-TATAATTGt
>Slou:SlouF7:+:3:19
-CTTAATTGa
>Slou:SlouF8:+:1:20
cTTTAGTAG-
>Slou:SlouF9:+:3:21
-TTTAATTGg
>Slou:SlouF10:+:3:22
-TTTAATTAc
>Slou:SlouF11:+:3:23
-GATAATTGg
>Btn:BtnE1:+:3:1
-TTTAATGGc
>Btn:BtnE2:-:3:2
```

```
cAGTAATGA-
>Btn:BtnE3:-:3:3
aCTTGATTA-
>Btn:BtnE4:+:3:4
-ATTAATGTa
>Btn:BtnE5:+:3:5
-CTTAATGGg
>Btn:BtnE6:-:3:6
cTGTAATTA-
>Btn:BtnE7:+:3:7
-GGTAATGAc
>Btn:BtnE8:+:3:8
-CTTAACGAc
>Btn:BtnE9:-:3:9
gCCTAATTA-
>Btn:BtnE10:+:3:10
-TATAATTGc
>Btn:BtnE11:-:1:11
-GATAATTAa
>Btn:BtnE12:-:3:12
tCGTAATGA-
>Btn:BtnF1:+:3:13
-TATAATGAt
>Btn:BtnF2:-:1:14
-CTTAATTAa
>Btn:BtnF3:+:3:15
-ATTAATGAc
>Btn:BtnF4:-:3:16
aTATAATGA-
>Btn:BtnF5:+:3:17
-GATAATTAg
>Btn:BtnF6:+:3:18
-ATTAATTAc
>Btn:BtnF7:+:3:19
-CTTAATGAc
>Btn:BtnF8:-:3:20
aCTTAATGA-
>Btn:BtnF9:-:3:21
cCTTAATGA-
>Btn:BtnF10:-:3:22
cGTTAATGA-
>Btn:BtnF11:-:3:23
aGTTAATGA-
>Dfd:dfd1:+:3:1
-CTTAATGAa
>Dfd:dfd2:+:3:2
-ATTAATGAc
>Dfd:dfd3:+:3:3
-TTTAATGAt
>Dfd:dfd4:+:3:4
-ATTAATGAc
>Dfd:dfd5:+:2:5
tGTTAATGAc
>Dfd:dfd6:+:3:6
-CTTAATTAg
>Dfd:dfd7:+:3:7
-ATTAATTAt
>Dfd:dfd8:+:3:8
-CTTAATTAg
>Dfd:dfd9:+:3:9
-ATTAATGGg
>Dfd:dfd10:-:3:10
tTTTAATGA-
>Dfd:dfd11:+:3:11
-TTTAATAGc
>Dfd:dfd12:-:3:12
cACTAATGA-
>Dfd:dfd13:-:2:13
tCGTAATGA-
>Dfd:dfd14:-:3:14
gCTTAATGG-
>Dfd:dfd15:-:3:15
tCGTAATTA-
>Dfd:dfd16:+:2:16
-AGTAATGAg
>Dfd:dfd17:+:3:17
-CTTAATGAa
>Dfd:dfd18:+:3:18
-CCTAATGAc
>Dfd:dfd19:-:3:19
aCCTAATGA-
>Dfd:dfd20:-:3:20
gGATAATGA-
>Dfd:dfd21:-:3:21
gACTAATGA-
>Dfd:dfd22:+:3:22
-GTTAATGAt
>Dfd:dfd23:-:3:23
cGTTAATGA-
>Dfd:dfd24:-:3:24
aATTAATGA-
>Scr:ScrE04:+:1:1
CCTTAATGA-
>Scr:ScrE05:+:2:2
ACATAATGA-
>Scr:ScrE08:+:2:4
TACTAATTA-
>Scr:ScrE12:+:2:7
CCTTAATGA-
>Scr:ScrF02:+:1:9
CGATAATGA-
>Scr:ScrF04:+:2:10
ACTTAATGA-
>Scr:ScrF06:+:2:11
CGCTAATGA-
>Scr:ScrF11:+:2:15
CATTAATGA-
>Scr:ScrF12:+:2:16
CGTTAATGA-
>Scr:ScrG1:-:2:17
CTTTAATTGc
>Scr:ScrG2:+:2:18
AATTAATGA-
>Scr:ScrG3:+:2:19
GTTTAATGA-
>Scr:ScrG5:+:2:20
CACTAATTA-
>Scr:ScrG6:-:2:21
CGTTAATTGc
>Scr:ScrG7:+:2:22
TGTTAATTA-
>Scr:ScrG8:+:1:23
CACTAATTA-
>Scr:ScrG9:+:2:24
CGTTAATTA-
>Scr:ScrG12:+:2:26
CGCTAATGA-
>Scr:ScrH1:+:1:27
TGTTAATGA-
>Scr:ScrH2:+:2:28
ACATAATGA-
>Scr:ScrH3:+:2:29
CGTTAATGA-
>Scr:ScrH4:+:2:30
TATTAATGA-
>Scr:ScrH9:+:2:32
TACTAATGA-
>Scr:ScrH10:+:2:33
ATTTAATGA-
>Scr:ScrH11:+:2:34
CACTAATGA-
>Zen:ZenC1:+:2:1
CCTTAATTA-
>Zen:ZenC3:+:1:2
GACTAATTA-
>Zen:ZenC4:+:2:3
GGCTAATTA-
>Zen:ZenC5:+:2:4
TAATAATGA-
>Zen:ZenC8:+:2:7
ACCTAATGA-
>Zen:ZenC9:+:2:8
GTTTAATGA-
>Zen:ZenC10:+:2:9
CCCTAATGA-
>Zen:ZenC11:+:2:10
CGCTAATGA-
>Zen:ZenD2:+:2:12
CCCTAATGA-
>Zen:ZenD3:+:2:13
GCTTAATGA-
>Zen:ZenD4:+:2:14
ACCTAATGA-
>Zen:ZenD5:+:2:15
TGCTAATTA-
>Zen:ZenD6:+:2:16
CCTTAATGA-
>Zen:ZenD7:+:2:17
TTTTAATTA-
>Zen:ZenD9:+:2:18
ACATAATGA-
>Zen:ZenD11:+:1:19
TGCTAATGA-
>Pb:Pb1:+:3:1
-TCTAATGAc
>Pb:Pb2:+:3:2
-GTTAATTAc
>Pb:Pb3:+:3:3
-TTTAATTAc
>Pb:Pb4:+:3:4
-GTTAATGAc
>Pb:pBG2:-:3:5
tCCTAATTA-
>Pb:pBG3:+:2:6
-TTTAATGAg
>Pb:pBG4:+:3:7
-TATAATTAc
>Pb:pBG5:-:3:8
tGTTAATTA-
>Pb:pBG6:+:3:9
-ATTAATTAc
>Pb:pBG7:-:3:10
aCATAATGA-
>Pb:pBG8:-:3:11
gACTAATGA-
>Pb:pBG9:-:3:12
cGTTAATGA-
>Pb:pBG11:+:3:13
-CTTAATGAg
>Pb:pBG12:-:1:14
-GCTAATTAa
>Pb:pBH1:+:3:15
-TTTAATTAc
>Pb:pBH3:+:3:16
-TATAATTAc
>Pb:pBH4:-:1:17
-GATAATTAt
>Pb:pBH6:-:3:19
tGCTAATGA-
>Pb:pBH7:-:1:20
-GCTAATTAa
>Pb:pBH8:-:3:21
gGTTAATGA-
>Pb:pBH9:+:1:22
cATTAATGA-
>Pb:pBH10:+:3:23
-TCTAATGAg
>Pb:pBH11:+:3:24
-TCTAATTAc
>Pb:pBH12:-:3:25
tGCTAATTA-
>Lab:LabE1:+:1:1
cgTTAATGA-
>Lab:LabE2:-:4:2
ccTTAATTA-
>Lab:LabE3:-:4:3
ggCTAATTA-
>Lab:LabE4:+:3:4
-aTTAATTAt
>Lab:LabE6:-:4:6
atTTAATTA-
>Lab:LabE7:+:3:7
-aTTAATTAg
>Lab:LabE8:-:4:8
agTTAATTA-
>Lab:LabE9:-:4:9
taCTAATTA-
>Lab:LabE10:+:4:10
--TTAAAGAa
>Lab:LabF1:+:3:12
-cTTAATGAc
>Lab:LabF5:-:4:16
tgTTAATTA-
>Lab:LabF9:+:3:20
-aTTAATGAc
>Lab:LabF11:-:4:22
gtCTAATGA-
>Lab:LabH6:-:4:25
tgTTAATTA-
>Lab:LabH7:-:4:26
tcTTAATTA-
>Lab:LabH12:-:4:31
acATAATGA-
>AbdA:AbdAG02:-:3:1
-tTTAATTAc
>AbdA:AbdAG03:+:4:2
cgTTAATGA-
>AbdA:AbdAG04:+:4:3
ttTTAATTA-
>AbdA:AbdAG06:+:1:5
--TTAATTAc
>AbdA:AbdAG07:+:4:6
tcTTTATTA-
>AbdA:AbdAG10:+:3:9
caCTAATTA-
>AbdA:AbdAG11:+:2:10
-cATAATTA-
>AbdA:AbdAG12:+:4:11
ttTTAATTA-
>AbdA:AbdAH03:-:3:14
-tTTAATGAc
>AbdA:AbdAH04:+:4:15
ttTTTATGA-
>AbdA:AbdAH05:+:4:16
taCTAATTC-
>AbdA:AbdAH06:-:3:17
-tTTAATTGc
>AbdA:AbdAH07:-:4:18
--TTAAAGAa
>AbdA:AbdAH08:+:4:19
atTTAATTA-
>AbdA:AbdAH09:+:4:20
cgCTAATGA-
>AbdA:AbdAH10:+:4:21
```

```
gtTTAATGA-
>AbdA:AbdAH11:-:3:22
-cTTAATTGc
>AbdA:AbdAH12:+:4:23
tcTTAATTAc
>Ap:ApA2:-:2:1
gGTTAATGAt
>Ap:ApA3:+:3:2
gACTAATTG-
>Ap:ApA4:+:3:3
gGTTAATTA-
>Ap:ApA5:+:3:4
aAATAATGA-
>Ap:ApA6:-:2:5
cGCTAATTAg
>Ap:ApA7:+:3:6
tGCTAATTG-
>Ap:ApA8:+:1:7
-GCTAATTAa
>Ap:ApA9:-:2:8
tCATAATTGg
>Ap:ApA10:-:2:9
cCTTAATTAg
>Ap:ApA11:+:1:10
-ACAAATTAa
>Ap:ApB3:+:3:12
cCTTAATGA-
>Ap:ApB4:-:2:13
aCTTAATTAg
>Ap:ApB5:-:3:14
-TTTAATGAg
>Ap:ApB6:+:1:15
-ACTAATTAa
>Ap:ApB7:+:3:16
cGTTAATGA-
>Ap:ApB8:+:3:17
cGCTAATTA-
>Ap:ApB9:+:3:18
gACTAATTA-
>Ap:ApB10:+:2:19
cGCTAATGA-
>Ap:ApB11:-:2:20
tGCTAATTAg
>Ind:IndC1:+:2:1
GATTAATTA-
>Ind:IndC2:+:1:2
CGCTAATGA-
>Ind:IndC3:+:2:3
ACCTAATGA-
>Ind:IndC4:-:1:4
TATTAAGTG-
>Ind:IndC5:-:1:5
CTCTAATTA-
>Ind:IndC7:+:2:7
CGTTAATGA-
>Ind:IndC8:+:2:8
GATTAATGA-
>Ind:IndC9:+:2:9
CGTTAATGA-
>Ind:IndC10:-:2:10
TACTAATTAc
>Ind:IndC11:-:2:11
TATTAATTAc
>Ind:IndC12:-:2:12
TGTTAATGAg
>Ind:IndD1:-:2:13
CCTTAATTAg
>Ind:IndD3:-:2:14
TTCTAATTAc
>Ind:IndD4:+:2:15
CACTAATGA-
>Ind:IndD5:+:2:16
TGCTAATTA-
>Ind:IndD6:+:2:17
TCCTAATTA-
>Ind:IndD7:+:2:18
CACTAATGA-
>Ind:IndD8:-:2:19
TATTAATTGc
>Ind:IndD9:+:2:20
CACTAATTA-
>Ind:IndD10:+:2:21
CCCTAATTA-
>Ind:IndD11:-:2:22
TGCTAATTAg
>CG18599:CG18599C1:-
:3:1
-tTTAATTGa
>CG18599:CG18599C2:+
:4:2
tgCTAATGA-
>CG18599:CG18599C3:+
:4:3
tgCTAATTA-
>CG18599:CG18599C4:+
:4:4
aaTTAATTA-
>CG18599:CG18599C5:+
:4:5
ctTTAATTA-
>CG18599:CG18599C6:+
:1:6
--TTAATTAa
>CG18599:CG18599C7:+
:3:7
aaTTAATTA-
>CG18599:CG18599C8:+
:4:8
ccCTAATTG-
>CG18599:CG18599C9:+
:4:9
gaTTAATGA-
>CG18599:CG18599C10:
+:4:10
atCTAATTA-
>CG18599:CG18599C11:
+:4:11
caCTAATGA-
>CG18599:CG18599C12:
+:4:12
cgCTAATTA-
>CG18599:CG18599D2:-
:2:13
ttATAATGAg
>CG18599:CG18599D9:+
:4:16
ccGTAATTA-
>CG18599:CG18599D10:
+:1:17
--TTAATCAc
>CG18599:CG18599D11:
+:4:18
acATAATGA-
>CG18599:CG18599F1:+
:4:19
cgTTAATTA-
>CG18599:CG18599F2:+
:4:20
taATAATGA-
>CG18599:CG18599F3:+
:4:21
cgCTAATTA-
>CG18599:CG18599F4:+
:4:22
cgTTAATTA-
>CG18599:CG18599F5:+
:3:23
caCTAATTA-
>CG18599:CG18599F6:+
:1:24
--CTAATTAg
>CG18599:CG18599F7:+
:3:25
gcTTAATGA-
>CG18599:CG18599F11:
+:4:27
ttATAATGA-
>CG18599:CG18599F12:
+:4:28
gcCTAATTA-
>Lbe:IbeG2:-:3:1
-CATAATCAt
>Lbe:IbeG3:+:3:2
aTCTAAGTA-
>Lbe:IbeG4:+:3:3
gGTTAACCA-
>Lbe:IbeG5:+:2:4
tCCTAATCAc
>Lbe:IbeG6:+:2:5
cGTTAAATGa
>Lbe:IbeG7:-:3:6
-GGTAATTAc
>Lbe:IbeG8:+:3:7
cTATAAGTA-
>Lbe:IbeG9:-:3:8
-TGTAACAAg
>Lbe:IbeG10:-:3:9
-GTTAACCAg
>Lbe:IbeG11:+:3:10
gCCTAATTA-
>Lbe:IbeG12:+:4:11
gATTAACTA-
>Lbe:IbeH1:+:3:12
cACTAACAA-
>Lbe:IbeH2:-:3:13
-ACTAACGAg
>Lbe:IbeH3:+:3:14
aCATAATCA-
>Lbe:IbeH4:+:3:15
cTTTAACAA-
>Lbe:IbeH5:+:1:16
-GCTAATTAa
>Lbe:IbeH6:-:3:17
-GTTAATTGg
>Lbe:IbeH7:-:3:18
-GTTAATTAg
>Lbe:IbeH8:-:1:19
-GATAACAA-
>Lbe:IbeH9:-:3:20
-GTTAATTGg
>Lbe:IbeH10:-:3:21
-TTTAACGAg
>Lbe:IbeH11:-:3:22
-CTTAACGAg
>Lbl:LblC1:-:3:1
-GATAATTAt
>Lbl:LblC2:-:3:2
-TATAATTAc
>Lbl:LblC3:+:3:3
gGATAATTG-
>Lbl:LblC4:+:3:4
cACTAATCA-
>Lbl:LblC5:+:1:5
-GCTAATTAt
>Lbl:LblC6:+:3:6
tGCTAATTG-
>Lbl:LblC7:+:1:7
-GTTAATTAa
>Lbl:LblC8:+:3:8
aACTAACGA-
>Lbl:LblC9:-:3:9
-CTTAATCAg
>Lbl:LblC10:-:2:10
-GATAATTGg
>Lbl:LblC11:-:2:11
-GTTAACAAg
>Lbl:LblC12:-:3:12
-CTTAACGAg
>Lbl:LblD1:-:3:13
-TTTAATTGg
>Lbl:LblD2:-:3:14
-CTTAACGAg
>Lbl:LblD3:-:3:15
-CTTAATTGg
>Lbl:LblD4:-:3:16
-GATAATTGg
>Lbl:LblD5:+:3:17
gACTAATGA-
>Lbl:LblD6:+:3:18
tGCTAATCA-
>Lbl:LblD7:+:1:19
-GCTAATTAa
>Lbl:LblD8:+:1:20
-GCTAATTAa
>Lbl:LblD9:+:3:21
gACTAATGA-
>Lbl:LblD10:-:3:22
-CTTAACGAg
>Lbl:LblD11:+:3:23
gACTAATGA-
>Eve:Eve-G2:+:2:2
aATTAAGTAa
>Eve:Eve-G3:+:2:3
tAATAATCGa
>Eve:Eve-G4:-:3:4
tTCTAATCA-
>Eve:Eve-G5:-:3:5
aGTAAATTA-
>Eve:Eve-G6:-:3:6
tTCTAACGA-
>Eve:Eve-G7:+:3:7
-TATAATGAt
>Eve:Eve-G8:+:2:8
tACTAACGAc
>Eve:Eve-G9:-:3:9
gGCTAATTG-
>Eve:Eve-G10:+:2:10
gTCTAATTGa
>Eve:Eve-G11:+:3:11
-GTTAATGTg
>Eve:Eve-G12:+:3:12
-CATAATGAg
>Eve:Eve-H1:-:1:13
-GTTAATTAa
>Eve:Eve-H2:+:3:14
-GTTAATGGg
>Eve:Eve-H3:+:3:15
-TTTAATGAc
>Eve:Eve-H4:-:3:16
gTTTAATGA-
>Eve:Eve-H5:-:3:17
gGCTAATTA-
>Eve:Eve-H6:-:3:18
tGCTAATTA-
```

```
>Eve:Eve-H7:-:3:19
cACTAATTA-
>Eve:Eve-H8:+:2:20
tACTAATTAc
>Eve:Eve-H9:-:3:21
tGCTAATGA-
>Eve:Eve-H10:+:3:22
-ATTAATGAg
>Eve:Eve-H11:+:3:23
-GTTAATGAc
>E5:E5C1:-:3:1
-CCTAATTGa
>E5:E5C2:-:2:2
aTTTAATTAa
>E5:E5C3:-:2:3
tGTAAATTAg
>E5:E5C4:-:1:4
gCTTAATGG-
>E5:E5C5:+:1:5
-ATAAATTAa
>E5:E5C6:+:3:6
tGTTAATAA-
>E5:E5C7:+:3:7
cCGTAATTA-
>E5:E5C8:+:2:8
gCTTAAGTA-
>E5:E5C9:-:3:9
-GCTAATTGa
>E5:E5C10:+:3:10
cTGTAATGA-
>E5:E5C11:+:3:11
tTATAATTA-
>E5:E5C12:-:2:12
tTCTAATTAa
>E5:E5D2:+:3:14
gTATAATGA-
>E5:E5D3:-:1:15
cACTATTGA-
>E5:E5D4:+:3:16
tGTTAATTG-
>E5:E5D5:+:3:17
aGCTAATGA-
>E5:E5D7:+:3:19
gGCTAATTA-
>E5:E5D8:+:3:20
tATTAATTA-
>E5:E5D9:+:3:21
aTCTAATTA-
>E5:E5D10:+:3:22
cACTAATGA-
>E5:E5D11:+:3:23
gACTAATGA-
>E5:E52C1:+:3:24
gACTAATAG-
>E5:E52C2:-:2:25
gGTTAATGGg
>E5:E52C3:+:3:26
gGCTAAATA-
>E5:E52C4:-:2:27
aCGTAATGAa
>E5:E52C5:-:3:28
-CTTAATAGt
>E5:E52C6:+:3:29
aTTAAATGA-
>E5:E52C7:+:3:30
aCTTAATAA-
>E5:E52C8:+:3:31
cAATAATGA-
>E5:E52C10:-:3:33
-ACTAATGAa
>E5:E52C11:+:3:34
gGGTAATTA-
>E5:E52C12:+:3:35
gTTTAATGA-
>E5:E52D1:+:3:36
aTGTAATGA-
>E5:E52D2:+:3:37
tTTTAATTA-
>E5:E52D3:+:3:38
gCATAATTA-
>E5:E52D4:+:3:39
cGTTAATAG-
>E5:E52D5:+:3:40
aACTAATTA-
>E5:E52D6:-:2:41
tCTTAATTGa
>E5:E52D7:+:3:42
cATTAATGA-
>E5:E52D8:+:3:43
cAATAATTA-
>E5:E52D9:+:3:44
aTCTAATGA-
>E5:E52D10:+:3:45
cGGTAATTA-
>E5:E52D11:-:3:46
-GCTAATTAc
>BH1:B1HG8:+:1:8
-GCTAATTGA
>BH1:B1HG9:+:1:9
-GTTAATTGA
>BH1:B1HG10:+:2:10
tGTTAAACGG
>BH1:B1HH9:-:2:20
-CTTAATTGC
>BH1:BH12E2:+:2:22
tTCTAAACGG
>BH1:BH12E3:-:2:23
-GTTAATTGG
>BH1:BH12E4:+:2:24
gGCTAATTGA
>BH1:BH12E5:+:2:25
aGTTAATAGG
>BH1:BH12E6:-:2:26
-GTTAATTGT
>BH1:BH12E7:-:2:27
-GTTAATTGA
>BH1:BH12E8:-:2:28
-CATAATTGC
>BH1:BH12E9:+:2:29
tGTTAAACGG
>BH1:BH12E10:+:2:30
tCTTAAACGG
>BH1:BH12E11:+:2:31
tCTTAAACGG
>BH1:BH12E12:+:2:32
aGATAATTGC
>BH1:BH12F1:+:2:33
gGATAATTGA
>BH1:BH12F2:+:2:34
aCTTAAACGT
>BH1:BH12F3:+:1:35
-GATAATTAA
>BH1:BH12F4:+:2:36
cTTTAAACGG
>BH1:BH12F6:-:2:38
-ATTAAATGT
>BH1:BH12F11:+:2:42
tTCTAATTGA
>BH2:BH2E1:+:1:1
-GATAAACGG
>BH2:BH2E2:+:1:2
-TTTAATTGC
>BH2:BH2E4:+:1:3
-CTTAATAGT
>BH2:BH2E5:+:1:4
-ACTAAATGG
>BH2:BH2E6:+:1:5
-TTTAATAGG
>BH2:BH2E7:+:1:6
-CTTAATGGC
>BH2:BH2E9:+:1:8
-CTTAAAAGG
>BH2:BH2E10:+:1:9
-ATTAATTGG
>BH2:BH2E11:+:1:10
-GATAAATGA
>BH2:BH2E12:+:1:11
-TCTAATGGG
>BH2:BH2F1:+:1:12
-GATAATTGG
>BH2:BH2F2:+:1:13
-CTTAAATGA
>BH2:BH2F3:+:1:14
-TGTAATTGG
>BH2:BH2F4:+:1:15
-CATAATTGG
>BH2:BH2F5:+:1:16
-GATAATTGG
>BH2:BH2F6:+:1:17
-TATAATTGC
>BH2:BH2F7:+:1:18
-GTTAATTGA
>BH2:BH2F8:+:1:19
-GTTAATTGC
>BH2:BH2F9:+:1:20
-GTTAATTGA
>BH2:BH2F10:+:1:21
-CTTAATTGA
>BH2:BH2F11:+:1:22
-CTTAATTGG
>CG11085:CG11085G1:+
:1:1
-CTTAATGGG
>CG11085:CG11085G3:+
:1:3
-TTTAATTGG
>CG11085:CG11085G4:+
:1:4
-TTTAATTGC
>CG11085:CG11085G5:+
:1:5
-TTTAATAGG
>CG11085:CG11085G6:+
:1:6
-CTTAATGGG
>CG11085:CG11085G7:+
:1:7
-TTTAATTAC
>CG11085:CG11085G8:+
:1:8
-TCTAATAGC
>CG11085:CG11085G9:+
:1:9
-CTTAATCGG
>CG11085:CG11085G11:
+:1:11
-CCTAATTGC
>CG11085:CG11085H2:+
:1:14
-GCTAATTGA
>CG11085:CG11085H4:+
:1:16
-GTTAATTGG
>CG11085:CG11085H6:+
:1:18
-ATTAAATGT
>CG11085:CG11085H11:
+:1:23
-ATTAATTGA
>CG34031:CG34031E1:+
:3:1
-TGTAATTGt
>CG34031:CG34031E2:+
:2:2
cTTTAATTGc
>CG34031:CG34031E3:+
:3:3
-GTTAATTAg
>CG34031:CG34031E4:-
:3:4
tTTTTATTG-
>CG34031:CG34031E5:+
:3:5
-ATTAAATGt
>CG34031:CG34031E6:-
:3:6
tTTTAATAG-
>CG34031:CG34031E7:-
:3:7
cTTTTATAG-
>CG34031:CG34031E8:+
:2:8
cTTTAATAGt
>CG34031:CG34031F3:+
:3:15
-GTTAATTGc
>CG34031:CG34031F4:+
:3:16
-CTTAATGGt
>CG34031:CG34031A1:+
:2:23
tATTAATTAg
>CG34031:CG34031A2:+
:3:24
-CTTAATAGa
>CG34031:CG34031A3:+
:2:25
gTTTAATTGg
>CG34031:CG34031A4:+
:2:26
gTTTAATTGc
>CG34031:CG34031A5:+
:3:27
-TTTAATTGc
>CG34031:CG34031A6:+
:3:28
-TCTAATTGt
>CG34031:CG34031A7:+
:2:29
gTTTAATTGc
>CG34031:CG34031A8:+
:3:30
-TTTTATAGt
>CG34031:CG34031A10:
-:3:32
tGTTAATTG-
>CG34031:CG34031A11:
+:3:33
-CTTAATTGt
>CG34031:CG34031A12:
+:3:34
-TTTAATAGg
>CG34031:CG34031B2:+
:3:35
```

```
-CTTAATTGc
>CG34031:CG34031B3:+
:3:36
-GTTAATTGa
>CG34031:CG34031B5:+
:3:37
-TTTAAATAg
>CG34031:CG34031B6:+
:3:38
-TTTAATAGa
>Hmx:HMXA2:+:4:1
tgTTAATTG-
>Hmx:HMXA3:+:4:2
agATAATTG-
>Hmx:HMXA4:+:4:3
cgCTAATTG-
>Hmx:HMXA5:-:4:4
--TTAAATGg
>Hmx:HMXA6:+:4:5
taTTAATTG-
>Hmx:HMXA7:+:4:6
ccTTAATTG-
>Hmx:HMXA8:+:4:7
tgTTAATTG-
>Hmx:HMXA12:+:1:10
--TTAATTG-
>Hmx:HMXE1:+:4:11
cgTTAATTA-
>Hmx:HMXE2:+:4:12
acTTAATCG-
>Hmx:HMXE3:+:4:13
ctCTAATTG-
>Hmx:HMXE4:+:4:14
acTTAATCG-
>Hmx:HMXE5:+:4:15
taTTAATCG-
>Hmx:HMXE6:+:4:16
acTTAATTA-
>Hmx:HMXE7:+:4:17
agTTAATTA-
>Hmx:HMXE8:+:4:18
caCTAATTA-
>Hmx:HMXE9:+:4:19
gtTTAATTG-
>Hmx:HMXE10:+:4:20
atTTAATTG-
>Hmx:HMXF4:-:3:25
-cTTAATTGc
>Hmx:HMXF8:-:3:28
-cTTAATTGc
>Unc4:Unc4C1:-:1:1
-TCTAATTAg
>Unc4:Unc4C2:-:3:2
cTGTAATTA-
>Unc4:Unc4C3:+:3:3
-CTTAATTCg
>Unc4:Unc4C4:-:1:4
-GCTAATTAt
>Unc4:Unc4C5:-:1:5
-CTTAATTAt
>Unc4:Unc4C6:+:3:6
-CTTAATAGa
>Unc4:Unc4C7:+:2:7
tTTTAATTGa
>Unc4:Unc4C8:-:3:8
tCTTAATTG-
>Unc4:Unc4C9:+:3:9
-CCTAATTGa
>Unc4:Unc4C10:+:3:10
-TGTAATTGa
>Unc4:Unc4C11:+:3:11
-GTTAATTGc
>Unc4:Unc4C12:+:3:12
-CCTAATTGa
>Unc4:Unc4D1:-:3:13
cACTAATTA-
>Unc4:Unc4D3:+:3:15
-AATAATTGg
>Unc4:Unc4D4:+:3:16
-CTTAATTGa
>Unc4:Unc4D5:+:3:17
-GTTAATTGa
>Unc4:Unc4D6:+:3:18
-CTTAATTGg
>Unc4:Unc4D7:+:3:19
-TTTAATTGg
>Unc4:Unc4D8:+:3:20
-GTTAATTGg
>Unc4:Unc4D10:+:3:22
-GTTAATTGa
>Unc4:Unc4D11:+:3:23
-CTTAATTAg
>Odsh:OdshC1:-:4:1
ttGTAATTA-
>Odsh:OdshC2:+:2:2
-tTTAATTTc
>Odsh:OdshC3:+:2:3
tgCTAATTAt
>Odsh:OdshC4:-:3:4
taCTAATTAa
>Odsh:OdshC5:-:4:5
aaCTAATTG-
>Odsh:OdshC7:-:3:6
agCTAATTA-
>Odsh:OdshC8:-:2:7
-cCTAATTAc
>Odsh:OdshC9:+:2:8
ccTTAATTGc
>Odsh:OdshC10:+:3:9
-gGTAATTAc
>Odsh:OdshC11:-:4:10
atTTAATTA-
>Odsh:OdshC12:+:3:11
-cCTAATTGt
>Odsh:OdshD1:+:3:12
-aTTAATTGt
>Odsh:OdshD2:+:2:13
atTTAATTGg
>Odsh:OdshD3:-:3:14
tgCTAATTAa
>Odsh:OdshD4:-:4:15
acTTAATTA-
>Odsh:OdshD5:-:2:16
-cCTAATTAa
>Odsh:OdshD6:-:2:17
-tCTAATTAa
>Odsh:OdshD7:-:4:18
cgCTAATTA-
>Odsh:OdshD8:+:3:19
-aCTAATTGa
>Odsh:OdshD9:+:3:20
-tGTAATTGa
>Odsh:OdshD10:-:2:21
-cTTAATTAa
>Odsh:OdshD11:+:3:22
-gGTAATTGa
>Dr:DrA2:+:1:1
CCTCAATTA-
>Dr:DrA3:+:1:2
AAGCAATTA-
>Dr:DrA4:+:1:3
GGCCAATTA-
>Dr:DrA5:+:1:4
AACTAATTA-
>Dr:DrA6:+:1:5
CTCCAATTA-
>Dr:DrA7:+:1:6
CACCAATTA-
>Dr:DrA8:+:1:7
GAGCAATTA-
>Dr:DrA9:+:1:8
GGGTAATTA-
>Dr:DrA10:+:1:9
GACTAATTA-
>Dr:DrA11:+:1:10
CGCTAATTA-
>Dr:DrB2:+:1:11
CTCCAATTA-
>Dr:DrB3:+:1:12
GGCCAATTA-
>Dr:DrB4:+:1:13
AAACAATTA-
>Dr:DrB5:+:1:14
AACCAATTA-
>Dr:DrB6:+:1:15
GAGCAATTA-
>Dr:DrB7:+:1:16
GACCAATTA-
>Dr:DrB8:+:1:17
CTCCAATTA-
>Dr:DrB9:+:1:18
GAGTAATTA-
>Dr:DrB10:+:1:19
CAGCAATTA-
>Dr:DrB11:+:1:20
CCCCAATTA-
>Dr:DrB12:+:1:21
GTCCAATTA
```

Table A.4

Mutual Information product comparing full 60 amino acid homeodomains and when ignoring positions 11 through 39 that are less likely to contact DNA directly.

MI Joint Rank Product

MI Row-wise ranks

MI(b,a):

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	8	2	17	10	42	3	22	44	49	40	32	33	16	18	31
2	3	34	26	9	44	8	18	41	46	20	13	37	31	16	2
3	5	10	7	17	41	1	9	28	47	23	25	19	30	18	12
4	28	22	13	16	47	4	3	23	52	27	20	43	31	18	11
5	16	17	21	31	45	1	8	48	24	12	35	25	13	29	4
6	18	28	22	8	60	41	20	43	23	3	17	55	31	25	1
7	16	30	20	22	52	2	11	34	46	25	32	26	17	31	21
8	20	18	13	7	48	5	10	44	46	33	43	6	36	17	28
9	22	13	10	11	47	6	8	43	48	28	29	20	35	18	16
10	9	20	15	3	44	6	25	38	49	14	32	33	8	22	30

	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
1	57	48	25	6	43	41	23	20	28	54	50	26	11	4	24
2	53	32	19	21	40	28	11	27	25	52	47	43	23	4	7
3	55	31	29	2	33	38	11	49	26	53	50	40	27	13	34
4	55	41	24	12	42	10	19	32	33	54	45	37	17	14	29
5	56	22	23	14	30	36	32	44	37	55	43	41	9	28	46
6	53	40	2	11	47	27	9	33	21	54	4	10	15	13	36
7	57	40	35	5	9	18	28	33	29	54	50	44	14	8	37
8	55	35	19	1	16	22	39	38	34	54	49	45	9	8	24
9	57	31	12	2	32	19	40	33	17	53	46	41	21	9	39
10	54	50	34	4	43	36	23	13	29	55	52	27	21	1	17

	31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
1	27	1	29	45	55	19	21	47	13	51	34	15	36	37	56
2	42	15	24	48	54	17	14	36	30	49	38	1	33	50	51
3	32	24	14	46	56	8	15	43	35	44	45	6	39	51	52
4	46	25	1	38	56	8	9	39	15	40	48	5	35	49	50
5	11	26	20	51	58	39	38	33	42	52	49	7	34	50	53
6	50	6	5	35	44	24	7	37	19	52	58	56	30	34	39
7	27	41	15	51	56	23	12	38	39	48	47	4	36	49	53
8	11	25	30	50	56	23	14	32	41	52	42	12	29	47	53
9	15	37	27	50	56	30	5	45	38	49	44	7	36	51	54
10	37	2	16	48	57	12	7	47	28	45	35	26	31	39	56

	46	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
1	7	39	58	52	30	59	38	59	5	46	14	53	35	12	9
2	22	45	58	56	39	59	29	59	6	57	12	55	35	5	10
3	16	42	58	57	21	59	37	60	4	48	3	54	36	20	22
4	30	36	58	57	34	60	26	59	21	51	6	53	44	2	7
5	5	3	54	57	6	60	47	59	19	18	10	40	2	15	27
6	51	29	49	48	59	45	57	45	12	32	38	42	26	14	16
7	3	24	58	55	6	60	10	59	1	43	7	45	42	13	19
8	3	26	57	58	2	59	37	60	4	40	27	51	21	15	31
9	3	24	58	55	1	59	42	60	4	34	14	52	23	26	25
10	11	40	58	53	24	59	41	59	10	46	19	51	42	5	18

MI Column-wise ranks

MI(b,a):

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	6	5	8	6	6	8	8	8	7	9	8	7	6	6	9
2	8	9	9	9	9	9	9	9	8	8	9	9	9	9	8
3	7	7	7	8	5	6	7	7	5	6	7	6	8	8	7
4	9	8	6	7	8	7	5	5	9	7	6	8	7	7	5
5	4	4	4	4	2	3	4	4	1	3	4	4	2	4	2
6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
7	2	3	3	3	7	1	3	1	3	2	2	3	1	3	3
8	3	2	2	2	3	4	2	3	2	4	3	1	4	2	4
9	1	1	1	1	1	2	1	2	4	1	1	2	3	1	1
10	5	6	5	5	4	5	6	6	6	5	5	5	5	5	6

	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
1	9	9	6	7	8	8	8	6	7	8	6	6	6	6	6
2	6	7	9	9	9	9	9	8	9	6	9	9	9	9	8
3	8	5	8	6	5	7	6	9	6	7	7	8	8	8	9
4	7	6	7	8	7	5	7	7	8	9	5	7	7	7	7
5	1	3	4	4	4	4	4	5	4	4	2	3	4	4	5
6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
7	5	4	3	3	1	3	1	2	3	3	4	4	2	3	3
8	2	2	2	1	2	2	3	3	2	2	3	2	1	2	1
9	4	1	1	2	3	1	2	1	1	1	1	1	3	1	2
10	3	8	5	5	6	6	5	4	5	5	8	5	5	5	4

	31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
1	6	6	8	8	5	8	7	9	5	8	6	8	6	6	8
2	9	9	9	9	6	9	9	8	9	9	8	9	9	9	7
3	7	8	7	6	8	7	8	6	8	6	7	6	8	8	5
4	8	7	5	5	7	6	6	5	7	4	9	5	7	7	6
5	4	3	4	4	3	4	4	4	4	7	5	4	4	4	3
6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
7	3	4	1	3	2	2	2	2	3	2	3	2	3	2	4
8	1	1	3	2	4	1	3	1	2	3	2	3	1	1	1
9	2	2	2	1	1	3	1	3	1	1	1	1	2	3	2
10	5	5	6	7	9	5	5	7	6	5	4	7	5	5	9

	46	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
1	6	7	9	4	7	10	8	10	6	7	8	6	6	7	7
2	9	9	10	7	9	8	9	8	9	9	9	9	9	9	9
3	7	8	7	10	6	6	7	6	7	6	7	8	7	8	8
4	8	6	6	6	8	3	5	3	8	8	5	7	8	6	5
5	4	1	1	5	4	2	4	2	4	2	4	1	1	4	4
6	10	10	5	9	10	1	10	1	10	10	10	10	10	10	10
7	3	4	4	2	3	4	1	4	1	4	1	2	4	1	2
8	1	3	3	8	2	5	2	5	3	3	3	3	3	2	3
9	2	2	2	1	1	7	3	7	2	1	2	4	2	3	1
10	5	5	8	3	5	9	6	9	5	5	6	5	5	5	6

Joint Rank Product = (row rank)*(col rank)

Color Key: products equal to 1 (i.e. top element in row as well as column) are red. Products <=4 are purple. Products <= 9 are blue

I choose 1, 4, and 9 as cut offs because rowRank_1*colRank_1 = 1, rowRank_2*colRank2 = 4 etc.

1	4	9

Joint Rank Product = (row rank)*(col rank)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	48	10	136	60	252	24	176	352	343	360	256	231	96	108	279
2	24	306	234	81	396	72	162	369	368	160	117	333	279	144	16
3	35	70	49	136	205	6	63	196	235	138	175	114	240	144	84
4	252	176	78	112	376	28	15	115	468	189	120	344	217	126	55
5	64	68	84	124	90	3	32	192	24	36	140	100	26	116	8
6	180	280	220	80	600	410	200	430	230	30	170	550	310	250	10
7	32	90	60	66	364	2	33	34	138	50	64	78	17	93	63
8	60	36	26	14	144	20	20	132	92	132	129	6	144	34	112
9	22	13	10	11	47	12	8	86	192	28	29	40	105	18	16
10	45	120	75	15	176	30	150	228	294	70	160	165	40	110	180

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	513	432	150	42	344	328	184	120	196	432	300	156	66	24	144
2	318	224	171	189	360	252	99	216	225	312	423	387	207	36	56
3	440	155	232	12	165	266	66	441	156	371	350	320	216	104	306
4	385	246	168	96	294	50	133	224	264	486	225	259	119	98	203
5	56	66	92	56	120	144	128	220	148	220	86	123	36	112	230
6	530	400	20	110	470	270	90	330	210	540	40	100	150	130	360
7	285	160	105	15	9	54	28	66	87	162	200	176	28	24	111
8	110	70	38	1	32	44	117	114	68	108	147	90	9	16	24
9	228	31	12	4	96	19	80	33	17	53	46	41	63	9	78
10	162	400	170	20	258	216	115	52	145	275	416	135	105	5	68

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
1	162	6	232	360	275	152	147	423	65	408	204	120	216	222	448
2	378	135	216	432	324	153	126	288	270	441	304	9	297	450	357
3	224	192	98	276	448	56	120	258	280	264	315	36	312	408	260
4	368	175	5	190	392	48	54	195	105	160	432	25	245	343	300
5	44	78	80	204	174	156	152	132	168	364	245	28	136	200	159
6	500	60	50	350	440	240	70	370	190	520	580	560	300	340	390
7	81	164	15	153	112	46	24	76	117	96	141	8	108	98	212
8	11	25	90	100	224	23	42	32	82	156	84	36	29	47	53
9	30	74	54	50	56	90	5	135	38	49	44	7	72	153	108
10	185	10	96	336	513	60	35	329	168	225	140	182	155	195	504

	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
1	42	273	522	208	210	590	304	590	30	322	112	318	210	84	63
2	198	405	580	392	351	472	261	472	54	513	108	495	315	45	90
3	112	336	406	570	126	354	259	360	28	288	21	432	252	160	176
4	240	216	348	342	272	180	130	177	168	408	30	371	352	12	35
5	20	3	54	285	24	120	188	118	76	36	40	40	2	60	108
6	510	290	245	432	590	45	570	45	120	320	380	420	260	140	160
7	9	96	232	110	18	240	10	236	1	172	7	90	168	13	38
8	3	78	171	464	4	295	74	300	12	120	81	153	63	30	93
9	6	48	116	55	1	413	126	420	8	34	28	208	46	78	25
10	55	200	464	159	120	531	246	531	50	230	114	255	210	25	108

Joint Rank Product Summary

	1	32						
	2	42						
Expected based on qualitative code	3	6						
5,6,7	4	33						
55,2,3	5	6	58	47	15			
51	6							
47,54,43	7	**54***	6	56	42	20	46	
47,50,54	8	19	46	**50***	12	28		
47,50,54	9	**50***	19	37	46	42	7	29
	10	29						

Ranking the joint rank product by row.

Purpose: find row rank of 'missed' qualitative code AA residues (shown in red)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	8	2	19	9	37	3	26	49	47	50	38	35	14	15	41
2	3	36	28	9	50	8	19	46	45	18	13	41	32	16	2
3	5	11	7	19	30	1	9	29	34	20	26	16	35	21	12
4	41	27	12	16	54	5	3	17	59	30	19	49	35	20	11
5	21	23	27	39	29	2	10	52	6	11	43	31	8	34	4
6	20	31	24	10	60	44	22	46	25	3	19	55	34	28	1
7	17	33	24	27	60	2	18	19	45	22	26	30	11	35	25
8	28	21	15	8	51	10	10	49	39	49	48	4	51	20	44
9	18	13	9	10	33	11	6	47	56	20	22	28	50	16	14
10	9	23	16	3	36	6	28	45	51	15	31	33	8	20	37

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	57	54	22	6	48	46	27	17	28	54	42	24	12	3	20
2	39	26	20	21	44	29	11	24	27	37	52	48	23	4	7
3	57	22	33	2	25	41	10	58	23	53	50	48	31	14	45
4	55	40	24	13	45	9	22	36	43	60	37	42	18	14	33
5	18	22	30	18	36	44	40	55	45	55	28	38	11	33	57
6	53	43	2	13	49	30	11	36	23	54	4	12	17	15	39
7	59	48	39	9	5	23	15	27	32	49	54	53	15	13	42
8	43	31	23	1	18	25	46	45	30	42	53	37	5	9	13
9	58	24	11	2	49	17	46	25	15	37	31	29	41	8	44
10	32	54	35	4	49	43	22	11	27	50	55	25	18	1	14

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
1	25	1	36	50	40	23	21	53	11	52	29	17	33	34	56
2	47	15	24	53	40	17	14	33	31	54	35	1	34	55	43
3	32	28	13	42	59	8	17	37	43	40	47	6	46	55	39
4	52	26	1	31	56	8	10	32	15	23	58	4	39	48	46
5	16	25	26	54	50	47	46	41	49	60	58	9	42	53	48
6	50	8	7	38	48	26	9	40	21	52	58	56	33	37	42
7	31	50	9	47	43	21	13	29	44	36	46	4	40	38	55
8	6	14	37	41	57	12	24	18	35	55	36	21	16	26	27
9	23	43	38	36	40	48	3	54	27	35	30	5	42	55	51
10	39	2	17	53	58	13	7	52	34	44	26	38	29	40	57

	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
1	6	39	58	30	31	59	43	59	5	45	16	44	31	13	10
2	22	51	60	49	42	56	30	56	6	59	12	58	38	5	10
3	15	49	54	60	18	51	38	52	4	44	3	56	36	24	27
4	38	34	50	47	44	29	21	28	24	57	6	53	51	2	7
5	5	2	17	59	6	36	51	35	24	11	14	14	1	20	32
6	51	32	27	47	59	5	57	5	14	35	41	45	29	16	18
7	5	36	56	41	12	58	7	57	1	52	3	33	51	8	20
8	2	33	56	60	3	58	32	59	7	47	34	54	29	17	40
9	4	34	52	39	1	59	53	60	6	26	20	57	31	44	19
10	12	41	56	30	23	59	47	59	10	46	21	48	42	5	19

Ignoring positions 11 through 39

MI(b,a):

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.12	0.15	0.11	0.12	0.05	0.14	0.10	0.04	0.03	0.06	0	0	0	0	0
2	0.08	0.04	0.04	0.07	0.02	0.07	0.06	0.03	0.02	0.06	0	0	0	0	0
3	0.12	0.11	0.11	0.09	0.05	0.15	0.11	0.07	0.04	0.09	0	0	0	0	0
4	0.08	0.09	0.12	0.11	0.03	0.14	0.15	0.09	0.02	0.08	0	0	0	0	0
5	0.25	0.25	0.24	0.20	0.14	0.45	0.30	0.09	0.23	0.28	0	0	0	0	0
6	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0	0	0	0	0
7	0.34	0.25	0.32	0.31	0.04	0.47	0.37	0.24	0.14	0.30	0	0	0	0	0
8	0.34	0.34	0.36	0.41	0.13	0.44	0.39	0.21	0.17	0.26	0	0	0	0	0
9	0.35	0.39	0.45	0.43	0.16	0.46	0.46	0.23	0.14	0.32	0	0	0	0	0
10	0.16	0.13	0.14	0.19	0.06	0.17	0.12	0.09	0.04	0.14	0	0	0	0	0

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
1	0	0	0	0	0	0	0	0	0	0.02	0.07	0.11	0.07	0.06	0.01
2	0	0	0	0	0	0	0	0	0	0.02	0.03	0.08	0.04	0.02	0.02
3	0	0	0	0	0	0	0	0	0	0.05	0.04	0.12	0.06	0.03	0.03
4	0	0	0	0	0	0	0	0	0	0.05	0.03	0.14	0.06	0.03	0.02
5	0	0	0	0	0	0	0	0	0	0.04	0.09	0.31	0.19	0.08	0.04
6	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
7	0	0	0	0	0	0	0	0	0	0.13	0.13	0.44	0.24	0.12	0.03
8	0	0	0	0	0	0	0	0	0	0.09	0.22	0.39	0.30	0.17	0.05
9	0	0	0	0	0	0	0	0	0	0.13	0.23	0.46	0.26	0.10	0.04
10	0	0	0	0	0	0	0	0	0	0.05	0.10	0.11	0.10	0.08	0.01

	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
1	0.12	0.06	0.00	0.02	0.08	0.00	0.06	0.00	0.13	0.03	0.11	0.02	0.07	0.12	0.12
2	0.05	0.02	0.00	0.01	0.03	0.00	0.04	0.00	0.08	0.01	0.07	0.01	0.03	0.08	0.07
3	0.10	0.05	0.00	0.00	0.09	0.00	0.06	0.00	0.12	0.04	0.12	0.01	0.06	0.09	0.09
4	0.07	0.06	0.00	0.01	0.07	0.00	0.08	0.00	0.10	0.02	0.14	0.02	0.04	0.16	0.14
5	0.32	0.35	0.03	0.01	0.31	0.00	0.11	0.00	0.24	0.25	0.30	0.16	0.36	0.26	0.22
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.01
7	0.44	0.30	0.00	0.02	0.43	0.00	0.37	0.00	0.50	0.17	0.41	0.16	0.19	0.36	0.32
8	0.56	0.31	0.01	0.01	0.58	0.00	0.24	0.00	0.45	0.24	0.30	0.11	0.33	0.36	0.28
9	0.53	0.34	0.01	0.04	0.69	0.00	0.23	0.00	0.47	0.27	0.39	0.08	0.34	0.33	0.33
10	0.15	0.08	0.00	0.02	0.12	0.00	0.08	0.00	0.16	0.05	0.13	0.03	0.08	0.17	0.13

Row-wise ranks

MI(b,a):

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	5	1	11	7	21	2	12	22	24	20	32	32	32	32	32
2	2	15	12	6	20	5	9	19	22	10	32	32	32	32	32
3	4	8	6	10	19	1	7	15	23	14	32	32	32	32	32
4	14	10	7	8	21	3	2	11	26	13	32	32	32	32	32
5	11	12	15	18	21	1	7	23	16	9	32	32	32	32	32
6	6	11	8	2	31	18	7	20	9	1	32	32	32	32	32
7	10	16	12	13	26	2	8	17	22	15	32	32	32	32	32
8	11	10	8	5	24	4	6	21	22	17	32	32	32	32	32
9	11	9	7	8	22	4	6	20	23	16	32	32	32	32	32
10	4	11	8	1	22	3	13	17	25	7	32	32	32	32	32

	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
1	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
2	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
3	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
4	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
5	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
6	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
7	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
8	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
9	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
10	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32

	31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
1	32	32	32	32	32	32	32	32	32	25	14	10	16	17	28
2	32	32	32	32	32	32	32	32	32	23	17	1	14	24	25
3	32	32	32	32	32	32	32	32	32	21	22	5	18	25	26
4	32	32	32	32	32	32	32	32	32	19	22	4	17	23	24
5	32	32	32	32	32	32	32	32	32	26	24	6	19	25	27
6	32	32	32	32	32	32	32	32	32	26	29	27	13	15	17
7	32	32	32	32	32	32	32	32	32	24	23	4	18	25	27
8	32	32	32	32	32	32	32	32	32	26	20	7	15	23	27
9	32	32	32	32	32	32	32	32	32	24	21	5	18	25	27
10	32	32	32	32	32	32	32	32	32	23	16	14	15	18	28

	46	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
1	4	19	29	26	13	30	18	30	3	23	9	27	15	8	6
2	11	21	29	27	18	30	13	30	4	28	8	26	16	3	7
3	9	20	29	28	12	30	17	31	3	24	2	27	16	11	13
4	15	18	29	28	16	31	12	30	9	25	5	27	20	1	6
5	4	3	28	29	5	31	22	30	14	13	8	20	2	10	17
6	25	12	24	23	30	21	28	21	3	14	16	19	10	4	5
7	3	14	29	28	5	31	7	30	1	20	6	21	19	9	11
8	2	13	28	29	1	30	18	31	3	19	14	25	12	9	16
9	2	13	29	28	1	30	19	31	3	17	10	26	12	15	14
10	6	19	29	27	12	30	20	30	5	24	10	26	21	2	9

Column-wise ranks

MI(b,a):

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	6	5	8	6	6	8	8	8	7	9	10	10	10	10	10
2	8	9	9	9	9	9	9	9	8	8	10	10	10	10	10
3	7	7	7	8	5	6	7	7	5	6	10	10	10	10	10
4	9	8	6	7	8	7	5	5	9	7	10	10	10	10	10
5	4	4	4	4	2	3	4	4	1	3	10	10	10	10	10
6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
7	2	3	3	3	7	1	3	1	3	2	10	10	10	10	10
8	3	2	2	2	3	4	2	3	2	4	10	10	10	10	10
9	1	1	1	1	1	2	1	2	4	1	10	10	10	10	10
10	5	6	5	5	4	5	6	6	6	5	10	10	10	10	10

	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
1	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
2	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
3	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
4	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
5	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
7	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
8	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
9	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

	31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
1	10	10	10	10	10	10	10	10	10	8	6	8	6	6	8
2	10	10	10	10	10	10	10	10	10	9	8	9	9	9	7
3	10	10	10	10	10	10	10	10	10	6	7	6	8	8	5
4	10	10	10	10	10	10	10	10	10	4	9	5	7	7	6
5	10	10	10	10	10	10	10	10	10	7	5	4	4	4	3
6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
7	10	10	10	10	10	10	10	10	10	2	3	2	3	2	4
8	10	10	10	10	10	10	10	10	10	3	2	3	1	1	1
9	10	10	10	10	10	10	10	10	10	1	1	1	2	3	2
10	10	10	10	10	10	10	10	10	10	5	4	7	5	5	9

	46	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
1	6	7	9	4	7	10	8	10	6	7	8	6	6	7	7
2	9	9	10	7	9	8	9	8	9	9	9	9	9	9	9
3	7	8	7	10	6	6	7	6	7	6	7	8	7	8	8
4	8	6	6	6	8	3	5	3	8	8	5	7	8	6	5
5	4	1	1	5	4	2	4	2	4	2	4	1	1	4	4
6	10	10	5	9	10	1	10	1	10	10	10	10	10	10	10
7	3	4	4	2	3	4	1	4	1	4	1	2	4	1	2
8	1	3	3	8	2	5	2	5	3	3	3	3	3	2	3
9	2	2	2	1	1	7	3	7	2	1	2	4	2	3	1
10	5	5	8	3	5	9	6	9	5	5	6	5	5	5	6

Ignoring positions 11 through 39:
Joint Rank Product = (row rank)*(col rank)

Joint Rank Product = (row rank)*(col rank)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	30	5	88	42	126	16	96	176	168	180	320	320	320	320	320
2	16	135	108	54	180	45	81	171	176	80	320	320	320	320	320
3	28	56	42	80	95	6	49	105	115	84	320	320	320	320	320
4	126	80	42	56	168	21	10	55	234	91	320	320	320	320	320
5	44	48	60	72	42	3	28	92	16	27	320	320	320	320	320
6	60	110	80	20	310	180	70	200	90	10	320	320	320	320	320
7	20	48	36	39	182	2	24	17	66	30	320	320	320	320	320
8	33	20	16	10	72	16	12	63	44	68	320	320	320	320	320
9	11	9	7	8	22	8	6	40	92	16	320	320	320	320	320
10	20	66	40	5	88	15	78	102	150	35	320	320	320	320	320

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320
2	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320
3	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320
4	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320
5	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320
6	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320
7	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320
8	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320
9	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320
10	320	320	320	320	320	320	320	320	320	320	320	320	320	320	320

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
1	320	320	320	320	320	320	320	320	320	200	84	80	96	102	224
2	320	320	320	320	320	320	320	320	320	207	136	9	126	216	175
3	320	320	320	320	320	320	320	320	320	126	154	30	144	200	130
4	320	320	320	320	320	320	320	320	320	76	198	20	119	161	144
5	320	320	320	320	320	320	320	320	320	182	120	24	76	100	81
6	320	320	320	320	320	320	320	320	320	260	290	270	130	150	170
7	320	320	320	320	320	320	320	320	320	48	69	8	54	50	108
8	320	320	320	320	320	320	320	320	320	78	40	21	15	23	27
9	320	320	320	320	320	320	320	320	320	24	21	5	36	75	54
10	320	320	320	320	320	320	320	320	320	115	64	98	75	90	252

	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
1	24	133	261	104	91	300	144	300	18	161	72	162	90	56	42
2	99	189	290	189	162	240	117	240	36	252	72	234	144	27	63
3	63	160	203	280	72	180	119	186	21	144	14	216	112	88	104
4	120	108	174	168	128	93	60	90	72	200	25	189	160	6	30
5	16	3	28	145	20	62	88	60	56	26	32	20	2	40	68
6	250	120	120	207	300	21	280	21	30	140	160	190	100	40	50
7	9	56	116	56	15	124	7	120	1	80	6	42	76	9	22
8	2	39	84	232	2	150	36	155	9	57	42	75	36	18	48
9	4	26	58	28	1	210	57	217	6	17	20	104	24	45	14
10	30	95	232	81	60	270	120	270	25	120	60	130	105	10	54

Summary ignoring positions 11 through 39:

Joint Rank Product Summary

Expected based on qualitative code	Nuc	AA's								
	1	2								
	2	42								
	3	6								
5,6,7	**4**	59								
55,2,3	**5**	6	58	47						
51	**6**									
47,54,43	**7**	**54***	6	56	52	42	46	59		
47,50,54	**8**	46	**50***	54*						
47,50,54	**9**	**50***	46	42	7	54*	3	4	6	9
	10	4								

Table A.5
Fly factors used to predict the human homeodomains and confidence scores.

Confidence Score	Human HD Query	fly HD Ref	Similarity Score	#Key Residue matches	Ref. Key Residues (5,47,50,54 &55)	#Sites Contributed by Ref
1	ALX3				RVQAK	
		Pph13	316	5	RVQAK	21
		Hbn	311	5	RVQAK	17
		Al	310	5	RVQAK	20
1	ALX4				RVQAK	
		Rx	315	5	RVQAK	27
		Al	314	5	RVQAK	20
		Pph13	311	5	RVQAK	21
	ARGFX				RVRFK	
	No predictions made					
1	ARX				RVQAK	
		Al	349	5	RVQAK	20
		Pph13	348	5	RVQAK	21
		Hbn	333	5	RVQAK	17
1	BARHL1				RTQTK	
		BH2	320	5	RTQTK	21
		BH1	310	5	RTQTK	21
1	BARHL2				RTQTK	
		BH2	314	5	RTQTK	21
		BH1	304	5	RTQTK	21
3	BARX1				RTQMK	
		Bsh	254	5	RTQMK	16
3	BARX2				RTQMK	
		Bsh	242	5	RTQMK	16
2	BSX				RTQMK	
		Bsh	339	5	RTQMK	16
2	CDX1				RIQAK	
		Cad	303	5	RIQAK	38
2	CDX2				RIQAK	
		Cad	308	5	RIQAK	38

2	CDX4				RIQAK	
		Cad	284	5	RIQAK	38
2	CRX				RVKAK	
		Oc	327	5	RVKAK	19
	CUTL1				RNHSR	
	No predictions made					
4	CUTL2				RNHSR	
		Ct	203	4	RNHMR	20
2	DBX1				RIQMK	
		CG12361	334	5	RIQMK	16
2	DBX2				RIQMK	
		CG12361	306	5	RIQMK	16
2	DLX1				RIQSK	
		Dll	332	5	RIQSK	23
2	DLX2				RIQSK	
		Dll	316	5	RIQSK	23
2	DLX3				RIQSK	
		Dll	304	5	RIQSK	23
2	DLX4				RIQSK	
		Dll	320	5	RIQSK	23
2	DLX5				RIQSK	
		Dll	313	5	RIQSK	23
2	DLX6				RIQSK	
		Dll	321	5	RIQSK	23
3	DMBX1				RVKAK	
		Gsc	260	5	RVKAK	22
		Ptx1	252	5	RVKAK	20
		Oc	243	5	RVKAK	19
1	DRGX				RVQAK	
		Al	312	5	RVQAK	20
		CG11294	308	5	RVQAK	15
		Pph13	295	5	RVQAK	21

1	EMX1				RVQTK	
		E5	311	5	RVQTK	43
		Ems	300	5	RVQTK	20
1	EMX2				RVQTK	
		E5	323	5	RVQTK	43
		Ems	307	5	RVQTK	20
3	EN1				RIQAK	
		En	283	5	RIQAK	23
		Inv	272	5	RIQAK	16
3	EN2				RIQAK	
		En	287	5	RIQAK	23
		Inv	276	5	RIQAK	16
3	ESX1				RVQAK	
		Hbn	265	5	RVQAK	17
		Repo	264	5	RVQAK	28
		Rx	260	5	RVQAK	27
2	EVX1				RVQMK	
		Eve	354	5	RVQMK	22
2	EVX2				RVQMK	
		Eve	358	5	RVQMK	22
2	GBX1				RIQAK	
		Unpg	338	5	RIQAK	21
2	GBX2				RIQAK	
		Unpg	337	5	RIQAK	21
2	GSC				RVKAK	
		Gsc	300	5	RVKAK	22
2	GSCL				RVKAK	
		Gsc	282	5	RVKAK	22
2	GSX1				RIQVK	
		Ind	300	5	RIQVK	21
2	GSX2				RIQVK	
		Ind	299	5	RIQVK	21
	HESX1				RIQAK	
	No predictions made					
3	HHEX				QTQAK	

		CG7056	267	5	QTQAK	26
2	HLX				RVQMK	
		H2	308	5	RVQMK	32
	HMBOX1				RNAKE	
	No predictions made					
2	HMX1				RIQNK	
		Hmx	346	5	RIQNK	20
3	HMX2				RTQNK	
		Hmx	321	4	RIQNK	20
2	HMX3				RIQNK	
		Hmx	345	5	RIQNK	20
	HNF1A				RNAKE	
	No predictions made					
	HNF1B				RNAKE	
	No predictions made					
2	HOXA1				RIQMK	
		Lab	315	5	RIQMK	16
3	HOXA10				RIQMK	
		AbdB	259	5	RIQMK	21
		Antp	248	5	RIQMK	16
		Scr	248	5	RIQMK	25
		AbdA	247	5	RIQMK	18
4	HOXA11				RIQMK	
		AbdB	251	5	RIQMK	21
		Scr	218	5	RIQMK	25
		Ftz	213	5	RIQMK	18
	HOXA13				RIQVK	
	No predictions made					
2	HOXA2				RVQMK	
		Pb	365	5	RVQMK	24
3	HOXA3				RIQMK	
		Scr	279	5	RIQMK	25
		Dfd	278	5	RIQMK	24
		Ftz	273	5	RIQMK	18
1	HOXA4				RIQMK	

		Dfd	355	5	RIQMK	24
		Scr	351	5	RIQMK	25
		Antp	334	5	RIQMK	16
1	HOXA5				RIQMK	
		Scr	361	5	RIQMK	25
		Antp	360	5	RIQMK	16
		Dfd	342	5	RIQMK	24
1	HOXA6				RIQMK	
		Antp	375	5	RIQMK	16
		Scr	351	5	RIQMK	25
		AbdA	341	5	RIQMK	18
1	HOXA7				RIQMK	
		Antp	390	5	RIQMK	16
		Scr	372	5	RIQMK	25
		AbdA	356	5	RIQMK	18
3	HOXA9				RIQMK	
		AbdB	272	5	RIQMK	21
		Scr	260	5	RIQMK	25
		Antp	254	5	RIQMK	16
2	HOXB1				RIQMK	
		Lab	303	5	RIQMK	16
	HOXB13				RIQVK	
	No predictions made					
2	HOXB2				RVQMK	
		Pb	365	5	RVQMK	24
3	HOXB3				RIQMK	
		Dfd	280	5	RIQMK	24
		Scr	276	5	RIQMK	25
		Ftz	275	5	RIQMK	18
1	HOXB4				RIQMK	
		Dfd	355	5	RIQMK	24
		Scr	351	5	RIQMK	25
		Antp	334	5	RIQMK	16
1	HOXB5				RIQMK	
		Scr	361	5	RIQMK	25
		Antp	360	5	RIQMK	16
		Dfd	342	5	RIQMK	24
1	HOXB6				RIQMK	

		Antp	374	5	RIQMK	16
		Scr	353	5	RIQMK	25
		AbdA	345	5	RIQMK	18
1	HOXB7				RIQMK	
		Antp	385	5	RIQMK	16
		Scr	353	5	RIQMK	25
		AbdA	345	5	RIQMK	18
1	HOXB8				RIQMK	
		Antp	342	5	RIQMK	16
		AbdA	315	5	RIQMK	18
		Ubx	310	5	RIQMK	20
3	HOXB9				RIQMK	
		AbdB	267	5	RIQMK	21
		Scr	258	5	RIQMK	25
		Antp	255	5	RIQMK	16
3	HOXC10				RIQMK	
		AbdB	254	5	RIQMK	21
		Dfd	250	5	RIQMK	24
		Antp	248	5	RIQMK	16
3	HOXC11				RIQMK	
		AbdB	265	5	RIQMK	21
4	HOXC12				RIQMK	
		AbdB	232	5	RIQMK	21
		Zen2	207	5	RIQMK	26
		Ftz	202	5	RIQMK	18
	HOXC13				RIQVK	
	No predictions made					
1	HOXC4				RIQMK	
		Dfd	359	5	RIQMK	24
		Scr	349	5	RIQMK	25
		Antp	332	5	RIQMK	16
1	HOXC5				RIQMK	
		Scr	346	5	RIQMK	25
		Antp	340	5	RIQMK	16
		Dfd	322	5	RIQMK	24
1	HOXC6				RIQMK	
		Antp	370	5	RIQMK	16
		Scr	343	5	RIQMK	25
		AbdA	341	5	RIQMK	18

1	HOXC8				RIQMK	
		Antp	334	5	RIQMK	16
		AbdA	307	5	RIQMK	18
		Ubx	302	5	RIQMK	20
3	HOXC9				RIQMK	
		AbdB	277	5	RIQMK	21
		Scr	268	5	RIQMK	25
		Antp	262	5	RIQMK	16
2	HOXD1				RIQMK	
		Lab	306	5	RIQMK	16
3	HOXD10				RIQMK	
		AbdB	257	5	RIQMK	21
		AbdA	247	5	RIQMK	18
		Scr	246	5	RIQMK	25
4	HOXD11				RIQMK	
		AbdB	254	5	RIQMK	21
		Ftz	221	5	RIQMK	18
		Scr	221	5	RIQMK	25
3	HOXD12				RIQMK	
		AbdB	229	5	RIQMK	21
	HOXD13				RIQVK	
	No predictions made					
3	HOXD3				RIQMK	
		Scr	277	5	RIQMK	25
		Dfd	276	5	RIQMK	24
		Ftz	273	5	RIQMK	18
1	HOXD4				RIQMK	
		Dfd	357	5	RIQMK	24
		Scr	353	5	RIQMK	25
		Antp	336	5	RIQMK	16
1	HOXD8				RIQMK	
		Antp	339	5	RIQMK	16
		AbdA	322	5	RIQMK	18
		Ubx	307	5	RIQMK	20
3	HOXD9				RIQMK	
		AbdB	281	5	RIQMK	21
		Scr	265	5	RIQMK	25
		AbdA	256	5	RIQMK	18

3	IRX1				KTARR	
		Mirr	309	4	RTARR	41
		Ara	300	4	RTARR	34
		Caup	300	4	RTARR	19
1	IRX2				RTARR	
		Ara	330	5	RTARR	34
		Caup	330	5	RTARR	19
		Mirr	329	5	RTARR	41
3	IRX3				KTARR	
		Mirr	302	4	RTARR	41
		Ara	301	4	RTARR	34
		Caup	301	4	RTARR	19
1	IRX4				RTARR	
		Mirr	352	5	RTARR	41
		Ara	335	5	RTARR	34
		Caup	335	5	RTARR	19
1	IRX5				RTARR	
		Ara	330	5	RTARR	34
		Caup	330	5	RTARR	19
		Mirr	329	5	RTARR	41
1	IRX6				RTARR	
		Mirr	344	5	RTARR	41
		Ara	340	5	RTARR	34
		Caup	340	5	RTARR	19
2	ISL1				RVQCK	
		Tup	353	5	RVQCK	16
2	ISL2				RVQCK	
		Tup	355	5	RVQCK	16
4	ISX				RIQAK	
		Unpg	204	5	RIQAK	21
1	LBX1				RTQAK	
		Lbe	332	5	RTQAK	22
		Lbl	327	5	RTQAK	23
3	LBX2				RTQAK	
		Lbl	287	5	RTQAK	23
		Lbe	280	5	RTQAK	22
2	LHX1				RVQSK	

		Lim1	330	5	RVQSK	18
2	LHX2				RVQAK	
		Ap	345	5	RVQAK	19
2	LHX3				RVQAK	
		Lim3	339	5	RVQAK	20
2	LHX4				RVQAK	
		Lim3	337	5	RVQAK	20
2	LHX5				RVQSK	
		Lim1	330	5	RVQSK	18
3	LHX8				RVQAR	
		Awh	253	5	RVQAR	40
2	LHX9				RVQAK	
		Ap	345	5	RVQAK	19
1	LMX1A				RVQAK	
		CG4328	317	5	RVQAK	30
		CG32105	308	5	RVQAK	19
1	LMX1B				RVQAK	
		CG4328	317	5	RVQAK	30
		CG32105	308	5	RVQAK	19
2	MEIS1				RNIRR	
		Hth	366	5	RNIRR	17
2	MEIS2				RNIRR	
		Hth	371	5	RNIRR	17
	MEIS3				DNIRR	
	No predictions made					
2	MEIS3P2				RNIRR	
		Hth	335	5	RNIRR	17
2	MEOX1				RVQMK	
		Btn	295	5	RVQMK	23
3	MIXL1				RVQAK	
		CG11294	234	5	RVQAK	15

		Pph13	231	5	RVQAK	21
		CG32 532	229	5	RVQAK	23
	MKX				RNARR	
	No predictions made					
2	MNX1				RIQMK	
		Exex	339	5	RIQMK	23
2	MSX1				RIQAK	
		Dr	333	5	RIQAK	21
2	MSX2				RIQAK	
		Dr	335	5	RIQAK	21
	NANOG				RTQMR	
	No predictions made					
	NANOGP1				RTQMR	
	No predictions made					
	NANOGP8				RTQMR	
	No predictions made					
2	NKX2-1				RIQYK	
		Vnd	326	5	RIQYK	19
2	NKX2-2				RIQYK	
		Vnd	363	5	RIQYK	19
3	NKX2-3				RIQYK	
		Vnd	293	5	RIQYK	19
		Bap	276	5	RIQYK	23
		Tin	255	5	RIQYK	16
2	NKX2-4				RIQYK	
		Vnd	326	5	RIQYK	19
3	NKX2-5				RIQYK	
		Vnd	289	5	RIQYK	19
		Bap	268	5	RIQYK	23
		Tin	250	5	RIQYK	16
3	NKX2-6				RIQYK	
		Vnd	283	5	RIQYK	19
		Bap	257	5	RIQYK	23
		Tin	252	5	RIQYK	16
2	NKX2-8				RIQYK	

		Vnd	337	5	RIQYK	19
2	NKX3-1				RIQYK	
		Bap	306	5	RIQYK	23
2	NKX3-2				RIQYK	
		Bap	333	5	RIQYK	23
2	NKX6-1				RVQTK	
		Hgtx	368	5	RVQTK	20
2	NKX6-2				RVQTK	
		Hgtx	371	5	RVQTK	20
4	NOBOX				RVQAK	
		Al	235	5	RVQAK	20
		CG4136	228	5	RVQAK	22
		PhdP	222	5	RVQAK	17
2	ONECUT1				RNMRR	
		onecut	316	5	RNMRR	15
2	ONECUT2				RNMRR	
		onecut	297	5	RNMRR	15
2	OTP				RVQAK	
		Otp	372	5	RVQAK	20
2	OTX1				RVKAK	
		Oc	360	5	RVKAK	19
2	OTX2				RVKAK	
		Oc	364	5	RVKAK	19
2	PBX1				RNGIR	
		Exd	350	5	RNGIR	17
2	PBX2				RNGIR	
		Exd	351	5	RNGIR	17
2	PBX3				RNGIR	
		Exd	346	5	RNGIR	17
2	PBX4				RNGIR	
		Exd	329	5	RNGIR	17
3	PDX1				RIQMK	

		Scr	279	5	RIQMK	25
		Dfd	271	5	RIQMK	24
		Antp	271	5	RIQMK	16
2	PITX2				RVKAK	
		Ptx1	374	5	RVKAK	20
2	PITX3				RVKAK	
		Ptx1	374	5	RVKAK	20
2	PKNOX1				RNIRR	
		Hth	285	5	RNIRR	17
2	PKNOX2				RNIRR	
		Hth	301	5	RNIRR	17
3	PROP1				RVQAK	
		CG32532	281	5	RVQAK	23
		Al	271	5	RVQAK	20
		Pph13	251	5	RVQAK	21
1	PRRX1				RVQAK	
		CG9876	329	5	RVQAK	20
		Pph13	301	5	RVQAK	21
1	PRRX2				RVQAK	
		CG9876	335	5	RVQAK	20
		Pph13	302	5	RVQAK	21
	Predicted				RNIHK	
	No predictions made					
2	RAX				RVQAK	
		Rx	388	5	RVQAK	27
2	RAXL1				RVQAK	
		Rx	374	5	RVQAK	27
	RHOXF1				RVKAR	
	No predictions made					
	RHOXF2				VIEAK	
	No predictions made					
	SATB1				RKQYY	
	No predictions made					

	SATB2				RKQYH	
	No predictions made					
3	SHOX				RVQAK	
		PhdP	280	5	RVQAK	17
		CG11294	276	5	RVQAK	15
		Otp	275	5	RVQAK	20
3	SHOX2				RVQAK	
		PhdP	280	5	RVQAK	17
		CG11294	276	5	RVQAK	15
		Otp	275	5	RVQAK	20
2	SIX1				SNKQR	
		So	361	5	SNKQR	27
2	SIX2				SNKQR	
		So	365	5	SNKQR	27
2	SIX3				TNKQR	
		Optix	372	5	TNKQR	27
2	SIX4				VNKQR	
		Six4	306	5	VNKQR	20
2	SIX5				VNKQR	
		Six4	307	5	VNKQR	20
2	SIX6				TNKQR	
		Optix	367	5	TNKQR	27
1	TGIF1				RNIRR	
		Vis	299	5	RNIRR	22
		Achi	298	5	RNIRR	23
1	TGIF2				RNIRR	
		Vis	311	5	RNIRR	22
		Achi	310	5	RNIRR	23
4	TGIF2LX				KNIRR	
		Vis	239	4	RNIRR	22
		Achi	238	4	RNIRR	23
		Hth	211	4	RNIRR	17
4	TGIF2LY				KNIRR	
		Vis	234	4	RNIRR	22
		Achi	233	4	RNIRR	23
		Hth	206	4	RNIRR	17

2	TLX1				RTQTK	
		C15	345	5	RTQTK	19
2	TLX2				RTQTK	
		C15	328	5	RTQTK	19
2	TLX3				RTQTK	
		C15	347	5	RTQTK	19
	VAX1				RSNFK	
	No predictions made					
2	VAX2				RVQTK	
		Ems	247	5	RVQTK	20
		E5	235	5	RVQTK	43
	VENTX				RTQMK	
	No predictions made					
1	VSX1				RVQAK	
		CG33980	328	5	RVQAK	13
		CG4136	296	5	RVQAK	22
		Rx	291	5	RVQAK	27
1	VSX2				RVQAK	
		CG33980	341	5	RVQAK	13
		CG4136	306	5	RVQAK	22
		Rx	303	5	RVQAK	27