

COMPREHENSIVE STUDY OF THE ZAD FAMILY OF ZINC FINGER  
TRANSCRIPTION FACTORS IN *DROSOPHILA MELANOGASTER*

by

Joseph Krystel

A Dissertation Submitted to the Faculty of  
The Charles E. Schmidt College of Science  
in Partial Fulfillment of the Requirement for the Degree of  
Doctor of Philosophy

Florida Atlantic University

Boca Raton, FL

August 2012

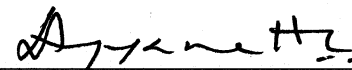
COMPREHENSIVE STUDY OF THE ZAD FAMILY OF ZINC FINGER  
TRANSCRIPTION FACTORS IN *DROSOPHILA MELANOGASTER*

by

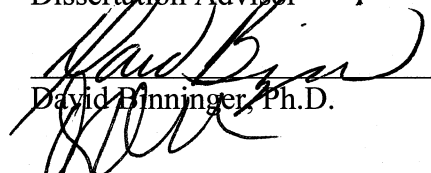
Joseph Krystel

This dissertation was prepared under the direction of the candidate's dissertation advisor, Dr Kasirajan Ayyanathan, Department of Biological Science, and has been approved by the members of his supervisory committee. It was submitted to the faculty of the Charles E. Schmidt College of Science and was accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

SUPERVISORY COMMITTEE:



Kasirajan Ayyanathan, Ph.D.  
Dissertation Advisor

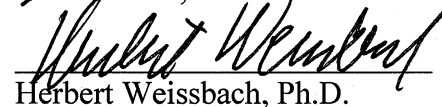


David Binninger, Ph.D.

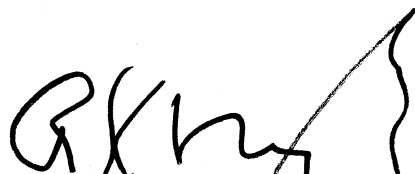
Tanja Godenschwege, Ph.D.



John Nambu, Ph.D.



Herbert Weissbach, Ph.D.



Rodney K. Murphey, Ph.D.  
Chair, Department of Biological Science



Gary W. Perry, Ph.D.  
Dean, Charles E. Schmidt College of Science



Barry T. Rosson, Ph.D.  
Dean, Graduate College

July 11, 2012

Date

## ACKNOWLEDGEMENTS

This author would like to thank Dr. Kasirajan Ayyanathan for his significant conceptual and practical additions to this work. I would like to thank Dr. Herbert Weissbach, Dr. Tanja Godenschwege and Dr. John Nambu for their contributions and guidance. I would also like to thank Dr. David Binninger for both his guidance as a dissertation committee member and for his generous supplying of fly genomic material. I would like to thank Dr. Theodore Haerry for kindly providing *Drosophila* embryonic cDNA libraries, S2 cell line and other fly reagents. I further thank Dr. Cindy Chiang and Edna Gamliel for assistance provided. Funding assistance from the Department of Biological Sciences, FAU as grant-in-aid to K.A. and from National Institutes of Health to K. A. (grant 5K01CA95620) to carry out this work is gratefully acknowledged. Lastly, I would like to thank my parents Jill Krystel and Philip Krystel for their immense personal support.

## ABSTRACT

Author: Joseph Krystel  
Title: Comprehensive Study of the ZAD Family of Zinc Finger Transcription Factors In *Drosophila melanogaster*  
Institution: Florida Atlantic University  
Dissertation Advisor: Dr. Kasirajan Ayyanathan  
Degree: Doctor of Philosophy  
Year: 2012

The zinc finger associated domain (ZAD) family of transcription factors from *Drosophila melanogaster* is not well described in the literature, in part because it is very difficult to study by traditional mutagenesis screens. Bioinformatic studies indicate this is due to overlapping functions remaining after a recent evolutionary divergence. I set out to use *in vitro*-binding techniques to identify the characteristics of the ZAD family and test this theory. I have constructed glutathione S-transferase (GST)-ZAD domain chimeric proteins for use in pull down protein binding assays, and GST-Zinc finger (ZnF) array domain chimera for electrophoretic mobility shift assays (EMSA). Protein binding assays indicated two putative conserved interactors, similar to the analogous KRAB system in mammals. Work is ongoing to isolate and identify these factors. DNA binding assays have provided a consensus binding site sequence for twenty four DNA binding domains (DBD) taken from twenty three independent ZAD proteins. The sequence results are consistent with previously reported work on CG7938, a bioinformatics study of genes regulated by CG17958 and CG11695, and unpublished work on mammalian

SNAG family transcription factors with similar DNA binding structures to CG11695.

Competitive bindings were carried out to show a specificity of binding conferred by the identified conserved positions. While the consensus binding sites show relatively few similarities, the predicted target genes identified by the consensus binding sites show significant overlap. The nature of this overlap conforms to the known characteristics of the ZAD family but points to a more positive selection to maintain conservation of function.

COMPREHENSIVE STUDY OF THE ZAD FAMILY OF ZINC FINGER  
TRANSCRIPTION FACTORS IN DROSOPHILA MELANOGASTER

List of Tables .....	ix
List of figures .....	x
Introduction.....	1
Transcription Factors in Nature .....	1
Mammalian Zinc Finger Protein Super-Families .....	5
Zinc Finger Associated Domain (ZAD) Proteins In <i>Drosophila</i> .....	6
Described ZAD Protein Functions .....	10
ZAD Proteins in Context .....	13
Research into Transcription Factors .....	15
ZAD Significance .....	18
Identification of ZAD Family Binding Sequences and Genetic Targets .....	22
Overview of the Target Gene Search.....	22
Methods utilized.....	23
Expression and Purification of GST-ZAD-ZnF Fusion Proteins.....	23
Binding Site Selection and Derivation of Consensus .....	24
Confirmation of Binding <i>in vivo</i> by Luciferase Assay .....	25
Dot Blot Analysis.....	26
Results and Discussion .....	27

Putative Target Genes .....	27
Correlation to Known ZAD Properties .....	30
DNA Binding Analysis .....	32
Development of an Improved BSS Protocol.....	34
Characterization of ZAD Domain and Identification of Co-Factors .....	37
Overview of Cofactor Analysis .....	37
Methods utilized.....	38
Protein Binding .....	38
Visualization .....	39
Elution Method .....	40
Extract Preparation.....	41
Labeled Binding.....	42
Yeast-2-Hybrid Screening .....	43
Results and Discussion .....	44
Materials and Methods.....	48
Materials .....	48
Methods.....	49
Protein Constructs.....	49
Protein Expression and Purification.....	49
DNA Binding Site Selection (BSS) .....	50
Protein Extract .....	52
Fluorography.....	54
DEAE Binding .....	55

Cloning and Sequencing of pGBKT7-ZAD-ZFP .....	55
Yeast Transformation.....	56
Yeast Protein Expression .....	56
Autoactivation and Toxicity Assays and Yeast Mating.....	58
Yeast Plasmid Preparation .....	58
DNA Sequencing .....	59
Dot Blot Analysis.....	59
Figures and Tables .....	62
Bibliography .....	112



## TABLES

Table 1. Archetypal ZAD Family Members Selected for BSS Analysis.....	106
Table 2. Primer Sequences Used to Amplify ZnF Domains from ZAD Family Members Used in BSS Analysis .....	107
Table 3. Consensus Binding Sequences Identified For 23 ZAD Family Members.....	108
Table 4. Oligonucleotide Sequences Used in Competitive Binding Experiments for CG7938, CG17958, CG12219, CG30020, and CG11695.....	109
Table 5. A Quantification of Predicted or Known Targets for ZAD Proteins Against Homeobox Containing Early Developmental Genes .....	110
Table 6. Primer Sequences Used to Amplify ZAD Domains from ZAD Family Members Used in Cofactor Analysis .....	111

## FIGURES

Figure 1. ZnF Domains identified from full length templates .....	62
Figure 2. ZnF Domains amplified and digested.....	64
Figure 3. Protein mini-induction experiment using positive GST-ZnF recombinant clones .....	65
Figure 4. GST-ZNF construction protein inductions were purified by GSH bead columns.....	66
Figure 5. Diagrammatic representation of each of the archetypal ZAD family members used in the second round of selections.....	67
Figure 6. Diagrammatic representation of each of the second set of GST-ZnF constructs .....	68
Figure 7. Purifications of GST-ZfP proteins on GSH affinity bead columns.....	69
Figure 8. GST-ZnF construct Proteins: Multiple independent clones .....	70
Figure 9. Binding site selection experiment .....	71
Figure 10. DNA-protein complexes from the first binding site selection against the random oligonucleotide library.....	72
Figure 11. Products from the final enriched EMSA selection .....	73
Figure 12. Restriction digests from CG14710 .....	74
Figure 13. Clones sent for sequencing.....	75

Figure 14. Sequence data from multiple independent clones from the CG30020	
binding site selection using EMSA.....	76
Figure 15. Sequence data from multiple independent clones from the CG12219	
binding site selection using EMSA.....	77
Figure 16. Sequence data from multiple independent clones from the CG7938	
binding site selection using EMSA.....	78
Figure 17. Sequence data from multiple independent clones from the CG17958	
binding site selection using EMSA.....	79
Figure 18. Sequence data from multiple independent clones from the CG11695	
binding site selection using EMSA .....	80
Figure 19. Binding site selection and competitions .....	81
Figure 20 Competition analysis of a protein selected under the modified binding	
site selection procedure.....	82
Figure 21. Competition analysis developed using a modified dot blot analysis.....	83
Figure 22. The genes predicted as targets for CG18555 and CG7928 .....	84
Figure 23. Relative incidents of potential target genes with primary functions .....	85
Figure 24. An analysis of the prevalence of secondary gene functions.....	86
Figure 25. Comparisons between five reported serendipity binding site sequences .....	87
Figure 26. A ClustalW comparison between the conserved DNA binding domains.....	88
Figure 27. A ClustalW comparison between the conserved DNA binding domains.....	89
Figure 28. A schematic representation of the coupled cold and hot binding protocol .....	90
Figure 29. Representative EMSA results from ZAD family members.....	91

Figure 30. Diagrammatic representations of the recombinant GST-ZAD fusion proteins.....	92
Figure 31. Diagrammatic representation of GST-ZAD protein binding assay .....	93
Figure 32. ZAD domains amplified by polymerase chain reaction (PCR) from cDNA .....	94
Figure 33. GST-ZAD protein inductions .....	95
Figure 34. Fractionated GST-ZAD protein induction.....	96
Figure 35. The Purification of GST-ZAD construct proteins expressed in E. coli cells visualized on a SDS PAGE gel.....	97
Figure 36. GST pull down assay to identify putative ZAD-interacting proteins.....	98
Figure 37. GST pull down assay diagram.....	99
Figure 38. A GST-pull down assay with the CG12219, CG11695, CG9233, CG10108, and CG2889 ZAD-GST constructs .....	100
Figure 39. Associated proteins from the fly cell nuclear extracts.....	101
Figure 40. Protein samples used in the pull-down assays.....	103
Figure 41. Protein binding pull-down assay repeated with <sup>35</sup> S.....	104
Figure 42. Diethylaminoethyl-agarose protein binding assay .....	105

## 1. INTRODUCTION

### 1.1 TRANSCRIPTION FACTORS IN NATURE

The intricacy of life has challenged our understanding again and again. It has done so not only in terms of the total complexity but also in how those diverse structures and behaviors arise. Many biologists once postulated that protein must be the molecule to store genetic material as it's more numerous amino acids would be needed to contain such vast amounts of information. The simple 4 base pairs of DNA molecules must be structural in nature. Before systematic genome sequencing efforts, it was widely expected that more complex organisms like humans and other primates would possess many more genes than their distant simple relations. With data in hand, the estimated 100,000 (Schuler et al., 1996) human protein coding genes dwindled to approximately 20-25,000, little more than the lowly fruit fly and well below that of rice's astonishing 43,000 (Carninci and Hayashizaki, 2007). Our current conceptualization now includes alternatively spliced gene variants and an ever expanding cadre of control mechanisms whose interactions precisely regulate fewer but more versatile genes.

The proper development and maintenance of functional organism requires genes to be expressed at specific times, coinciding with appropriate external and internal conditions. The expression of these genes must be tightly regulated to ensure proper

development and maintenance of the organism. To ensure that these genes are transcribed in a spatiotemporal fashion; a whole suite of regulatory apparatus are available. These regulatory elements, including both pre-transcriptional and post-transcriptional systems, function to activate and/or silence the target genes when necessary. One major pre-transcriptional system includes several large families of genes that code transcription factor proteins. These transcription factors are among the best-studied and most significant players in that control. Each transcription factor can interact with environmental cues, signal molecules, cofactors, each other, histones and DNA to increase or decrease gene expression. They do so in part by modifying the protein framework containing DNA and the physical structure of the DNA itself. Through these modifications, the transcription factors may act to prevent or facilitate the physical access of the transcriptional machinery to the DNA sequence (Morse, 2007). Transcription factors bind to DNA and modify the local chromatin structure to either facilitate or hinder the access of the transcriptional machinery (Morse, 2007). They may do so alone or in conjunction with other transcription factors and co-factors. A single transcription factor may regulate an entire suite of genes, effecting major phenotypic changes in the organism. Transcription factors thus have a multiplicative effect, allowing for more combinations of a given number of protein genes. Much of the variation in structures and responses in higher eukaryotes can be attributed to the complexity afforded by transcription factors.

The human genome consists of approximately 25,000 genes. Roughly 2,500 of those genes, or about 10%, serve to regulate the expression patterns of the remaining 22,500 genes (Consortium, 2004; Lodish, 2004). Of these ~2,500 regulatory genes

coding for transcription factors, approximately one third of them code for zinc finger proteins (Huntley et al., 2006). This ratio is relatively consistent between many other eukaryotes. Zinc finger proteins are named for the presence of a particular type of zinc chelating domains. They are the most abundant and one of the best understood DNA binding domains coded for by eukaryotic organisms (Fu et al., 2009). One of the most common type of transcription factor in humans and all other eukaryotes is the C<sub>2</sub>H<sub>2</sub> zinc finger transcription factor (ZFP) (Duan et al., 2008). First identified in the late 1980's, these proteins are so named for their DNA binding domain that consists of a tandem array of C<sub>2</sub>H<sub>2</sub> zinc finger domains (Klug and Rhodes, 1987). They are further grouped based on the presence of one of several different possible effectors domains that provide the regulatory activity; in the form of transcriptional activation, transcriptional repression, or as a basis for recruiting additional members with these activities. Possible effector domains and their associated families include the KRAB domain, BTB/POZ domain, ZAD domain, and SCAN domain (Collins et al., 2001). Understanding these transcription factors is key to understanding how complex organisms develop. Discovery, structure, and biomedical applications are recently reviewed (Klug; Klug). Zinc-finger domains contain a series of very well conserved amino acid residues that interact with a zinc atom. Zinc finger proteins may be grouped according to which of these conserved sequences are present. Two of these groups are the Cys4 (C4) and Cys6 (C6) types, which are recognized by the specific placement of four or six cysteine residues (Falquet et al., 2002; Witte and Dickson, 1990). A third group of zinc finger proteins contain two conserved cysteine and two conserved histidine residues. Members of this group are similarly named Cys2His2 (C2H2) type zinc fingers. Each C2H2 domain is 25 to 30

amino acids in length, with the four conserved amino acids in pairs on either end of the domain acting in a tetrahedral binding of zinc cation and interacts with a set number of DNA nucleotides (Rosenfeld and Margalit, 1993). The zinc fingers work in a modular fashion, with each member able to extend the consensus recognized by the array. Amino acid replacement studies of zinc fingers have shown these well-conserved residues to be vital to the function of the domains. Even a single amino acid replacement may result in a total loss of function (Croizatier et al., 1992).

The classical transcription factor architecture involves two key structures. Each functions in a modular fashion, able to operate independently of the other. The first is an effector domain that imparts the transcriptional regulation activity by recruiting the necessary machinery to modify the local chromatin structure. Transcription factors are grouped into families based on the presence of different conserved effector domains. The second structure is a DNA binding domain that will interact with a specific DNA sequence to properly position of transcription factor near the genes it will regulate. In the zinc finger proteins the tandem arrays of  $C_2H_2$  zinc fingers serve the DNA binding function. This modular ZnF DNA binding array and effector domain architecture has proven to be very versatile that has undergone strong positive selection in most of the higher eukaryotes. The only significant difference seems to be the specific family of zinc finger proteins that has been selected for expansion, which varies from one lineage to another.

$C_2H_2$  zinc fingers functions are possible through the close interactions to the divalent cation, which provides the necessary compact and stable fold structure needed to recognize variations in the DNA molecule's major groove consistent with specific base



pair sequences (Klug and Rhodes, 1987; Miller et al., 1985). Each domain typically interacts with three adjacent nucleotides, primarily by directly contacting with the amino acids at positions -1, +3 and +6 relative to the beginning of the alpha helix. This region has been termed the recognition helix. In some instances other residues may influence the binding or even directly contact the DNA (Fairall et al., 1993). Larger sequences can be recognized by multiple ZnF domains expressed in an array, in which case both the variable positions within the ZnF and the overall framework of the array contribute to the specific sequence recognized (Fu et al., 2009; Isalan et al., 1998). Zinc fingers are not only DNA binding domains, but are also known to interact with proteins, RNA, and other small molecules (Krishna et al., 2003; McCarty et al., 2003) and to also serve as nuclear localization sequences (Mingot et al., 2009). Further, it is possible for arrays that mediate a DNA binding function to also be involved in protein-protein interaction or homodimer formation (Brayer et al., 2008; Brayer and Segal, 2008; Jauch et al., 2003; Mackay and Crossley, 1998).

## **1.2 MAMMALIAN ZINC FINGER PROTEIN SUPER-FAMILIES**

Zinc finger proteins may also be grouped into families based on the presence of conserved amino-terminus domains. Approximately 400 of the human zinc finger proteins belong to the Kruppel Associated Box (KRAB) domain zinc finger protein superfamily (Huntley et al., 2006). This very large family of transcription factors has been implicated in several biological processes. Each member, with very few exceptions, is structured in a similar manner with one conserved amino-terminus KRAB domain and

one or more tandem arrays of C<sub>2</sub>H<sub>2</sub> zinc fingers towards the C-terminus. Similarities discussed in later sections. The zinc finger arrays bind to specific DNA binding sites and the KRAB domain interacts with the Ring finger-B boxes-Coiled-Coil (RBCC) domain of the KRAB Associated Protein 1 (KAP-1). KAP-1 serves as a universal cofactor for the KRAB domain transcription factors. It functions as a molecular scaffold for recruiting a protein complex that coordinates the histone deacetylation, methylation, and heterochromatin protein 1 (HP1) deposition needed to silence the target gene (Ayyanathan et al., 2003).

### **1.3 ZINC FINGER ASSOCIATED DOMAIN (ZAD) PROTEINS IN *DROSOPHILA***

The analogous family to the *Drosophila* ZAD family in mammals is the KRAB superfamily. These proteins share a number of similarities beyond the basic C<sub>2</sub>H<sub>2</sub> zinc finger architecture. The ZAD ZFPs are present in roughly the same ratio as KRAB, making up approximately one third of the total zinc fingers and one tenth of the total transcription factors in *Drosophila* (Chung et al., 2002) (Chung et al., 2002). Both families also display a high degree of lineage specific enrichment and clustering at distinct chromosomal locations (Chung et al., 2007; Hamilton et al., 2003). Structurally, KRAB and ZAD proteins are only differentiated by the type of effector domain present. That effector domain is located at the protein amino-terminus in each case. That effector is a known protein-binding domain that has been either shown or predicted to recruit

chromatin modifying complexes. Like other zinc finger proteins, they also share the C-terminal zinc finger arrays for DNA binding.

The ZAD-ZFP family displays a high degree of lineage specific enrichment and clustering at distinct chromosomal locations, a pattern that has also been observed in the KRAB superfamily (Chung et al., 2007; Hamilton et al., 2003). Of the total 98 different ZAD protein-coding genes present in the *Drosophila* genome, each is present in a single copy. However, at least four ZADs do exhibit alternative splicing. Most members are located on the third chromosome (N=54) with the rest primarily present on chromosome two. A full 28% of the 326 ZFPs in *Drosophila* are conserved through eukaryotes from *C. elegans* to *Homo sapiens*. In comparison, only one ZAD protein has been found through vertebrate genomes (Chung et al., 2007). This makes ZAD proteins both the most abundant ZFP family and one of the most dipteran-specific (Duan et al., 2008; Chung et al., 2007). ZAD proteins are also enriched within closely related mosquito lineages, with only 9 of the 98 *Drosophila melanogaster* ZADs identified as being present at the speciation event between the *Drosophila* genus and the *Anopheles* genus (Chung et al., 2007) with one other possible paralogue between *Drosophila* and *Culex* (Curwen et al., 2004).

ZAD-ZFP architecture closely resembles that of KRAB proteins. They have a very well conserved amino-terminus domain common to all ZADs, with a series of variable zinc finger arrays towards the C-terminus. The ZAD domain is a Cys<sub>2</sub>Cys<sub>2</sub> zinc finger domain (Hamilton et al., 2003). Only three members of the ZAD family show any significant variation from the standard N-terminal ZAD domain architecture. The ZAD domain in CG6689 is preceded by a 90 amino acid C<sub>2</sub>CH type (THAP) zinc finger

domain that is similar to the DNA binding domain of *Drosophila P* element transposase (Roussigne et al., 2003). This places the ZAD domain 165 amino acids from the N-terminus, as compared to the median distance of 11 seen among other ZAD proteins. There are several other members with the ZAD domain similarly far from the N-Terminus, but no other member with an identifiable domain distal to ZAD. The ZAD domains present in Molting Defective (CG34100) and GATAd (CG5034) appear to have been bifurcated by an insertion (Krystel et al., 2009). Molting Defective's ZAD domain matches the consensus sequence from amino acid residues 62-86 and 216-156, with an apparent 129 amino acid insertion (Krystel et al., 2009). GATAd's ZAD domain matches the consensus sequence from amino acid residues 11-34 and 213-285, with an apparent 178 amino acid insertion (Krystel et al., 2009).

Like KRAB, the ZAD C-terminus zinc finger arrays are primarily C<sub>2</sub>H<sub>2</sub> zinc fingers, with a few interesting variations. Alternate types of zinc fingers, if present at all, are mostly contained within the C<sub>2</sub>H<sub>2</sub> arrays and are generally few in number. The median value those ZAD ZFPs with any alternates is presence of one zinc finger. However, they may have as many as five non-C<sub>2</sub>H<sub>2</sub> zinc fingers, as is the case in CG8145 (Krystel et al., 2009).

A closer examination of the composition of ZAD ZFPs provides some noteworthy results. If I consider any zinc finger domains separated by less than 25 amino acids (the size of one full zinc finger domain) as in array; the majority of ZAD ZFPs contain only one zinc finger array (75/98 including isoforms) with a median number of five C<sub>2</sub>H<sub>2</sub> zinc fingers present (Krystel et al., 2009). A further 12 ZAD ZFPs contain two zinc finger arrays, and ten more ZAD ZFPs contain three or more arrays. In ZAD ZFPs with two

zinc finger arrays, the standard structure consists of one array of length similar to those seen in one array ZADs and one smaller than median array (<4) or even single zinc finger set apart from the first array. Single zinc fingers are actually more common than a second small array. On the extreme end, CG32575 has seventeen zinc fingers spread across eleven different strings, both small arrays and in isolation (Krystel et al., 2009).

Interestingly, four of the ZAD ZFPs have no C<sub>2</sub>H<sub>2</sub> zinc finger arrays at all. With no significant DNA binding functions, these ZAD ZFPs may serve to inhibit the function of other ZAD ZFPs through sequestration or competition for the cofactors. Similarly structured mammalian KRAB proteins have been shown to function in this manner. Approximately one quarter of ZAD ZFPs also contain nuclear localization sequences; this includes fifty percent of the ZAD ZFPs lacking C<sub>2</sub>H<sub>2</sub> arrays. This higher rate of NLS within non-C<sub>2</sub>H<sub>2</sub> containing ZAD proteins may be related to the recently identified examples of intrinsic NLS activity within C<sub>2</sub>H<sub>2</sub> Arrays (Krystel et al., 2009).

There are also a few other instances of potentially notable domains within the ZAD family. The protein produced by CG1647 appears to contain a domain similar to the HSP20 heat shock protein domain (Krystel et al., 2009). Also present are possible DNA binding domains such as the previously mentioned THAP domain in CG6689 and CG6813, which contains a copper fist domain. Copper fists are metal chelating structures similar to zinc fingers, but associating with divalent copper cation instead of zinc and are predicted to have DNA binding functions (Thorvaldsen et al., 1993).

The modular domains that comprise the majority of transcription factors can often remain active when paired with the other necessary domains from a heterologous transcription factor. The DNA binding domain will continue to position the protein at the

DNA binding site without its normally associated effector domain (Brent and Ptashne, 1985; Liu et al., 2001). The overall backbone of the zinc finger array has been shown to play a role in DNA recognition. In most cases this has been primarily an effect within the array between adjacent zinc fingers (Fu et al., 2009). Instances of ZAD proteins in which regions outside of the array were significant in maintaining binding activity included only members shown to require dimerization to become functional (Jauch et al., 2003; Payre et al., 1997). These two members were also identified as self-interacting in high throughput yeast-2 hybrid screenings. No other ZAD members have been reported to form such a structure. The interchangeability of effector and DNA binding domains has been exploited by a number of groups to create chimeric transcription factors that combine an effector domain of desired activity with a zinc finger array sufficient to target the gene to be regulated.

#### **1.4 DESCRIBED ZAD PROTEIN FUNCTIONS**

Very few of the functions of ZAD containing zinc finger proteins in *Drosophila* have been identified. The lack of knowledge about ZAD proteins is in part due to their resistance to the commonly utilized mutagenesis screens. Approximately half of the ZAD proteins are available in knockout or significant knockdown lines, by way of P element insertions and RNAi expression, but most do not present a discernable phenotype. Studies suggest that a relatively recent expansion of ZAD proteins may have resulted in the conservation of function across the ZAD ZFP family (Chung et al., 2002). Overlapping functions would explain difficulty in elucidating their functions, and the

lineage specific enrichment seen in ZAD ZFPs may support the theory of a recent expansion of the family. They have further suggested that the expansion of ZAD proteins may be associated with the development of adaptive structures, specifically the merostic ovary which shares a close phylogenetic correlation to ZAD expression (Chung et al., 2007). Corollary to this theory would be the anticipation that those few ZAD proteins with a severe and notable phenotype are the exceptions that have acquired a necessary and not merely adaptive function. This is in part supported by the fact that only three of the nine ZAD proteins with necessary functions in *Drosophila* are conserved between closely related dipterans (Chung et al., 2007); as well as the observation that the least characteristic ZAD proteins, as is the case with GATAd and Molting Defective, often possess a noticeable phenotype or function.

Two of the well-described ZAD ZFPs are Grauzone and Serendipity-delta. Both of them have been shown to act in transcription regulation, more specifically to serve as transcription activators. Grauzone has been found to activate the gene Cortex, which encodes an Anaphase-Promoting-Complex (APC) subunit (Chen et al., 2000; Chu et al., 2001; Harms et al., 2000). Similarly, Serendipity has been shown to activate the bicoid gene involved in egg polarity (Payre et al., 1994). The ZAD domains of both Serendipity-delta and Grauzone have been shown to function as protein binding domains involved in forming Ser-d/Ser-d and Grau/Grau homodimers (Jauch et al., 2003; Payre et al., 1997).

Molecular interactions of the ZAD domain in Grauzone has been examined (Jauch et al., 2003). A Grau-ZAD-GST construct containing the amino acid residues 2 to 90 from the amino-terminus of Grauzone that comprises of the ZAD domain was made.

Later, they performed both glutaraldehyde chemical crosslinking experiments and multi-angle-laser-light-scattering, following size exclusion chromatography. Both procedures yielded strong evidence of homo-dimerization, with no higher oligomeric states indicated. This supported their hypothesis that the crystal structure of Grauzone's ZAD domain would be most stable forming a head to tail dimer. This model places the ZADs in such a fashion that the largest conserved surface patches coincide with the dimer interface. This caused the interface to contain 72.5% non-polar amino acid residues, giving the interface a very different character than the rest of the protein's surface (Jauch et al., 2003). High throughput yeast-2 hybrid screening has also identified Grauzone as interacting with Grauzone. This has not been seen in similar screening on many of the cryptic ZAD proteins (Breitkreutz et al., 2008; Giot et al., 2003).

Another study compared the binding characteristics of the wild type and several mutant Serendipity proteins by utilizing the reporter plasmids pTKCAT with one to four tandem repeats of the Serendipity-Delta Consensus Binding Sites (SDBS) (Payre et al., 1997). Their results indicated binding to the Specialized Chromatin Structure (SCS) occurred in a manner consistent with dimer formation. Through systematic deletion analysis they identified locations vital to the formation of this homodimer. They found that mutations in either the first amino-terminus zinc finger domain (ZAD) or the sixth zinc finger in the array of C<sub>2</sub>H<sub>2</sub> zinc fingers caused the loss of this dimer form. This is especially interesting, because all six C<sub>2</sub>H<sub>2</sub> zinc fingers in Serendipity-Delta are also known to be necessary for the specific binding of the SDBS. This means that the sixth C<sub>2</sub>H<sub>2</sub> zinc finger motif functions both as a protein-binding domain (with ZAD) and as a DNA binding domain. This study clearly supports a head to tail homodimer formation



model, but with ZAD-C<sub>2</sub>H<sub>2</sub> interaction, unlike the ZAD-ZAD interaction observed in Grauzone (Jauch et al., 2003).

Two other described ZAD ZFPs are termed Poils-au-dos and Zeste-white 5. Both of these ZAD ZFPs have been shown to act as transcription factors, more specifically as transcription repressors. Poils-au-dos is found to inhibit the transcription of the Achaete and Scute genes. The Achaete and Scute genes code for the production of transcription factors responsible for the proper arrangement of the large bristles on the *Drosophila* natum. Poils-au-dos also collaborates with Hairy and extramacroschaetae repressors to dominantly inhibit Achate and Scute. It was further seen to have a strong genetic interaction with the Punt (put) and Thickveins (TKV) mutants, suppressing the wild type phenotype or enhancing the mutant phenotype of heterozygote mutants (Gibert et al., 2005).

Zeste-white 5 has been shown to play a role in the Specialized Chromatin Structure (SCS) domain related nuclear protein complex that blocks the enhancer-promoter interaction of the 87A7 heat shock domain. This study concluded that both the amino-terminus zinc finger domain (ZAD) and the C2H2 zinc finger arrays were necessary for this insulator activity; the arrays act to locate and bind specific sequences of DNA and the ZAD domain is involved in specific protein-protein interactions needed to confer the transcriptional activity (Gaszner et al., 1999).

## **1.5 ZAD PROTEINS IN CONTEXT**

*Drosophila melanogaster* is one of the most utilized model organisms for genetic

and molecular studies. The ease of growth, availability of powerful techniques, and relatively high incidence of homology with human disease states has contributed to this status. Because of the prevalence of its use, filling in gaps in our current understanding take on a special importance. *Drosophila* dedicate nearly 1% of their genome to creating ZAD proteins (closer to 10% of their total transcription factors), and yet they are at best poorly represented in the literature (Drysdale, 2008). The family as a whole cannot be identified as possessing any conserved activity (transcriptional repression or activation), and the targets and functions of ~90% of the members are totally unknown. The application of proven techniques used in similar mammalian systems instead of more traditional *Drosophila* techniques will allow the vigorous testing of current theories as to ZAD origins and function.

Various zinc finger protein families have been found to be expanded within different eukaryotic lineages. While the most prominent examples are the KRAB domain in mammals and the ZAD domain in some insects, the positive selection of one or more of these families of versatile transcription factors has occurred independently across the spectrum of eukaryotic lineages. Even at a species specific level, the number of unique ZnF proteins can be quite high; 55.6%, 43.9%, 76.8%, and 21.5% respectively in *B. mori*, *D. melanogaster*, *C. elegans*, and *H. Sapiens* (Duan et al., 2008). Each expansion has shown a similarity of formation, with an uneven clustering on the chromosomes as described in L(3)Neo38, Tiptop, BR-C, Fru, Hkb, Ab, Ken, and Sens in nematodes (Duan et al., 2008; Haerty et al., 2008), KRAB, SNAG, and BTB in mammals (Collins et al., 2001; Huntley et al., 2006), and ZAD in dipteran insects (Chung et al., 2007; Chung et al., 2002), coinciding evolutionarily with the potential development of novel adaptive

structures and phenotypes. Understanding the development of the ZAD family of zinc finger proteins will provide insight into the evolutionary history and formation of lineage specific features far beyond *Drosophila*.

## **1.6 RESEARCH INTO TRANSCRIPTION FACTORS**

Molecular transcription factor research has focused primarily on the two modular domains and their functions. Knowing where a transcription factor binds and the effect it has on gene expression goes a long way towards understanding the role it plays in biologic processes. This is not an exhaustive study, as more complex interactions with weak effectors displacing powerful ones and similar contextual situations will still occur.

Identifying the specific nucleotide sequence bound by a transcription factor is an important early step in characterizing both its molecular and biologic functions. This may be carried out by either selecting for the DNA sequence that most efficiently binds the protein or by locating the native targets within the genome. Both approaches have advantages and disadvantages.

Techniques to identify sequences, which efficiently bind a protein may be constructed in different ways, but in general they involve repeated selections of a random collection of oligonucleotides against the protein of interest. Because only a very small number of the total sequences present in a random library will bind the protein, this method requires multiple selection rounds and a very powerful examination technique to identify the signal. The archetypal method for this approach used radio-labeled double stranded oligonucleotides with  $^{32}\text{P}$ . The oligonucleotides each contain a variable region

large enough to represent the expected binding site and short enough to make analysis reasonably possible. Zinc finger arrays generally recognize three or four nucleotides per finger participating in the binding (Choo and Klug, 1994; Wu et al., 1995). The labeled family of oligonucleotides is then bound under near cellular conditions to a purified form of the protein. That protein may be obtained by way of antibody selection or generated in an ex-vivo expression system and then purified by affinity chromatography. Once bound, the combined protein and DNA sample is then run on a non-denaturing polyacrylimide gel. Oligonucleotides that are unbound will pass at normal speed through the gel whereas those incorporated into a larger DNA-Protein complex will be retarded in their movement. This shift separates the two populations of molecules. Those that bind the protein may then be recovered from the gel and amplified by PCR into a new library enriched in molecules that bind the protein. The binding is then repeated multiple times to further enrich the library until a point when it contains primarily those members that efficiently bind the protein. That library is then sequenced and an analysis of the sequences for shared motifs will identify a consensus binding sequence. Several protocols involving a filtering on immobilized protein have been reported in the past, using different techniques such as nitrocellulose filters or southwestern blot analysis (Swirnoff and Milbrandt, 1995; Thiesen and Bach, 1990).

Identifying a consensus binding in this method does not directly identify genes within the organism that are bound and regulated by the protein. Other *in vivo* conditions such as the presence or absence of cofactors, the binding of other more strongly associated transcription factors, and epigenetic factors may all play significant rolls in this process. It will however provide a means to identify potential target genes, even

those that may not be associated with the transcription factor under the growth conditions typically used. It also provides information about the molecular function of the zinc fingers in the array. Because zinc finger arrays are modular in function, this information can be invaluable in constructing artificial transcription factors for genetic research (Beerli et al., 2000).

The second approach for identifying binding sequences for a transcription factor is to collect and isolate those sequences bound in-vivo by the protein. A common protocol would involve binding the protein to a sample of naked genomic DNA and then exposing the bound DNA to a very low concentration of DNAase. Those portions bound by the protein will be shielded from the nuclease activity and will not be degraded. Either affinity chromatography or antibody precipitation can be used to isolate these fragments of DNA, which may then be sequenced (Payre and Vincent, 1991). This method does immediately provide the researcher with regions targeted by the transcription factor under the conditions used. However, the sequence may include nucleotide unnecessary to the binding activity. Bases that are physically covered by the protein may not be directly involved in the protein-DNA interaction. Those sequences may be maintained in the genome by positive selection due to targeting by another transcription factor with a binding site that overlaps for regulatory reasons. This has been shown to be possible in genes such as MRF4, where overlapping binding sites for TBP and MEF2 are contained in one region (-26 to -15) with specific nucleotides being required for each binding. (Naidu et al., 1995). They may also be conserved as binding sites of another protein that acts as a binding site competitor for regulation of

transcription factor activity. This is seen in the binding of EF1 to the E2 Box (Sekido et al., 1994).

Other more gene specific methods are also possible. If a known target gene has been identified through other means, a sequential deletion of its upstream regulatory region can reveal the region necessary for the binding activity. If then combined with EMSA against oligos tailored to mimic that region, a single sufficient binding sequence may be characterized. An example of this method is described in a study by Harms' group (Harms et al., 2000).

An additional strategy has been used, beginning with a known DNA sequence and presenting to it- by means such as phage display- a wide array of different C<sub>2</sub>H<sub>2</sub> zinc finger domains (Dreier et al., 2001). This method does not provide any direct information about a particular transcription factor, but it does allow for the creation of zinc finger libraries to construct artificial transcription factors. The method is complicated by the contextual effect wherein adjacent zinc fingers slightly affect the binding activity, so multiple rounds of design may be required before the desired binding is achieved (Greisman and Pabo, 1997; Isalan et al., 1998).

## **1.7 ZAD SIGNIFICANCE**

*Drosophila melanogaster*. dedicate nearly 1% of their genome to creating ZAD proteins and closer to 10% of their total transcription factors (Benson et al.). They are more numerous, specific to the *Drosophila* lineage, and more expressed in the critical early embryo development period when compared to other families of transcription

factors (Adryan and Teichmann; Adryan and Teichmann, 2006). Yet they are at best poorly represented in the literature (Drysdale, 2008). The family as a whole cannot be identified as possessing any conserved activity, either transcriptional activation or repression. The targets and functions of over 80% of the members have not been reported in the literature. The lack of knowledge about ZAD proteins is in part due to their resistance to the commonly utilized mutagenesis screens. Approximately half of the ZAD proteins are available in knockout or significant knockdown lines, by way of P-element insertions and RNAi expression, but most do not present a discernable phenotype. Only a single ZAD appears to be present at the time of divergence between crustaceans and holometabolous insects. Since then the ZAD family has quickly grown to contain many of the transcription factors in each of the members; 29 within *Apis mellifera*, 75 within *Tribolium castaneum*, 86 within *Bombyx mori*, 98 within *Drosophila melanogaster*, and 147 within *Anopheles gambiae*. This relatively recent expansion of ZAD proteins may have resulted in the conservation of function across the ZAD ZFP family. Overlapping functions would explain difficulty in elucidating their functions, and the lineage specific enrichment seen in ZAD ZFPs may support the theory of a recent expansion of the family. This expansion in ZAD proteins may be associated with the development of adaptive structures, specifically the meroistic ovary, which shares a close phylogenetic correlation to ZAD expression (Chung et al., 2007; Chung et al., 2002).

If this theory is correct, I would anticipate that those few ZAD proteins with a severe and notable phenotype are the exceptions that have acquired a necessary and not merely adaptive function. This is in part supported by the fact that only three of the nine ZAD proteins with necessary functions in *Drosophila* are conserved between closely related

dipterans (Chung et al., 2007); as well as the observation that the least characteristic ZAD proteins, as is the case with GATAd and Molting Defective, often possess a noticeable phenotype or function. I found in the course of categorizing ZAD proteins (unpublished data) that the ZAD domains present in Molting Defective (CG34100) and GATAd (CG5034) appear to have been bifurcated by an insertion. Molting Defective's ZAD domain matches the consensus sequence from amino acid residues 62-86 and 216-156, with an apparent 129 amino acid insertion. GATAd's ZAD domain matches the consensus sequence from amino acid residues 11-34 and 213-285, with an apparent 178 amino acid insertion (Bateman et al., 2002; Bateman et al., 2004).

The remaining ZAD proteins should then be clustered around the pathways needed to produce the adaptive structures. Identifying the small cluster of ZADs with similar functions and knocking their expression down in tandem should reveal the previously cryptic functions. Given their early developmental and neural expression patterns, those previously masked phenotypes may offer excellent model systems for other neuronal development research.

Transcription factors and associated genes have been prolific targets for small molecules and mutations in transcription factors are common in many disease states. In addition to the implications inherent in understanding the control systems that regulate biological development, further understanding of the molecular functions of these transcription factors has biomedical and research technique potential. In recent years several groups have been working on ways to adapt the very versatile ZFP system to design customized transcription factors for specific applications. Because the DNA binding domain acts independently of the effector and the individual zinc finger domains



can be assembled into an array to target a larger less common sequence, it is possible to target a domain containing any desirable activity to nearly any sequence properly positioned regulate a gene of interest (Gogos et al., 1992; Liu et al., 2001). That activity may be in the form of a transcriptional repressor or activator or even nuclease activity as a precursor for using homologous recombination (Choo et al., 1997)(Bae 2003, Fu 2008, Choo 1997, Reviewed in Cathomen 2008).

## 2. IDENTIFICATION OF ZAD FAMILY CONSENSUS BINDING SEQUENCES AND GENETIC TARGETS

### 2.1 OVERVIEW OF THE TARGET GENE SEARCH

Our first avenue of investigation into the ZAD family was to identify the DNA sequences recognized by the individual members. From there I then located the genetic positions of those sequences and determined what genes are thus likely under the control of a ZAD family transcription factor. By doing so I sought to understand the developmental importance of the ZAD family. Their large numbers and expression in the key developmental stages of the *Drosophila* embryo indicates a significant function. However, that function has remained elusive in all previous studies. Knowing the overall character of those genes targeted by the family as a whole, I could then infer their developmental significance. Additionally, knowing specific target genes for those members currently completely un-described in the literature would allow for more specific analysis in future studies to avoid whatever pitfalls prevented previous mutagenesis screens from identifying the mutant phenotypes. If the binding sequences or target genes were shared amongst a cohort of ZAD proteins, it would also allow for co-knockdown experiments to reveal those same phenotypes in more detail.

## 2.2 METHODS UTILIZED

### A. EXPRESSION AND PURIFICATION OF GST-ZAD-ZNF FUSION PROTEINS

Initial binding site selections were performed with the DNA binding zinc finger arrays from the following five ZAD proteins: CG7938 (sry- $\beta$ ), CG17958 (sry- $\delta$ ), CG30020, CG11695, and CG12219. The zinc finger arrays were amplified by PCR from cDNA clones provided by Open BioSystems. A diagrammatic representation of the GST-ZnF constructs is shown in **Figure 1A**. A comparison including the natural forms of a ZAD member and a KRAB member is included in **Figure 1B**. Initial PCR products can be seen in **Figure 2**. The products were then ligated into a pGEX-4T2 plasmid vector for expression of GST-affinity tagged construct proteins. Ligated plasmids were transformed into DH5a *E. coli* cells for bulk plasmid expression. Plasmids were checked for correctly sized insert by restriction enzyme digestions and correctly constructed plasmids were transformed into cells of the BL21 strain of *E. coli* for protein production. Multiple independent clones were checked for correct size and solubility of protein. Protein inductions from multiple independent clones are shown in **Figure 3**. The best expressing clones were taken for further study. Large protein inductions were purified on GSH-agarose bead columns. Representative purification on SDS-PAGE gel analysis is shown in **Figure 4**. A second round of GST-ZnF constructs for the additional **21** ZAD family members were built using PCR amplified zinc finger domains from a collection of *Drosophila* embryonic cDNA libraries. Representations of the structures of those additional members are shown in **Figure 5** and diagrammatic representations of the second round of GST-ZnF constructs in **Figure 6**. Detailed information about the

archetypal members selected is presented in **Table 1** with the primer sequences used for amplification of the zinc finger arrays in **Table 2**. SDS-PAGE gel analysis of the purified protein eluted from GSH columns is shown in detail for several members in **Figure 7** and in brief for all members in **Figure 8**. Proteins eluted from the column were dialyzed using 5kDa cutoff dialysis tubing to increase concentration and also to remove the reduced glutathione before performing the binding assay.

## **B. BINDING SITE SELECTION AND DERIVATION OF CONSENSUS**

The GST-ZnF fusion proteins were used to select members of the random oligonucleotide library that efficiently bound the protein. Those species recruited by the protein were retarded in their movement through a non-denaturing polyacrylamide gel. Each shifted complex was eluted from the gel and amplified by PCR to create a new protein specific library enriched in that fraction of the original 68 billion ( $4^{18}$ ) possible oligonucleotides that efficiently formed a protein-DNA complex. Representative gel shifts from the binding site selections are shown in **Figure 9**. This enrichment process was then repeated three additional times to produce a highly enriched library. DNA PAGE Gels purifying the enriched library after the first and fourth rounds are shown in **Figure 10**. The final enriched library for each protein was then PCR amplified, digested for the inbuilt restriction sites, and cloned into pUC18 plasmid vectors for sequencing. A collection of DNA page gel images of the released insert from restriction enzyme digests of the pUC18 plasmids from multiple independent clones of the first 5 selections are shown in **Figure 11**. The second group of selections utilized a modified protocol detailed

below. The same is shown for multiple members from the second round of BSS analysis under the improved protocol in **Figure 12 (full gel)** and **Figure 13 (brief)** respectively. The sequence data for CG8319, CG9797, and CG31365 contained sequences of low quality. Those sequences available for each did not support a consensus that matched our selection criteria. The consensus sequences for all other members are summarized in **Table 3**.

A degree of plasticity was expected in developing the binding site consensus. This is consistent with previously published work describing that an excess of ZFP protein will incorporate less ideal binding sequences at a modest rate (Choo et al., 1997). To verify the binding site selection results, I produced the oligonucleotides presented in **Table 4**. Four separate oligonucleotides were created for each of five proteins. Two of the complementary oligonucleotides were annealed to provide the wild type binding-site consensus (WT) and the remaining two were annealed to provide the mutant binding-site consensus (MUT). Both WT and MUT double stranded oligonucleotides were used in increasing concentrations to compete the binding of a Y-<sup>32</sup>P ATP labeled wild type binding site. Results for all the five consensus sequences show efficient competition with wild type as compared to mutant populations.

### **C. CONFIRMATION OF BINDING IN VIVO BY LUCIFERASE ASSAY.**

To test the activity of the ZAD effector domain and to show in vivo binding of ZAD proteins to both the consensus binding site and the putative target genes, I have developed a two-part assay system using both prokaryotic expression vectors and S2

drosophila cell culture cells (data not shown). This system uses a luciferase reporter under the control of either a low activity minimal fly promoter (Act5C-43) or a medium activity basic fly promoter (Act5C-361). Upstream of each promoter I insert either the consensus binding site in multimer form or one out of a selection of putative binding regions from the Drosophila genome. Drosophila S2 cells transfected with these plasmids will then be treated with a ZAD-TAT fusion protein. The TAT peptide will penetrate the cell membrane and concentrate the ZAD protein into the nucleus. This then allows for the binding of multiple ZAD-TAT fusion proteins whose regulatory effect can be observed when compared to treatments without ZAD addition. The substitution of the predicted binding sequence from a drosophila natural promoter in place of the multimerized consensus will confirm both binding and regulatory activity for that member. Our ZAD-TAT fusion protein will be produced in a prokaryotic expression system. I am producing proteins as both His6-ZAD-TAT and ZAD-TAT-His6 fusions to maximize the uptake into the cell and minimize the interference with the N-terminal ZAD effector domain. This work is currently ongoing.

#### **D. DOT BLOT ANALYSIS**

As an additional means of functionally testing the binding between the consensus sequence and GST-ZnF construct proteins, I developed a competition assay utilizing a dot blot type system. Construct proteins are bound with biotin labeled wild type oligonucleotides in the presence on increasing concentrations of unlabeled wild type and scrambled competitors. The complex is immobilized on a protein binding membrane and

visualized with a streptavidin-enzyme conjugate. Work with this radioisotope free method to confirm binding is continuing with the remaining ZAD family proteins.

## **2.3 RESULTS AND DISCUSSION:**

### **A. PUTATIVE TARGET GENES**

Sequence data for each protein were analyzed separately, with the consensus being derived from the longest region containing >50% of the sequences members and having >50% sequence homology over a sequence of at least 6 bp in length. Most of the consensus sequences were well above the minimum thresholds. This is of equivalent stringency as similar reported work in the ZAD family (Payre and Vincent, 1991). In that case, of the seven sequences found, they selected the four most similar and derived the consensus from each position matching in three of the four sequences. A representative consensus alignment and our binding site consensus for each of the five initial selections are shown in **Figures 14, 15, 16, 17, and 18**

Autoradiographs of the competitive EMSA results are shown in **Figure 19**. These results confirm that the sequence is recognized by the DNA binding domain and also that the nucleotides at each of those conserved positions is relevant to that binding activity. Similar confirmations are in progress for the remaining family members. The Autoradiograph of one such member whose consensus was derived from the modified BSS protocol is also shown in **Figure 20**. Dot blot analysis method has been used to

confirm the binding activity and specificity of two additional ZAD proteins shown in **Figure 21**.

With confirmed binding activity to the consensus sequences, I began a comprehensive bioinformatics search to identify potential target genes for each member. I examined several available databases on *Drosophila* genomes to locate potential *in vivo* targets for each ZAD family member. Sequences within one base pair of our consensus or matching a specifically selected sequence that appeared immediately upstream (<60 nucleotides) of the promoter or within regions previously reported as being transcription factor binding sites were considered as putative binding sites. This was done through a combination of the National Center for Biotechnology Information's BLAST, the Swiss Institute of Bioinformatics' Eukaryotic Promoter Database, and the publicly available data from BioBase's Transfac Database.

I identified 291 potential target genes from the ZAD members investigated. These genes were categorized according to their reported function. Of primary interest were those genes related to transcription and translation, neural and sensory genes and developmental genes. Most of the previously characterized ZAD members regulated genes involved in these three areas. Of the 291 putative targets, 222 could be grouped into these three categories. The remaining genes are sorted into the most commonly seen types; cell cycle, metabolism/molecule biosynthesis, membrane transport and unknown /other. Detailed breakdowns of the target gene functions for two members, CG18555 and CG7928 are shown in **Figure 22**. A set diagram of the collective results for all 23 ZAD proteins can be seen in **Figure 23**. Those target genes that have functions in Transcription/Translation (**A**), Neural and Sensory systems (**B**) and Development (**C**) and



also possess additional functions in Cell Cycle, Metabolism/Molecule biosynthesis, and membrane transport are broken down in **Figure 24**.

I further examined the data for related genes and those genes predicted to be the target of multiple ZAD proteins. The most significant cluster of results was found in the upstream regulators of Achate and Scute. These regulators include a series of Homeobox containing transcription factors that are active in the early development of the fly. This pathway already includes two of the reported ZAD targets: Serendipity Delta on Bicoid and *poils-au-dos* on Achaete/Scute (Gibert *et al.*, 2005; Payre *et al.*, 1994). Shown in **Table 5** is a listing of key members in and related to this pathway and the number of predicted targets for ZAD protein members. Nearly every member of the pathway contains at least one putative ZAD protein-binding site, with many having multiple sites from one or several different ZAD proteins.

Even if demonstrated *in vivo*, simply being the target of the ZAD proteins does not necessarily denote a redundancy of function. Overlap of expression patterns of ZAD proteins is critical. The result on the pathway as a whole is subject to very complex dynamics. The ZAD domain may not result in the same regulation on each gene. Competition by transcription factors that possess dominant transcriptional effects as well as the presence or absence of cofactors are the two possible means by which the same effector domain bound to a slightly different position could exert divergent control. However, the clustering of so many binding sites in a single pathway is consistent with what I would expect to see if the conserved genetic function model is accurate. Experiments to confirm the binding to and determine the nature of regulation are currently in progress.

## B. CORRELATION TO KNOWN ZAD PROPERTIES.

In this study, I have identified distinct binding sites for 23 of the 98 ZAD family members. In this selection were two proteins with previously reported *in vivo* binding sites determined by nuclease protection assays. I found the ideal binding sites to be more diverse than expected. The previously reported work showed Syr- $\delta$  and Syr- $\beta$  each binding a 13 base pair region with 10 positions conserved between the two Syr-ZAD genes. This would have been more indicative of the first model of duplicated molecular function. When comparing the binding site selected consensus sequences, I find that while each similar to their respective reported *in vivo* sites, they are less similar to each other. **Figure 25** illustrates the data comparisons. I observed similar results for the remaining ZAD consensus binding sites. The binding sites did not fall into distinct families of proteins binding similar target sites. While there was a degree of similarity in the binding sites of some members (CG10366/CG1792, CG7928/CG10267, CG7938/CG10321), they differed at some of the most conserved and therefore expectedly important positions. The proteins do not seem to group into distinct families with clearly overlapping binding characteristics. It is possible that other ZAD members able to substitute for each protein tested in this cohort are present in the remaining 70 ZAD members. However, the process used to select the 23 members used in this study was designed to yield the best chance for finding similar members by selecting those with the highest structural similarity.

The binding site selected consensus for CG11695 was identical to a known transcription factor binding site; 5'-CACRTG-3'. This sequence is the same consensus

reported for the E-box recognizing proteins and including the SNAG domain proteins, 5'-CANNTG-3' (Malik *et al.*, 1995) or 5'-CACRTG-3' (Desbarats *et al.*, 1996).

Bioinformatics work on the SNAG domain proteins Snail, Smuc, Slug, and Sct in mammals has shown a very well conserved DNA binding motif. This conserved region encompassed the second and third C<sub>2</sub>H<sub>2</sub> zinc fingers from Snail and the third and fourth C<sub>2</sub>H<sub>2</sub> zinc fingers from Slug, Smuc and Sct. This region is of an appropriate length to mediate the binding of six nucleotides (Desjarlais and Berg, 1992) and was identified in a bioinformatics study as the mostly likely region mediating the sequence specific DNA binding for each transcription factor (Unpublished data, Cindy Chiang). A Clustal analysis between this conserved region in the *Drosophila* Snail and Slug proteins and the zinc finger array from CG11695 shows a very high degree of homology with the fourth and fifth C<sub>2</sub>H<sub>2</sub> zinc finger domains of CG11695. This analysis utilized the web based clustalw tool available from European Bioinformatics Institute. This homology includes 4 of the 6 amino acid residues predicted in previous work to directly contact the DNA at positions -1, 3, and 6 relative to the start of the A-helix (Choo and Klug, 1994). A summary of the analysis of conserved zinc finger domains within the Snag family is shown in **Figure 26** while the comparison of *Drosophila* SNAG family members' putative DNA binding motif is compared to that of CG11695 in **Figure 27**.

Biogrid also reports a yeast-2-hybrid interaction between the ZAD proteins CG11695 and Grau. Grau is a known regulator of the *Drosophila* gene *cortex* with the identified binding sequence 5'-TCACTGTA-3' (Chen *et al.*, 2000; Harms *et al.*, 2000). Immediately upstream of the Grau binding site is a sequence within one base pair of two independent clones shown to bind CG11695. While this is supportive of a cooperative

regulation, future work will need to be conducted to identify any actual binding to these regions or direct gene regulation.

### **C. DNA BINDING ANALYSIS:**

The expansion of particular families of zinc finger transcription factors in various higher eukaryotes has been well described in the literature. Less understood is the reason why one family and effector domain is expanded in one lineage when a different family dominates in the next. Understanding these differences leading to this variation is complicated by the cryptic nature of the ZAD family members. While being a significant portion of the *Drosophila* regulatory apparatus and being expressed in the very important early developmental stages; mutagenesis studies have been unable to show phenotypes for the vast majority (>80%) of members (Drysdale et al., 2008). This facet of ZAD proteins has received a higher degree of speculation in the literature in recent years, with *in silico* studies strongly suggesting the particular evolutionary history of ZAD proteins has left them with enough overlapping function between members to mask many phenotypes (Chung *et al.*, 2007). Our study can for the first time directly address this theory experimentally.

Our results appear consistent with the current cryptic nature of many ZAD proteins. Multiple ZAD proteins are targeting either the same gene or closely related genes. The knockout or knockdown of one of those members still leaves other ZADs targeting members on the same pathway. More surprising is the nature of this overlap. I had expected the recent divergence to having resulted in more identical or overlapping

binding sites where the ZnF domains had not yet had time to change. I instead observed ZAD members possessing relatively divergent DNA binding sites targeting different regions near the promoter of a single gene or of members in the same pathway. This indicates to us a more positive selection to maintain the redundancy of function. I therefore postulate that there is a distinction in function between the members associated with that positive selection that is being masked by the current evolutionary state of the family. This may also indicate similar selective pressures are at work in other dipteran insects with similar lineage specific expansions in the ZAD family as well as the expansion of homologous families across the eukaryotic taxa. Future work knocking out multiple members in this cluster should help elucidate the specific functions. A reexamining of null mutants for possible changes to patterning and sensory bristle development would also be useful, as these are the developmental steps most closely associated with the specific Homeobox containing transcription factors most prevalent in our predicted targets.

Understanding the development of the ZAD family of zinc finger proteins will therefore provide insight into the evolutionary history and formation of lineage specific features far beyond *Drosophila*. It is likely other ZAD containing genomes have undergone a similar evolution, and may be better understood through the *Drosophila* model system. This is also true of other homologous families that have arisen in disparate species; including the KRAB proteins in humans, with their strong cancer and biomedical implications. Beyond the evolutionary insights, *D. melanogaster* is also one of the most utilized model organisms for genetic and molecular studies. The ease of growth, availability of powerful techniques, and relatively high incidence of homology with

human disease states contributes to this status. With this prevalence of use, filling in gaps in our current understanding takes on a special significance

#### **D. DEVELOPMENT OF AN IMPROVED BSS PROTOCOL**

*Protocol published in the Journal of Biomolecular Techniques.*

Through the course of this study, I found it necessary to develop a new protocol for selecting the consensus binding sites for transcription factors. The new protocol was as effective and powerful as the original radiolabeled BSS methods used in our lab but required significantly less radiation use. The overview of this technique is diagrammatically represented in **Figure 28**. As seen in the protocol, an enriched library obtained after four rounds of unlabeled selection using GSH affinity beads to immobilize each respective protein of interest. This pre-enriched library was the radiolabeled and an EMSA performed as in a standard protocol. By creating a new tailored library specifically enriched in sequences binding each protein, the number of labeled binding rounds could be significantly reduced. Our lab previously used between four and six labeled EMSA selections per protein and reports in the literature utilize as many as seventeen. This number was reduced to only two in the new protocol with no loss of selection power. A four hour exposure of that initial labeled selection for three ZAD proteins can be seen in **Figure 29**. Original pre-enriched and selected libraries were selected against a GST control and as expected no complexes were observed.

I found this modified technique is particularly well suited for adoption in labs that are currently using traditional radiolabeled binding site selection protocols. The

similarities in the pre-enrichment rounds to the standard EMSA enrichment allows for easy transitions. Affinity tagged fusion proteins and their matching affinity beads are already widely used in these studies as a means of making and purifying large quantities of the proteins and do not need to be purchased only for the pre-enrichment purposes. Eluted fractions from these cold bindings were PCR amplified using the same primers as conditions already required for the EMSA enrichment rounds. Our protocol requires no significant apparatuses or materials not already available in a lab equipped for isotopic BSS. It also requires no additional skills or training beyond those already employed in the preexisting methods. For essentially no cost in terms of funding, time, or training, a lab may transition from the traditional methods to this modified protocol and reduce the overall radiation usage.

All other alternative BSS methods replace the powerful selective function of the radioisotope. This may be done by replacing radioisotopes for visualizing the EMSA by the incorporation of fluorescent or affinity tags into the complex partners. This additional step is subject to the limitations of the process, including the inefficiency of related enzymes such as terminal deoxynucleotidyl transferase (TdT) and yields a product that may physically interfere with the complex formation. Other methods may also use an entirely different complex separation technique such as immobilization of the protein. This is similar to our pre-enrichment procedure but without the coupled isotopic rounds of selection it must be paired with a high throughput sequencing and computational analysis such as is reported in Reiss and Mobley. (Reiss, 2011). The cost of these systems put them beyond the reach of many labs. It is also possible to wholly remove the need for complex-free probe separation by performing each potential binding

independently on microarray such as is reviewed in Wang et al. (Wang 2011) This method is limited by the overall length of the sequence used in each reaction and requires technology and apparatus not necessarily available at all institutions.

This protocol has been successfully used in our lab to identify the consensus binding sequences for more than 20 additional ZAD family members and has been used by other colleagues in the laboratory to select sequences for other *Drosophila* and mammalian zinc finger transcription factors. Sequences selected in this manner interacted as strongly and specifically as sequences identified in our lab by the traditional all labeled BSS method. It was also possible to perform selections on nearly three times as many transcription factors per label order or to order approximately 67% less label for a given set of selections. Our label use efficiency increased by more than the 50% reduction in rounds of labeled binding. This is because the overall length of time to complete the BSS protocol was also reduced by 50%. This greatly reduced the loss of effective counts to decomposition and eliminated complications caused by kinasing with partially decomposed label. The advantages of reducing the total label required and the time in which the laboratory must house radioactive isotopes are significant; including decreased costs of label, decreased exposure times for personnel and fewer survey and storage requirements because the laboratory can clear of isotope sooner.



### 3. CHARACTERIZATION OF ZAD DOMAIN AND IDENTIFICATION OF CO-FACTORS.

#### 3.1 OVERVIEW OF COFACTOR ANALYSIS

To characterize the ZAD family of transcription factors and to examine the current theories as to their resistance to mutagenesis screenings, I began a screening to identify novel ZAD domain interacting protein partners. Previous bioinformatics work I performed found that the ZAD proteins did not possess the necessary molecular machinery to modify chromatin structure and exert regulatory control over gene expression. I therefore theorized that the ZAD family members- as is seen in the analogous systems in mammals- were recruiting a cofactor to act as a scaffold to assemble the necessary machinery for gene regulation. Previous ZAD protein interaction studies were limited to those few members previously well characterized in the literature and a high throughput yeast-2 hybrid screen using a large selection of known *Drosophila* gene products. I set out to use *in vitro* binding techniques to identify the ZAD interacting partners. From this work I have isolated two such partners, with the full characterization of each still ongoing.

## 3.2 METHODS UTILIZED

### A. PROTEIN BINDING

My initial assays used GST-tagged ZAD domains from five representative ZAD family members, CG12219, CG11695, CG9233, CG10108 (phyl), and CG2889. Early constructs including those used for binding are shown in **Figure 30**. The protein binding scheme is diagrammed in **Figure 31**. Each of these members contains a classical ZAD domain with E values ranging from  $8.1e^{-9}$  to  $2e^{-23}$  when compared to the consensus sequence using PFM tool. Each of the ZAD domains were amplified from clones purchased from Open BioSystems with the primers shown in **Table 6** and PCR products in **Figure 32**. Each domain was inserted in frame into a pGEX 4T-2 plasmid vector which contributed a GST affinity tag to the resulting construct protein. Multiple independent clones were produced and checked for the production of size-matched soluble protein under IPDG induction. A selection of expressing clones and GST control are shown in **Figure 33**. Clones were screened for members producing both the largest amount of induced protein and the largest fraction of that protein in a soluble form. Gels comparing the amount of protein in soluble and insoluble (inclusion body containing) fractions are shown in **Figure 34**. Proteins were produced in bulk under maxi prep conditions and dialyzed as previously described. The construct proteins were expressed in a soluble form and immobilized on GSH-affinity beads. Elutions obtained from these beads were analyzed on SDS-PAGE and are shown in **Figure 35**.

These constructs were used as baits in a pull-down assay to fish out interacting

proteins from a total soluble proteome extracted from the S2 fly cell line. I selected the S2 embryonic cell line because it highly expresses both the ZAD family members and many other proteins from the embryonic stages, which I anticipated would include any currently unidentified ZAD interacting proteins. Two controls were maintained, first a GST tag without a ZAD domain to account for any non-ZAD related binding, and secondly a binding without S2 cell extract to account for any contributions from the *E. coli* expression vector. Multiple replicates visualized by silver staining showed two potential protein partners, one at approximately 51kDa and a second at approximately 40kDa. Our initial selections proved difficult to analyze. The large concentrations of construct protein relative to all others resulted in weak visualization of non-construct protein. An example of these initial gels visualized by way of a standard silver staining protocol is shown in **Figure 36**. I began a series of modifications to improve the power of our assay.

## **B. VISUALIZATION**

Our first modification was to the staining protocol. Our initial method called for the application of 40 ml of staining solution containing 2 grams of silver nitrate for 15 minutes. By increasing both the relative concentration of silver to 3 grams per staining and increasing the exposure to 45 minutes I successfully visualized the non-bait proteins present in each sample as seen in **Figure 37**.

The improved silver staining protocol brought to light additional concerns with the pull down assay. Each sample contained a heavy background of proteins originating

in the prokaryotic expression vector and a relatively low concentration of proteins bound from the S2 cell lysate. In order to increase the amount of interacting proteins I switched from using a whole cell lysate to a nuclear extract from S2 cells. I also increased the number of cells used for each extraction, decreased the volume of each extraction, and increased the fraction of that extraction loaded onto the columns. Combined with a higher percentage gel and longer run time, increased sensitivity and specificity of the screens was achieved (**Figure 38**).

### **C. ELUTION METHOD**

Initially all samples were recovered by transferring the washed GSH-affinity beads from each binding into a sample of SDS containing loading buffer and denaturing the proteins at 95°C for 3 minutes. Samples were then hard spun at 14k RPM for 5 minutes and samples from the supernatant were run on SDS PAGE gels. In order to minimize the contribution of unrelated proteins I increased the concentration of BSA for blocking of non-specific binding and switched to recovering the proteins by elution with reduced glutathione. In combination these changes greatly improved the strength of signal from our binding assay. I was able to identify several promising regions that appeared contain proteins originating from the S2 nuclear extract and binding to the GST- ZAD construct as seen in **Figure 39**. By altering the SDS PAGE running conditions to a longer run under lower polyacrylamide percentage I were able to improve the resolution between 30 and 100 kDa. I was then able to positively identify two series of bands showing the expected characteristics. Each was present only in columns

containing GST- ZAD fusion proteins loaded with S2 nuclear extract. Controls only containing bait protein and not loaded with nuclear extract lacked the band, identifying it as originating in the S2 nuclear extract. Columns bound to GST protein only before S2 extract loading similarly did not contain the bands, confirming that neither the GSH affinity beads nor the GST portion of the fusion protein was sufficient to bind the unknown proteins. Loading controls comparing the relative abundance of construct and GST control proteins loaded into reactions shown in **Figure 40**.

#### **D. EXTRACT PREPARATION**

Our next step was to produce the interacting proteins in sufficient quantity to be subjected to sequence analysis. I began a series of protocols in further increase our signal until I could produce sufficient protein to be detectable via coomassie staining. Our main focus was on further concentrating the amount of protein loaded into the binding from the S2 nuclear material. The high abundance of bait from our prokaryotic expression system did not appear to be the limiting factor.

The total volume of S2 nuclear extract added to each binding reaction was limited by the salt conditions. Nuclear extraction buffer utilizes a fairly substantial concentration of NaCl and our binding conditions called for a salt concentration approximating that of the *in vivo* binding conditions between 100 and 130 mM NaCl. So any significant increase in extract addition would require the a more substantial increase in total volume or produce an unacceptably high salt concentration. I solved this issue through the use of size exclusion spin columns. This was possible because I knew the proteins of interest

were greater than the 10kD size discriminated through the column. Samples were spun to remove the nuclear extract buffer. The protein samples were then brought up to half their original volume in desalinated ICLB buffer. However, this method only marginally increased the amount of protein isolated in a given binding.

## **E. LABELED BINDING**

An additional avenue of investigation was used in the pull-down assay approach to cofactor identification efforts. The first utilized an  $^{35}\text{S}$  labeled proteome from *Drosophila* S2 cells. Cells were grown in a methionine/cysteine starvation media and then subsequently were supplemented with a mixture of  $^{35}\text{S}$  labeled methionine and cysteine. This labeled proteome was then selected against our ZAD-GST construct protein in a similar pull-down assay as to that previously described. The resultant products were separated on an SDS-PAGE gel. The Gel underwent flurography to convert the  $^{35}\text{S}$  signal into light to be recorded on an X-ray film. The resultant film can be seen in **Figure 41**. This method will not visualize any protein originating from the prokaryotic expression system, fully removing that source of background. The results are consistent with the initial pulldown assay with our two putative ZAD binding proteins at ~40 and 51kDa originating from the S2 nuclear extract.

I next used a diethylaminoethyl cellulose (DEAE) column to fractionate the S2 cell nuclear extract through ion exchange chromatography. I were able to duplicate the previous binding results with the fractions eluted under salt conditions between 100mM and 400mM NaCl with a peak at 200mM NaCl. While this again improved our signal

ratio it did not provide the orders of magnitude increase I required before sequence analysis. It was determined that a more powerful purification method such as HPLC would be required before a pull down assay of sufficient sensitivity could be made for this particular interaction. Initial DEAE fractionation and subsequent pull down assays have shown some promise, with efficient elution of a ZAD interacting protein in elution buffers containing between 200 and 300mM NaCl concentrations. Results of a binding assay performed with a DEAE fractionated S2 cell proteome is shown in **Figure 42**.

## **F. YEAST-2-HYBRID SCREENING**

I pursued another entirely independent approach to identify ZAD interacting partners. I used the Matchmaker Gold Yeast-Two Hybrid system available from Clontech. This system incorporates multiple redundant selection markers to greatly reduce the incidence of false positives seen in other yeast-two hybrid assays. I developed clones expressing baits of either ZAD domains or full length ZAD proteins in frame with the pGBKT7 contributed Gal4 DNA binding domain. Inserts were amplified from OBS clones of each respective ZAD protein CDNA visualized on an EtBr agarose gel. Construct plasmids were transitioned through DH5 $\alpha$  *E. coli* into (Bait) yeast cells. Initial selections against CG11695 have shown positive interactions against clones containing sequences from the genes of four *Drosophila* genes: calmodulin, myocyte enhancer factor 2, CG7053, and RPA-interacting protein alpha. Selections for other ZAD members have proven inconclusive thus far.

## 2.3 RESULTS AND DISCUSSION

It was our initial theory that the ZAD domain was responsible for recruiting a cofactor or series of cofactors responsible for exerting transcriptional control. This theory was supported by the current state of the literature and by comparisons to analogous systems in other eukaryotes. The ZAD domain had been characterized as a protein binding domain with no direct transcriptional activity. While ZAD members had individually been shown to act as transcription factors, no previous study predicted or identified the ZAD domain as containing any traditional chromatin modifying element. Nor did our full database of ZAD proteins, their sequences, expression, and predicted structures, show any consistent example of identifiable chromatin modifying activity in other regions of the proteins. Nearly all members contained only the typical DNA binding C2H2 zinc finger arrays and the protein binding ZAD domain.

While the molecular function of the ZAD domain for Syr and Grau were well characterized as homo-dimerization domains, I theorized this to be a deviation from its typical function. In both examples the ZAD domain interacted with a nearly c-terminal zinc finger domain to form head to tail homo-dimers. However, no other ZAD family protein has been shown to form a dimer in either direct molecular studies or yeast-2 hybrid screens reported by Biological General Repository for Interaction Datasets. Syr and Grau are also two of only nine *Drosophila* ZAD proteins with essential functions and the only two known to be activators at the molecular level. So while this specific activity seemed to be limited to these two members, it indicated to us that the ZAD domain is likely a protein binding structure. This was consistent with what is seen in the analogous



KRAB super family of proteins. There the protein binding N-terminal KRAB domain KRAB domain interacts with the Ring finger-B boxes-Coiled-Coil (RBCC) domain of the KRAB Associated Protein 1 (KAP-1). KAP-1 serves as a universal cofactor for the KRAB domain transcription factors. It functions as a molecular scaffold for recruiting a protein complex that coordinates the histone deacetylation, methylation, and heterochromatin protein 1 (HP1) deposition needed to silence the target gene (Ayyanathan et al., 2003).

Our results are consistent with the theory that the ZAD domain is recruiting a cofactor to actually exert its transcriptional regulatory activity. A broad selection of members are consistently interacting with a small population of uniquely sized proteins. Those proteins originate from and are present in *Drosophila* embryonic cells and their interaction withstands a very stringent 500mM NaCl concentration. However, without a purification method powerful enough to produce those partners in sufficient quantity for sequence analysis it is not possible to conclusively state that the proteins identified in our screens and the cofactors are one in the same. Future studies utilizing other concentration and purification methods such as high pressure liquid chromatography or utilizing an entirely different approach like Yeast-2 hybrid screening will be required for full characterization of these interactions.

As mentioned, previous yeast-2 hybrid screens have been performed on ZAD proteins with the data available through the Biogrid database. It is these studies that tell us of the self interacting nature of Syr and Grau. However, this study only included a large but limited selection of known *Drosophila* gene products that were tested in a pair wise fashion. This made for a more practical high throughput screen design but excluded

many possible interactions from the assay. Because all of the ZAD members were included, this study is significant for the identification of ZAD-ZAD homodimer and heterodimer formations. However, it is not able to identify any single or family of cofactors that could be responsible for the direct regulation of gene expression. My yeast-2 hybrid screen utilized a library of prey plasmids constructed with a normalized *Drosophila* cDNA library. The plasmid library was purchased from Clontech. My positive interacting proteins appear consistent with the known functions of CG11695 as it interacts with Grau to possibly control in the early embryonic development of *Drosophila* (Giot et al., 2003). However, neither my screen nor the earlier works have allowed for the full characterization of the necessary cofactors of the ZAD domain.

Our selection, while it includes other potential partners not contained in previous studies, the results are complicated by the incorporation of normally un-translated regions of the mRNA into the plasmid library. Future studies utilizing yeast-2 hybrid assays to determine cofactors of the ZAD domain must either include many more known and predicted gene products from *Drosophila* or utilize a cDNA derived prey library in a true high throughput methodology.

Through the course of my investigations into the ZAD proteins, it became evident that attempts to categorize the entire family would require a database of information on all members. Using only select members that were well characterized in the literature to draw any conclusions was insufficient. This was because the best characterized members were in fact often the least typical in form and function. I therefore developed just such a database that included all 98 known ZAD members and all of their individual isoforms. The database included known nucleotide and amino acid sequences, the ZAD domain

regions, zinc finger arrays, identifiable protein motifs, mutant allele phenotypes, alternative splicing, known genetic and protein interactions, restriction maps for cloning and all other relevant data. It was through this database (available upon request) that I was able to show a lack of consistent DNA binding motifs outside of the zinc finger arrays, a total lack of chromatin modifying structures and select the classical archetypal members for my research. Each of these tasks was a necessary prerequisite to this and future studies of the ZAD family.

## 4. MATERIALS AND METHODS

### 4.1 MATERIALS

Materials for binding site selection include but are not limited to ZAD-ZFP full length (Open BioSystems), primers and oligonucleotides (Integrated DNA Technologies), all restriction enzymes and their buffers as well as Quick Ligase Buffer and ligase (New England Biolabs), and BL21 and DH5 $\alpha$  competent *E. coli* cells (Invitrogen). PCR reactions were done using Taq Bead Hot Start Polymerase (Promega), DMSO, and mineral oil (Sigma-Aldrich). Qiagen Gene Clean Kit was used to extract DNA. Various reagents were purchased from Fisher BioReagents (ampicillin, kanamycin, urea, PCI/CI, proteinase inhibitors), EMD (lysozyme, glycerol), and Sigma-Aldrich (PMSF, SDS). Anti-myc and anti-mouse IgG antibodies were from Promega. Clontech Matchmaker Gold kit supplied all materials for transformation (TE buffer, lithium acetate, Yeastmaker Carrier DNA, competent cells), yeast two-hybrid assay including cloning vectors and yeast strains, yeast media and *Drosophila* Normalized Mate & Plate prey library.

## 4.2 METHODS

### A. PROTEIN CONSTRUCTS

Protein constructs expressing the ZAD and ZnF domains of each ZAD family member were built using products PCR amplified from *Drosophila* cDNA clones purchased from Open BioSystems Inc. (Huntsville, AL) or from 0-4, 4-8, 0-8, and 0-12 hour *Drosophila* cDNA libraries utilizing the primers in **Table 1** (ZnF) and **Table 6** (ZAD). ZAD constructs were created for protein interaction studies and ZnF constructs for DNA binding site selections. Each product was purified on an agarose gel, sequentially digested for the endonucleases restriction sites built into each primer, and directionally ligated into a similarly digested pGEX 4T-2 or pGEX 4T-1 plasmid vectors. Ligated GST-domain fusion vectors were transitioned through *E. coli* DH5 $\alpha$  to produce sufficient quantities of supercoiled plasmid for transformation into *E. coli* BL21 competent cells for protein expression.

### B. PROTEIN EXPRESSION AND PURIFICATION

Each independent clone was then cultured for mini-plasmid DNA preps. The plasmids were then checked for the correct insert size by restriction endonuclease digestion and gel electrophoresis. Confirmed recombinant plasmids from each pGEX-ZAD construction were then transformed into BL21 *E.coli* host for protein expression. Transformed BL21 cells were cultured in LB/Amp/Kan media and tested via IPTG (1 mM) induction at 37<sup>0</sup>C for protein production. A non-insert bearing pGEX-4T2 plasmid

transformed culture was used as a control to check for the proper size protein production. Two expression clones were then selected from each construction for maxi-protein production by inducing a 500 ml culture with 0.1 mM IPTG at 30<sup>0</sup>C. Expressed proteins were released from the cells by lysozyme treatment followed by sonic disruption. The soluble fractions of each protein, predicted to represent the functional form, were then bound on a GSH bead column, eluted in a 15mM reduced glutathione containing Tris-buffered elution buffer and were later dialyzed to remove the reduced glutathione. Dialysis was conducted in three rounds using Spectra/por membranes (5kDa cutoff) with two six hour rounds in 1L of .1mM PMSF containing PBS and one twelve hour round in 1 L of 10% glycerol and .1mM PMSF containing PBS.

### **C. DNA BINDING SITE SELECTION**

Initial binding site selection experiments were conducted with GST-ZnF proteins for CG11695, CG12219, CG30020, CG7938, CG17958, (amplified from clones purchased from Open BioSystems) and a GST control as described (Peng et al., 2002). Each protein was combined with a <sup>32</sup>P-ATP end-labeled 49mer oligonucleotide library. The library consisted of oligonucleotides of the species 5'-agacGGATCCattgca-NNNNNNNNNNNNNNNNNNNN-ctgtccGAATTCgga-3'; each member contained a random 18-N central region that was flanked by known primer targets with imbedded BamHI and EcoRI restriction sites (Restriction sites are underlined). The protein-DNA binding was conducted in Nuclear Extract Binding Buffer (20 mM HEPES, 75 mM NaCl, 0.5 mM DTT, 10% glycerol, 0.5 mM MgCl<sub>2</sub>, and 50 μM ZnSO<sub>4</sub>) (NEBB) and ran on a

5% poly-acrylamide gel for electrophoretic mobility shift assay (EMSA). The GST-ZnF-oligonucleotide complexes were electro-eluted from the gel and amplified by PCR using the known 18mer flanking sequence primers. The resulting products were run on a 10% PAGE gel with marker and stained in a solution of 0.0125% EtBr. The PCR products were cleaned and prepared for use by proteinase-K treatment, phenyl-chloroform-isoamyl alcohol extraction, chloroform-isoamyl extraction followed by ethanol precipitation. These enriched libraries were then used in the second round of mobility shift assays with each of the purified GST-ZnF proteins. Original unenriched library was used against the GST control and no complex was observed. This process was repeated for a total of four rounds of enrichment. The products obtained from the final enrichment were amplified by PCR, and digested with EcoRI and BamHI restriction enzymes present flanking the known 18-mer ends of each oligonucleotide. These digested products were then ligated into pUC18 plasmid vector for cloning and transformed into DH5 $\alpha$  cells. Multiple independent clones were produced from each GST-ZnF binding. Mini-plasmid preps were conducted using each clone. The plasmid DNA was checked for the presence of an insert by way of enzyme digestion. The resulting 15-18 positive clones for each construct were then sent to ICBR Genomics Core (Univ. of Florida, Gainesville) for cycle sequencing in a 96-well format.

The additional 23 ZnF clones amplified from CDNA libraries were selected using a hybrid cold and hot binding technique. Proteins were bound in 1.5ml microcentrifuge tubes containing 10ul of GSH beads supplemented with 10ul of G75 sepharose beads to increase the volume for rinses. The unbound proteins were removed with three 800ul washes of PBS wash buffer (137 mM NaCl, 2.7 mM KCl, 10 mM sodium phosphate

dibasic, 2 mM potassium phosphate monobasic, 1mM PMSF, 0.5% BSA, 0.5mM DTT). Samples were rotated at 25<sup>0</sup>C for 2 minutes, centrifuged at 2,000 RPM for 5 minutes, and the supernatants were removed by suction pump. The samples were then rotated at 4<sup>0</sup>C for at least 30 minutes in PBS wash buffer to block non-specific binding. The previously described annealed random 49mer oligonucleotide library was then bound to the immobilized protein in 1x NEBB wash buffer [20 mM HEPES, 75 mM NaCl, 0.5 mM DTT, 10% glycerol, 0.5 mM MgCl<sub>2</sub>, and 50 μM ZnSO<sub>4</sub> 1mM PMSF, 0.5% BSA, 0.5mM DTT) rotated for 30 minutes at 4<sup>0</sup>C followed by 30 minutes at 25<sup>0</sup>C. DNA-Protein complexes were eluted from the beads in 20ul of 15mM reduced glutathione containing Tris elution buffer. Oligonucleotides in the elutions were than amplified by the known primer targets to create a library enriched in the sequences that effectively bind the protein construct. This enriched library was used in a second round of enrichment, with the products then used for a third and fourth round. The fourth round library was then taken for two more rounds of labeled binding site selection as described above.

#### **D. PROTEIN EXTRACT**

*Drosophila* Schneider line 2 (S2) cells were grown at 25<sup>0</sup>C in Shields and Sang M3 medium (Sigma) containing 10% insect medium supplement (Sigma), 2% fetal bovine serum and 1x penicillin and streptomycin. Cells were then transferred to a methionine and cysteine deficient media for 20 min for starvation. Cells were then supplemented with <sup>35</sup>S labeled methionine and cysteine for 2 hours to label the fly cell proteome. Cells were then lysed in Insect Cell Lysis Buffer (Tris 10 mmol/L pH 7.5,



NaCl 130 mmol/L, Triton X-100 1 %) for 1 hour at 4<sup>0</sup>C. The lysate was then fractionated with a 14,000 rpm spin for 30 minutes, and total soluble protein content was then collected with the supernatant. In parallel, bacterial cell extracts were prepared from the five GST-ZAD-constructs and the GST control as follows: fifty-milliliter aliquots from the 500 ml LB/Amp/Kan/IPTG maxi-inductions were spun down. The cells were resuspended in 4ml of PBS buffer with lysozyme and incubated for 30 minutes at 4<sup>0</sup>C. Phenylmethylsulfonyl fluoride (PMSF) was added to 1 mM concentration, and the samples were incubated for an additional 30 min at 4<sup>0</sup>C. The lysed cell suspensions were disrupted by sonication (8 cycles, at 50% duty cycle, each cycle was 1 minute pulse and 1 minute cooling). The sonicated samples were spun down. The supernatants were collected and passed through a 0.45 m syringe filter. Finally the GST-pull down assays were performed as described (Ryan et al., 1999) and is briefly mentioned below: 10ul of 50% GSH sepharose affinity bead slurry was bound with 600ul of either filtered GST-ZAD protein or filtered GST control protein extract from the preparations mentioned above. Nonspecific binding sites were blocked with 0.5% bovine serum albumen (BSA), and the bound GST and GST-ZAD proteins were associated with the fly cell lysate prepared as above for 1 h at 25<sup>0</sup>C. The beads were then washed five times in 1ml volumes of binding buffers, each containing increasing concentrations of NaCl (100 mM, 250 mM, and 500 mM). This step was carried out to remove any loosely bound proteins from the beads and to assess the strength of interaction between the ZAD domain and any hitherto unknown ZAD-domain interacting protein. All proteins retained on the beads were then released by heating the beads in sodium dodecyl sulfate(SDS)-sample buffer at 95<sup>0</sup>C for 3 minutes. The samples were spun down to pellet the beads in order to make for

an efficient supernatant removal. The soluble proteins were then electrophoresed on SDS-Polyacrylamide gels.

## **E. FLUOROGRAPHY**

Gels from Protein interaction assays were subjected to fluorography to intensify the <sup>35</sup>S signal and better visualize the construct binding proteins. Gels were dehydrated by rocking in 100% DMSO for one hour. The gels were then impregnated with 22% PPO (2,5-Diphenyloxazole) in DMSO by rocking for one hour. The DMSO was removed and the PPO within the gel was precipitated by soaking in water followed by 30 minutes of continuous water washes to remove excess PPO. Gels were then exposed to x-ray film for autoradiography.

This assay was also performed with unlabeled S2 lysate and visualized by silver staining as a preliminary step. Gels were fixed successively in solution A (50% methanol; 10% acetic acid) for one hour and in solution B (10% methanol; 7% acetic acid) for over night. The gels were then treated in 100 ml of 10% glutaraldehyde solution, washed six times in distilled water, and incubated with 150 ml of substrate solution containing 0.15N NaOH, 2.5% NH<sub>4</sub>OH and 2% of silver nitrate for 30 min. The gels were thoroughly rinsed in ddH<sub>2</sub>O three times for 15 minutes each and developed in 250 ml solution containing 0.002% formaldehyde, 0.005% citric acid. Developed gels were fixed in 50% methanol; 10% acetic acid and then preserved in 50% ethanol; 20% glycerol and finally dried on a water permeable cellophane membrane for long-term storage.

## **F. DEAE BINDING**

Two milliliters of DEAE sepharose beads were packed into columns and washed with PBS buffer. The beads were then rinsed in a salt free Tris buffer (50mM Tris, 0.1% TX-100, 1mM PMSF). Proteins were diluted in either the same salt free Tris buffer or in ICLB lacking NaCl to bring the total salt concentration to less than 50mM. The labeled S2 protein extracts were then bound to the column. Columns were then washed in salt free Tris buffer and collected. Proteins were then sequentially eluted from the column in Tris buffer solutions containing 50mM, 100mM, 200mM, 300mM, 400mM and 500mM salt concentrations (50mM Tris, 0.1% TX-100, 1mM PMSF). All the elutions were brought to 100mM NaCl and used in a protein binding pull down assay as described earlier.

## **G. CLONING AND SEQUENCING OF PGBKT7-ZAD-ZFP BAIT**

The DNA of the constructs to be transformed into the yeast vector pGBKT7 was PCR amplified, along with the DNA of the cloning and control vectors supplied in the Clontech Matchmaker Gold kit. pGBKT7-DNA Binding Domain and pGBKT7-Activation Domain are cloning vectors, and pGBKT7-53, pGADT7-T, and pGBKT7-Lam are control vectors.

The DNA of these was gene cleaned using a Qiagen kit, and digested with *EcoRI* and *BamHI* to release the insert of the desired construct. These were gene cleaned again and ligated into pGBKT7 vector, transformed into competent *E. coli* DH5 $\alpha$  cells,

recovered for one hour in 900 $\mu$ L of LB, and 200 $\mu$ L was plated onto LB/kan plates and incubated at 37°C overnight. Colonies were mini plasmid prepared and digested to check for the proper inserts. These digests were run on a 1% agarose gel, and for Slug and Scratch domains, a 10% DNA polyacrylamide gel. At least three unique clones for each construct were obtained.

The DNA samples that showed a positive recombinant clone were subjected to RNase cleaning, and PCI/CI purified. Positive clones were midi prepared, RNase and PCI/CI cleaned, and spotted into a 96-well microtiter plate for sequencing by the ICBR Genomics Core at UF.

## **H. YEAST TRANSFORMATION**

Competent yeast cells were prepared using the Yeast Two-Hybrid Matchmaker Gold yeast as stated in the protocol using TE buffer and lithium acetate. The plasmid DNA was transformed into the cells with the supplied Yeastmaker Carrier DNA. All steps were followed as stated in the yeast transformation protocol, and cells were plated on SD/-Trp to determine transformation efficiency.

## **I. YEAST PROTEIN EXPRESSION**

Transformed yeast glycerol stocks were grown in SD/-Trp at 30°C, 225rpm for about 40 hours, and cultures were poured into 15mL conical tubes filled halfway with ice to chill the cells then centrifuged at ~1400rpm for five minutes. The supernatant was

decanted, an additional 5mL of ice water resuspended the pellet, and the culture was centrifuged again under the same conditions. The pellet was processed for protein extraction using cracking buffer stock solution (8M urea, 5% w/v SDS, 40mM Tris-HCl pH 6.8, 0.1mM EDTA, 0.4 mg/mL bromophenol blue) which was used to make a prewarmed 60°C cracking buffer (0.1M DTT, ~4.4x PMSF, aprotinin (0.37mg/mL), leupeptin (0.03mM), and pepstatin (0.1 mg/mL)). Samples were heated to 70°C for ten minutes, vortexed vigorously for one minute, and placed on ice for one minute, for a total of ten cycles, adding PMSF and proteinase inhibitor cocktail every other cycle. Samples were centrifuged at 14,000 rpm for five minutes to pellet debris and unbroken cells while supernatants were combined and transferred to new tubes.

The supernatant was boiled for three minutes and run on a 12% SDS-polyacrylamide gel to ensure the transformed yeast cells were producing proteins. The completed gel was run for 4 hours at 250mA, and the membrane was rinsed in blocking solution (1x PBS, 0.2% Tween, 5% nonfat milk powder) for one hour. Membrane was then washed in a rinsing solution (1xPBS and 0.2% Tween) for five minutes. A 1:3333 myc antibody in PBS, 5% BSA, and 1x PBS primary antibody solution coated the membrane for one hour, then four washes in the rinsing solution were done. The secondary antibody (1:10,000 antimouse conjugate, 1x PBS, 0.2% Tween20, and 1% nonfat dry milk) was followed by four additional rinses. The membrane was then soaked in 0.1M Tris Cl, pH 9.5 before a substrate (0.1M Tris Cl, pH 9.5, NBT, and BCIP) was added. Membrane was incubated until bands developed and kept in water with 20mM EDTA to save.

## **J. AUTOACTIVATION AND TOXICITY ASSAYS AND YEAST MATING**

Each of the glycerol stocks made as a result of a transformation was plated on selective plates to ensure the bait did not autonomously activate the reporter genes without a prey protein.

pGBKT7-bait was grown in 3mL of 2x YPDA at 225rpm and 30°C for 24 hours, and 10µL was transferred into 25mL SD/-Trp and grown to an O.D.600 of ~0.8. This culture was spun down at 1000rpm for 5 min, resuspended in 2mL 2x YPDA and added to a 1L flask containing 25mL of 2x YPDA/kan (50µg/mL) and 1mL of a normalized human cDNA library. The mixture was incubated for 20 hours at 30°C at 40rpm. The mating was centrifuged at 1000rpm for 10 minutes, decanted, rinsed in 25mL 0.5x YPDA to resuspend the pellet, centrifuged at 1000rpm for 5minutes, decanted, and finally, the pellet was resuspended in 5mL of 0.5x YPDA with 50µg/mL of kan. Ten DDO/-Trp/-Leu/X-α-gal/A plates were plated with 200µL of the mated mixture and spread with glass beads. The plates were incubated at 30°C for about three days. Colonies that were positive under all conditions were saved in glycerol stocks for later use.

## **K. YEAST PLASMID PREPARATION**

Glycerol stocks were used to grow midi-amount cultures in 25mL YPDA broth at 30°C, 225rpm. Cultures were centrifuged for 5 minutes at 7000rpm, the supernatant was decanted, and then 1 mL of lysis buffer (2% Triton X100, 1% SDS, 100mM NaCl, 1mM Na<sub>2</sub>EDTA) was added to the cell pellets to resuspend them in. 0.2g of acid washed glass

beads and PCI were added to the samples, and the tubes were vortexed for 2 minutes. Samples were PCI/CI purified as well as digested with RNase for 1 hour at 37°C. Final samples were again PCI/CI purified and dissolved in water to minimize salt content.

Electrocompetent DH5 $\alpha$  cells were used to electroporate plasmids into. To make electrocompetent cells for transformation, DH5 $\alpha$  streaked on an LB plate was grown overnight at 37°C. One colony from the plate was inoculated into 5mL of SOB medium and grown to an O.D. between 0.5 and 1.0. The culture was chilled on ice for 15 minutes then centrifuged at 4960rpm for 15 minutes. The cell pellet was resuspended in ice-cold water, centrifuged for another 15 minutes, and this was repeated. The cell pellet was next resuspended in 20mL ice-cold water with 10% glycerol, centrifuged at 4960rpm, and repeated, and finally resuspended in 3mL volume in 10% glycerol. Aliquots were taken and frozen at -80°C immediately.

Plasmids were diluted 1/10 from the midi preps (to further reduce salt concentration which hinders electroporation). From this, a 40:1 ratio of cells to DNA was chilled and combined in an electroporation cuvette (0.2 cm gap). Mixtures were pulsed in the Bio-Rad Gene Pulser II and Pulse Controller Plus apparatus (25 $\mu$ F, 2.5kV, and 200  $\Omega$ ). Cells were recovered in SOC media for 1 hour, at 225 rpm, at 37°C and then plated on SOC/ampicillin overnight at 37°C.

Colonies were then mini plasmid prepared, and to ensure that a plasmid was inserted into the cells, restriction endonuclease digestion with *NdeI* and *EcoRI*. After one hour in a 37C water bath, samples were run on a 0.7% agarose gel.

## **L. DNA SEQUENCING**

Positive clone mini preps were RNase purified then sent to the ICBR Genomics Core at the University of Florida (Gainesville) and sequenced in duplicate. Each clone was sequenced with the T7 promoter 5'-TAATACGACTCACTATAGGG-3' and 3' pGADT7 Activation Domain sequencing primer 5'-AGATGGTGCACGATGCACAG-3'. Sequences were then analyzed and genes were searched possessing those sequences.

## **M. DOT BLOT ANALYSIS**

GST-ZnF construct proteins were bound to biotinylated wild type consensus oligonucleotides in the presence of unlabeled wild type or scrambled oligonucleotides. The binding was conducted in 1x NEBB-NaCl and .25x PBS with 3% BSA for 10 minutes at room temperature and 20 minutes at 4<sup>0</sup>C. The full binding was then immobilized on a Milipore Immobilon-p membrane that had been saturated with 1x PBS + 1% Tween-20. After immobilization, the membrane was UV crosslinked and washed three times for 20 minutes in 1x PBS + 3% BSA.

Bio-labeling was conducted with the Biotin 3' End DNA Labeling Kit from Thermo Scientific. For visualization with horse radish peroxidase I utilized the Chemiluminescent Nucleic Acid Detection Module from Themro Scientific with the following modifications- blocking extended to 60 minutes, incubation with conjugate increased to 60 minutes and dilution of conjugate reduced from 1:400 to 1:1000. For visualization with alkaline phosphatase the membrane was incubated for 16 hours in 1x

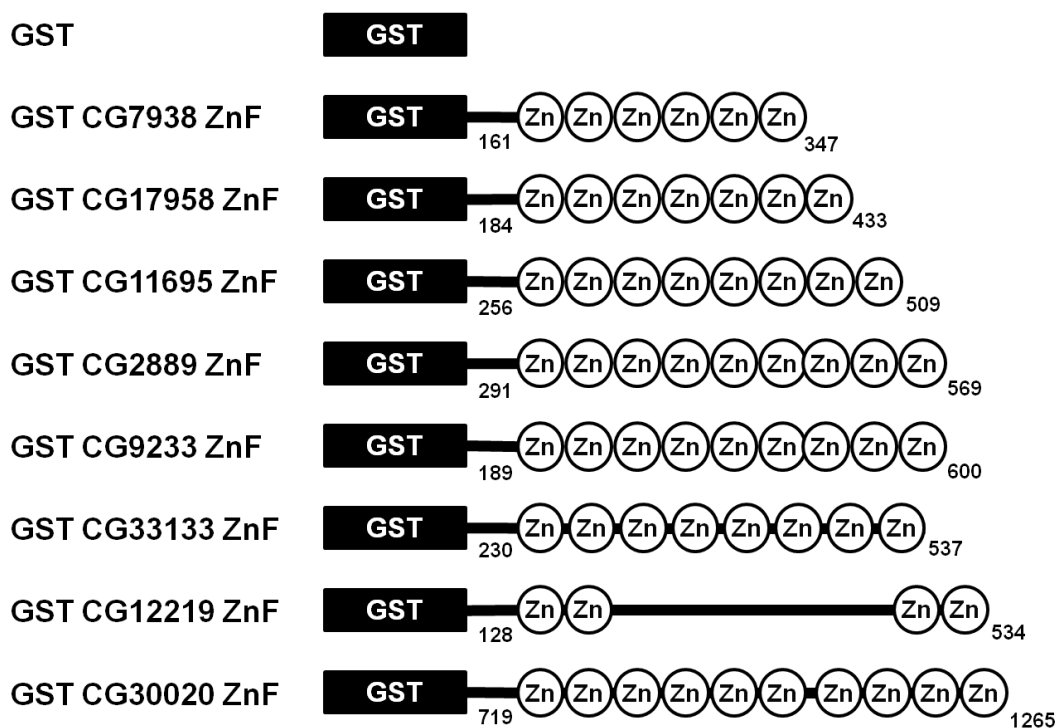


PBS + 3% BSA with a 1:5000 dilution Streptavidin conjugate, washed twice in 1x TTBS for 20 minutes, once in TBS for 20 minutes, equilibrated in 100mM tris 9.5 and incubated in a nitro blue tetrazolium and 5-Bromo-4-chloro-3-indolyl phosphate solution in 100mM tris 9.5.

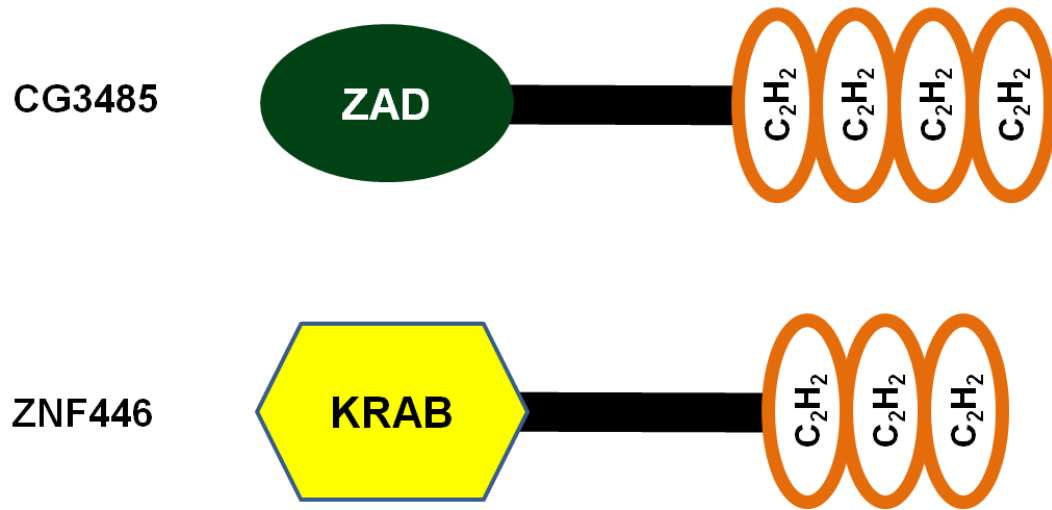
## 5. FIGURES AND TABLES

**Figure 1. ZnF Domains identified from full length templates.** **A.** DNA segments corresponding to the indicated amino acids were PCR amplified by using the full-length ZAD genes and fused in frame with the Glutathione-S-transferase tag to generate the respective recombinant fusion proteins. **B.** A diagrammatic representation of the natural protein of one ZAD member (CG3485) and one KRAB member (ZNF446).

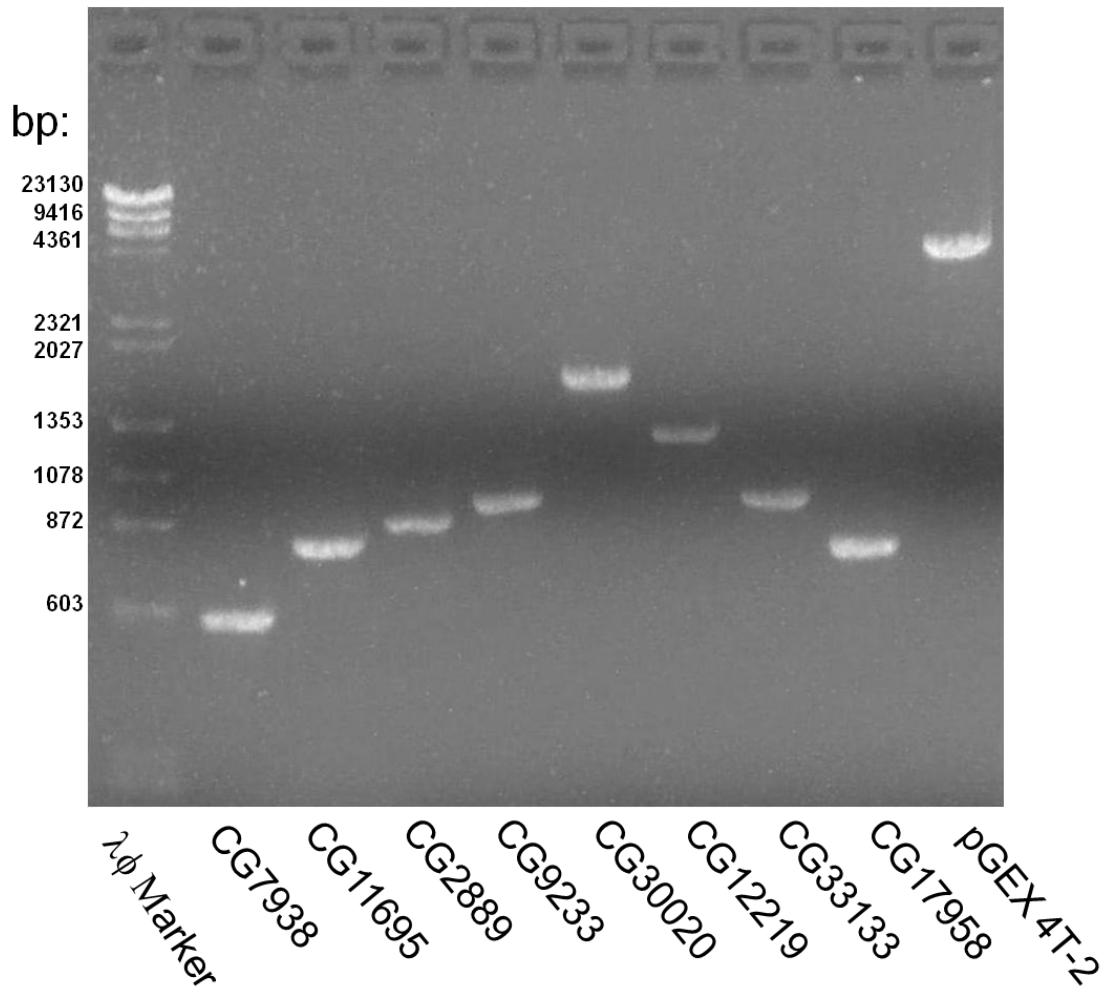
**A.**



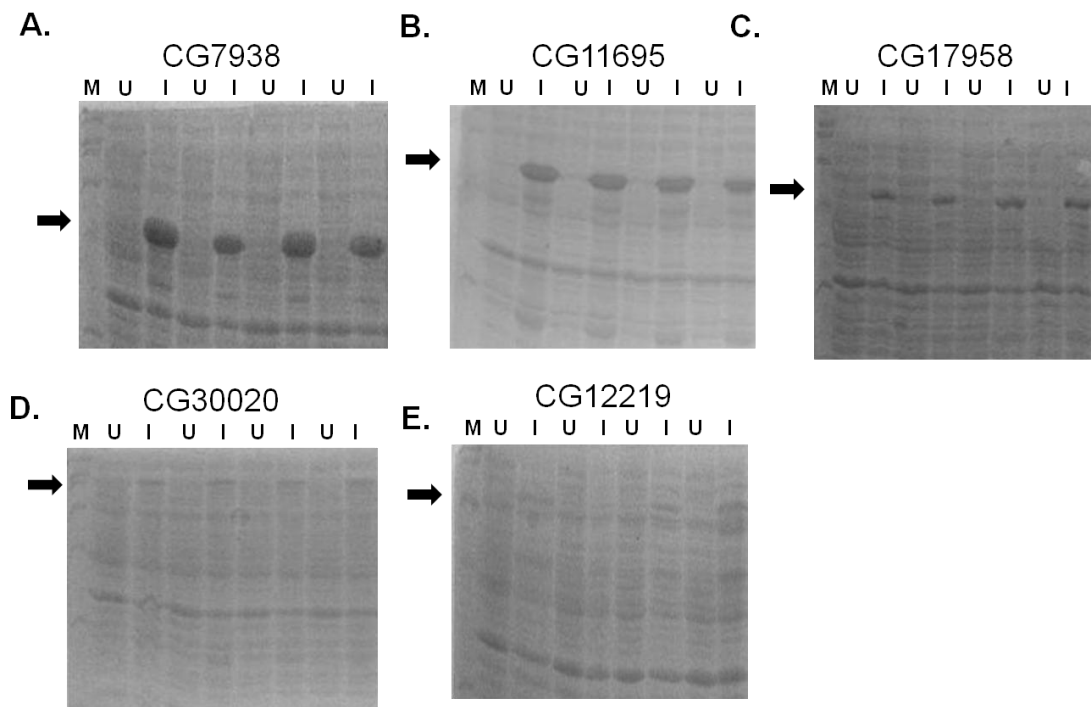
B.



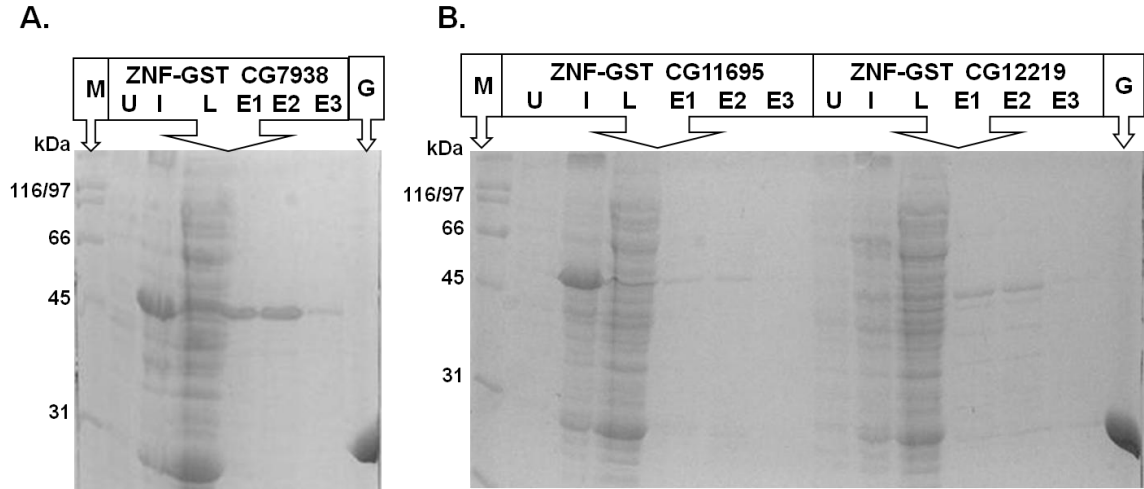
**Figure 2. ZnF Domains amplified and digested.** Zinc finger arrays from each of nine ZAD proteins. Domains were amplified from the cDNA samples purchased from Open Biosystems with the primers in **Table A2**. Each ZnF product was digested, as was a complementary pGEX plasmid vector. CG10108 and CG11371 are not present because these ZAD proteins lack the tandem zinc finger arrays predicted to act in DNA binding.



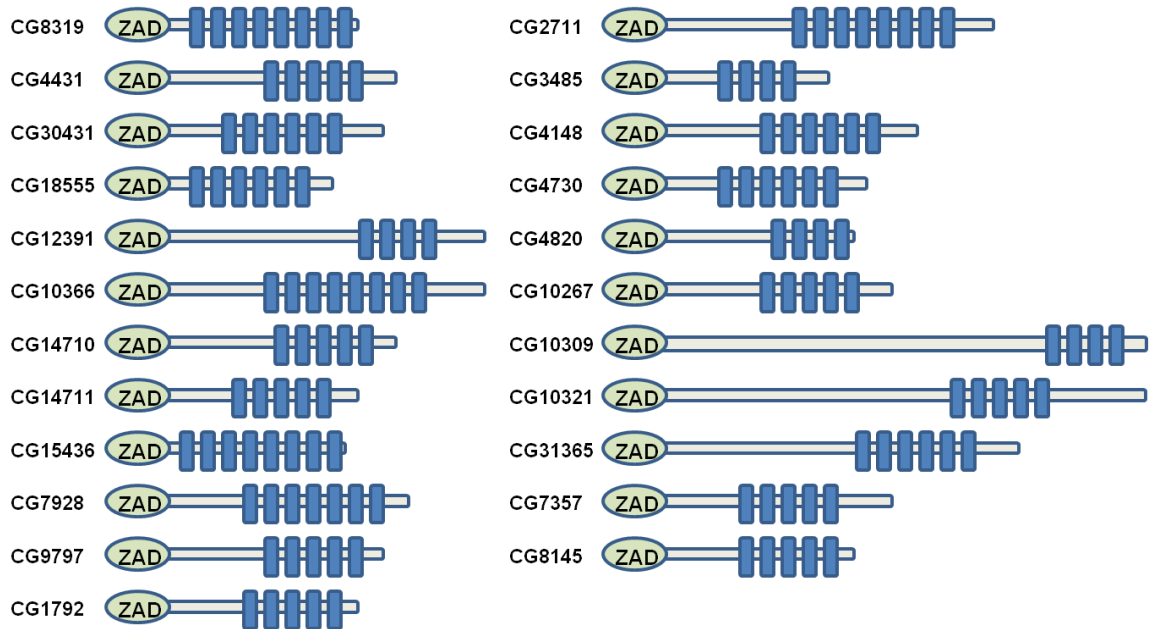
**Figure 3. Protein mini-induction experiment using positive GST-ZnF recombinant clones.** Induced and uninduced samples from independent clones of each GST-ZNF construct were processed and run on 12% SDS PAGE gels. Shown here are the inductions for four clones from CG7938 (A), CG11695 (B), CG17958 (C), CG30020 (D), and CG12219 (E). Each gel also contains broad range marker (M), and the expected size band is indicated by an arrow.



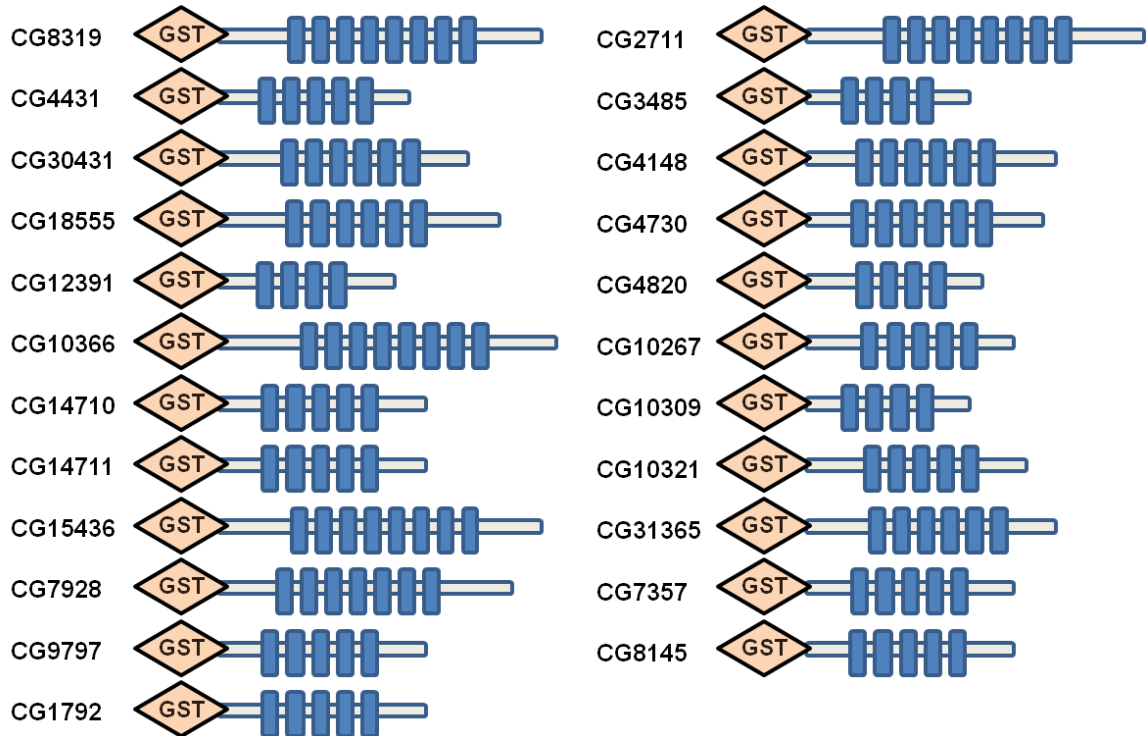
**Figure 4. GST-ZNF construction protein inductions were purified by GSH bead columns.** Panels containing the resulting elutions were run on SDS-PAGE gels and stained with coomassie blue. Each panel shows protein extract from uninduced(U) and induced(I) cell cultures, a sample of the raw cell extract(L) and aliquots from each of the three column elutions(1,2,3). Broad range marker(M) and purified GST protein(G) were also included. Shown here are the panels for CG7938 (A) and CG11695 and CG12219 (B).



**Figure 5. Diagrammatic representation of each of the archetypal ZAD family members used in the second round of selections.** Drawn to scale of amino-acid length in 5' to 3' orientation.



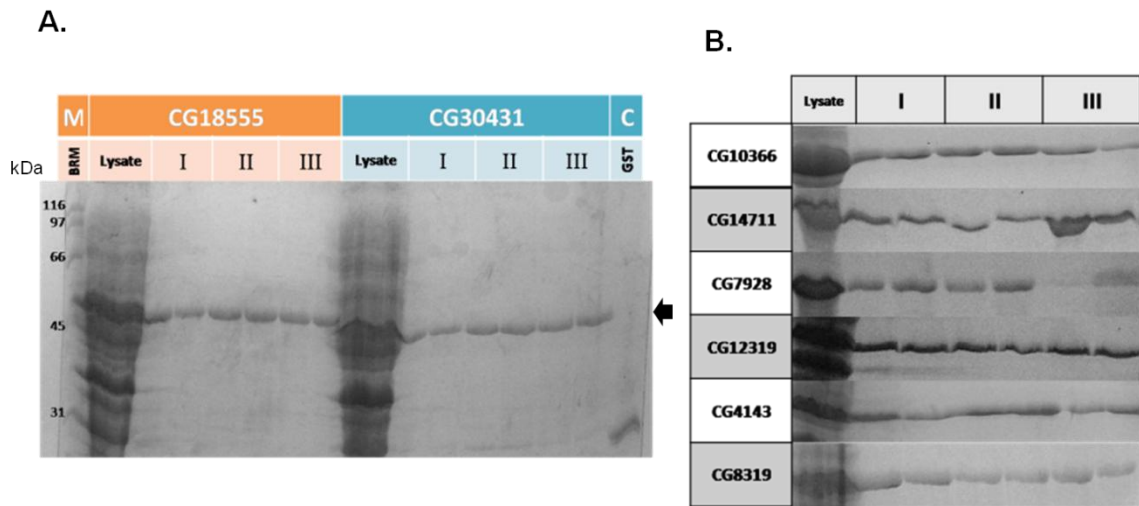
**Figure 6. Diagrammatic representation of each of the second set of GST-ZnF constructs.** Each construct consisted of the pGEX contributed GST affinity tag, a 15 amino acid linker region, and the full C2H2 zinc finger array from each of the selected classical ZAD family members.



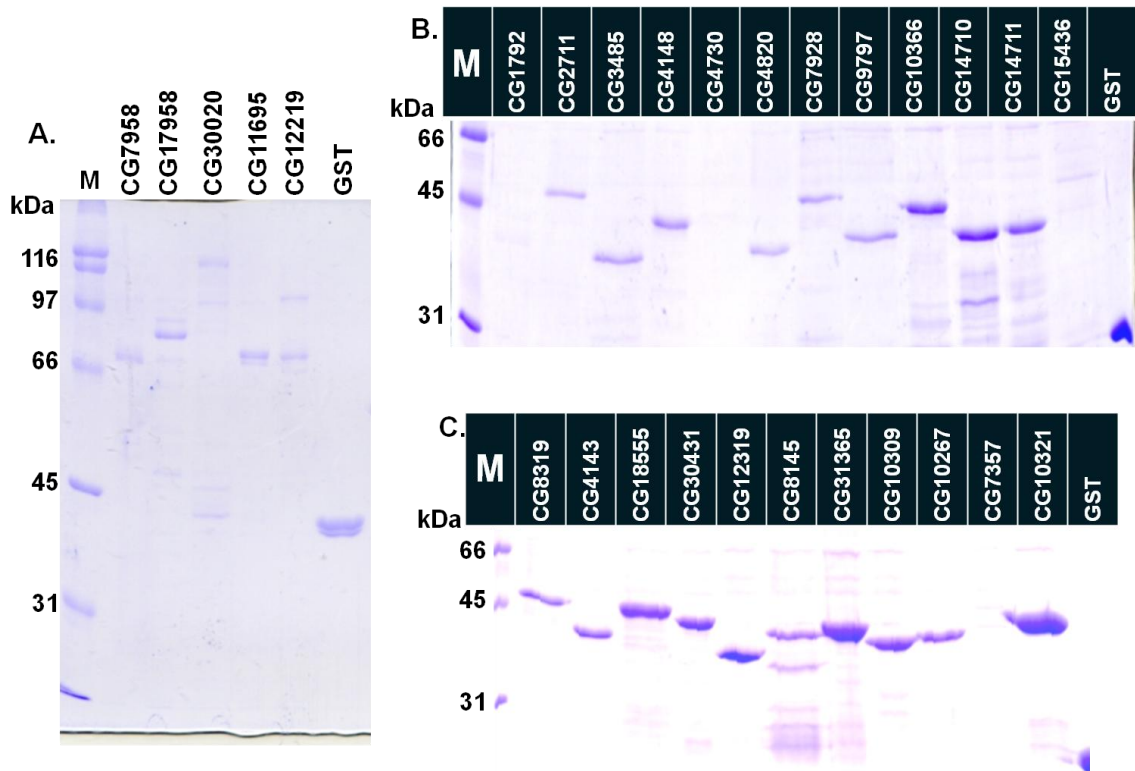


**Figure 7. Purifications of GST-Zfp proteins on GSH affinity bead columns.**

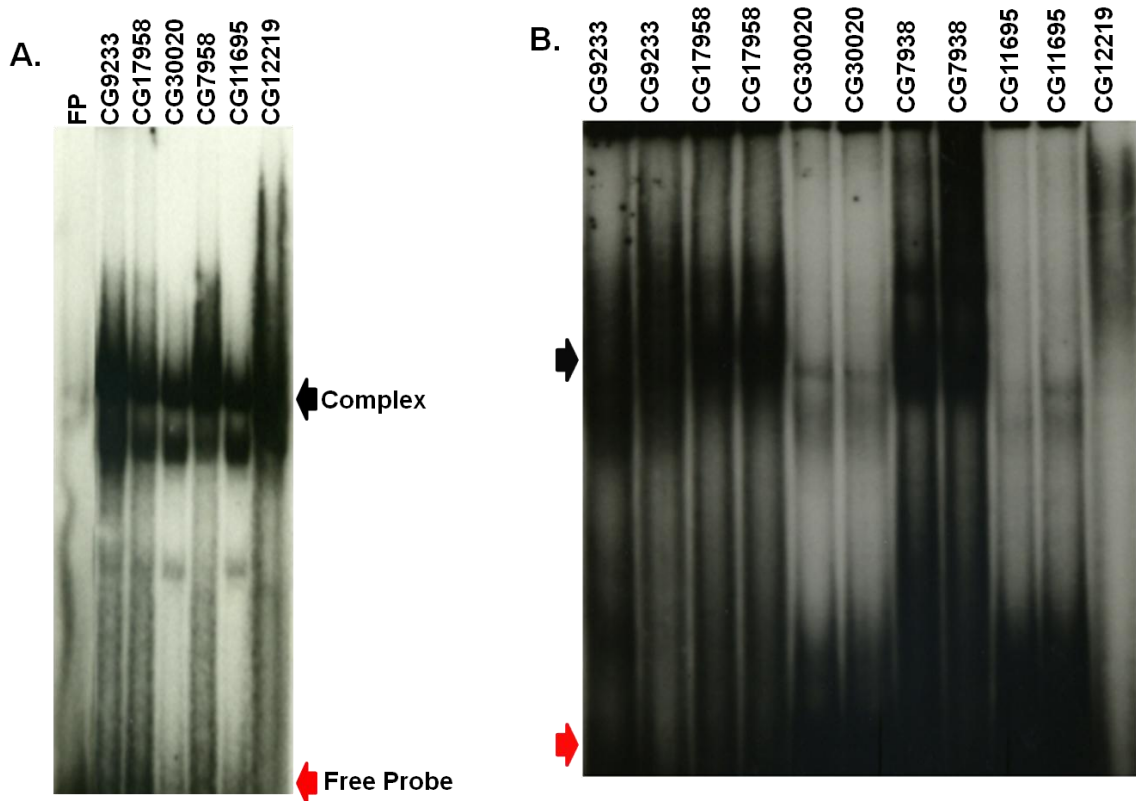
Purifications shown for CG18555, CG30431 (A) , CG10366, CG14711, CG7928, CG12319, CG4143 and CG8319 (B). First, second, and third elutions from two independent protein inductions and purifications shown with broad range marker and whole cell lysate from protein inductions for comparison.



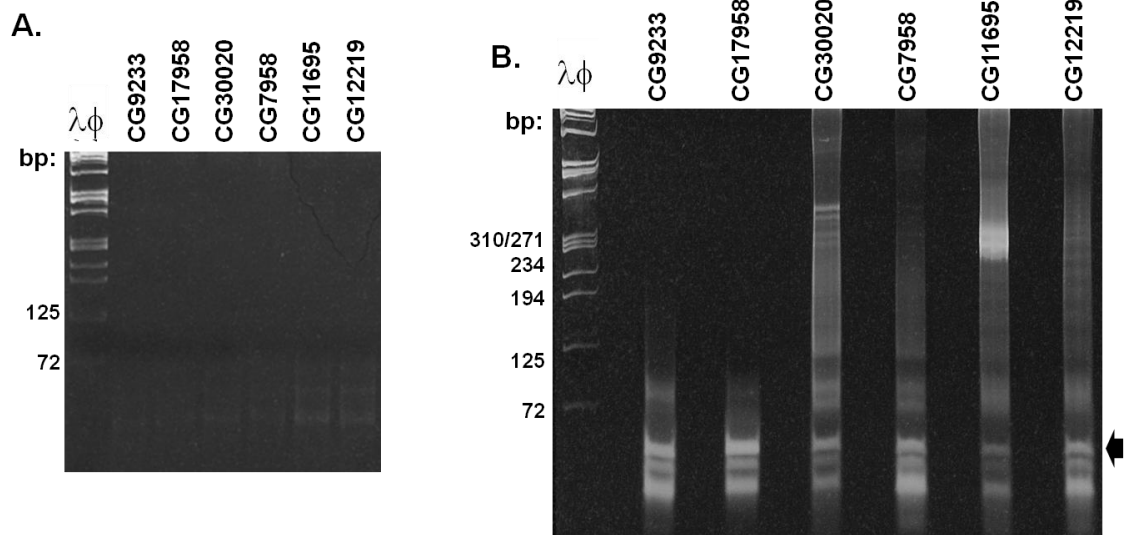
**Figure 8. GST-ZnF construct Proteins: Multiple independent clones expressing each GST-ZnF construct protein were checked for expression of correctly sized and soluble protein.** Protein extracts were purified on GSH-sepharose columns, eluted in reduced glutathione containing buffer and dialyzed to remove reduced glutathione. Purified proteins for the initial five ZAD family members examined (**A**) and all additional members (**B**, **C**) are shown with Broad Range Protein Marker and purified GST affinity tag as control.



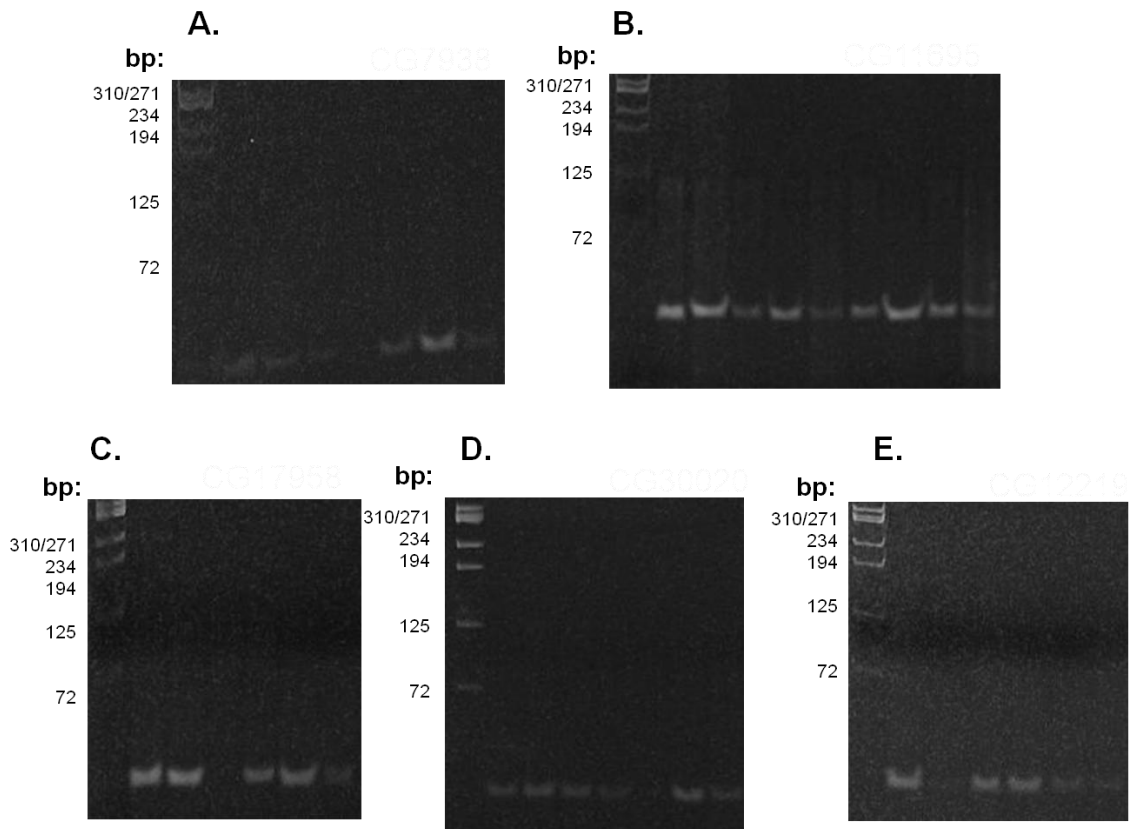
**Figure 9. Binding site selection experiment.** GST-ZNF constructs were combined with an N18 (49mer) oligonucleotide library. Each oligonucleotide contained an 18N region flanked by known 18mer regions containing a known sequence for PCR primers and restriction sites for cloning. Sequences that interact with the GST-ZNF construct were retarded in an electrophoretic mobility shift assay. Bands containing the protein DNA complex were eluted from the gel and re-amplified by PCR. This enriched library was used for the second round of selection, and the process repeated for four total selection cycles. The first selection with the unenriched library (A) is shown with a 60-hour exposure. The third enriched selection is shown with an 18-hour exposure (B).



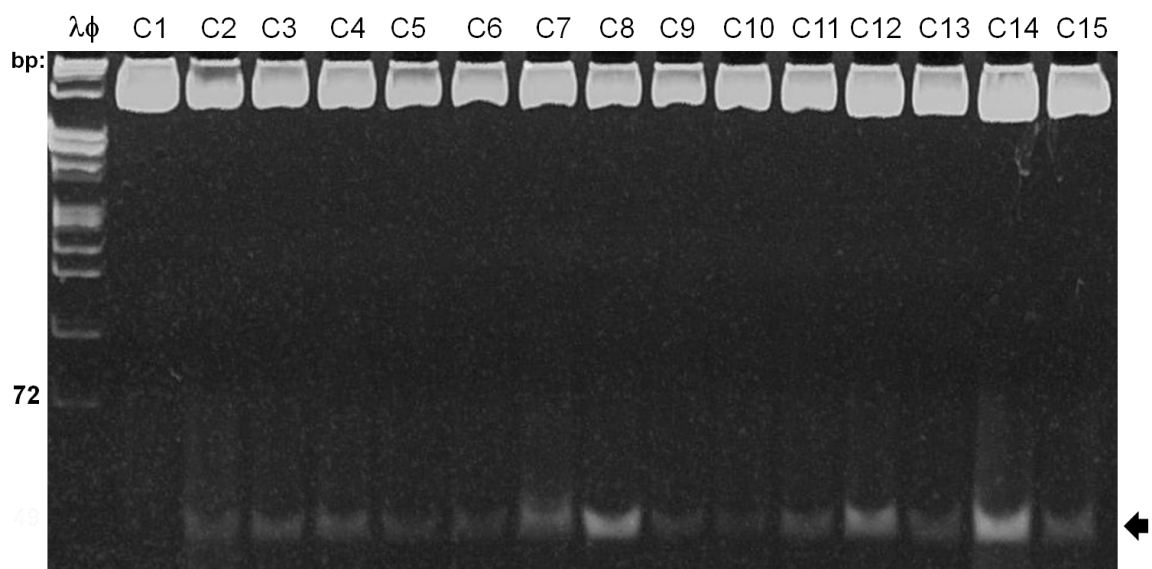
**Figure 10. DNA-protein complexes from the first binding site selection against the random oligonucleotide library.** Each sample was treated with proteinase K, and cleaned with PCI and CI extractions. They were then amplified by PCR and purified by size on a 4% DNA PAGE gel, visualized under UV illumination after staining in a buffer of 1% EtBr for 20 minutes (A). This enriched library was used for the subsequent round of binding site selection. The process was repeated for a total of four selections. The complex from the final selection (B) was purified by size on a 4% DNA PAGE gel, visualized under UV illumination after staining in a buffer of 1% EtBr for 20 minutes. The samples were then digested by restriction enzymes, ligated into a plasmid vector and then transformed into *E. coli* DH5a cells.



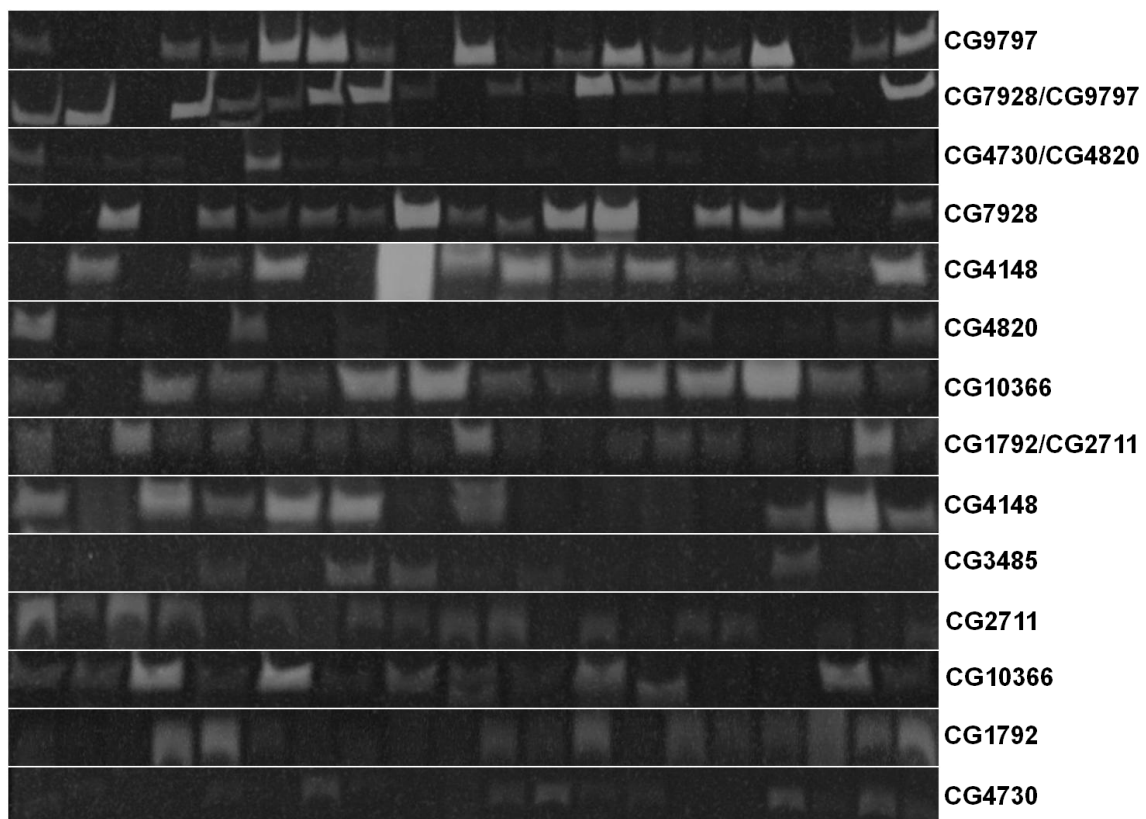
**Figure 11. Products from the final enriched EMSA selection.** Samples were amplified by PCR and ligated into pUC18 vector. Ligations were transformed into DH5 $\alpha$  cells for mini plasmid preparation. Before sequencing, each independent clone was checked for the presence of a cloned insert by restriction enzyme digestion of the poly-linker region. All positive clones were sent to (UF) for sequencing. Shown are portions of the CG7938 (A), CG11695 (B), CG17958 (C), CG30020 (D), and CG12219 (E) digests



**Figure 12. Restriction digests from CG14710.** Restriction digests of 15 independent clones from the population of oligonucleotide sequences efficiently binding the CG14710 ZfP construct proteins sent for sequencing and consensus analysis.



**Figure 13. Clones sent for sequencing.** A compilation of insert sequences released from independent clones of multiple BSS assays sent for sequencing and consensus analysis.



**Figure 14. Sequence data from multiple independent clones from the CG30020**

**binding site selection using EMSA.** Each table shows the consensus above with the full sequence information or a summary of that information below. Below: A visual summary of the relative abundance of each nucleotide at a given position in the sequence.

Consensus	<b>G</b>	<b>R</b>	<b>G</b>	<b>C</b>	<b>R</b>	<b>C</b>	<b>G</b>	<b>R</b>
	<b>G</b>	A	<b>G</b>	<b>C</b>	A	<b>C</b>	<b>G</b>	<b>G</b>
	<b>G</b>	<b>G</b>	<b>G</b>	<b>C</b>	A	<b>C</b>	<b>G</b>	<b>G</b>
	<b>G</b>	A	T	<b>C</b>	A	<b>C (C)</b>	<b>G</b>	<b>G</b>
	<b>G</b>	T	<b>G</b>	<b>C</b>	T	<b>C</b>	<b>G</b>	A
	<b>G</b>	<b>G</b>	<b>G</b>	<b>C</b>	<b>C</b>	<b>C</b>	<b>C</b>	A
	<b>G</b>	<b>G</b>	<b>G (T)</b>	<b>C</b>	<b>G</b>	<b>C</b>	<b>C</b>	A
	<b>G</b>	<b>G</b>	<b>G</b>	<b>C</b>	<b>G</b>	<b>C</b>	T	<b>G</b>
	<b>G</b>	<b>G</b>	T	<b>C</b>	<b>G</b>	<b>C</b>	<b>G</b>	<b>C</b>
A	0	2	0	0	3	0	0	3
T	0	1	2	0	1	0	1	0
<b>C</b>	0	0	0	8	1	8	2	1
<b>G</b>	8	5	6	0	3	0	5	4
%	100	88	75	100	75	100	63	88





**Figure 15. Sequence data from multiple independent clones from the CG12219 binding site selection using EMSA.** Each table shows the consensus above with the full sequence information or a summary of that information below. Below: A visual summary of the relative abundance of each nucleotide at a given position in the sequence.

Consensus	R	G	T	R	T	G	G	A	G
A	G	T	G	T	G	G	A	G	
G	G	T	A	T	G	G	A	G	
G	G	-	G	T	G	G	A(T)	G	
G	G	-	A	T	G	G	G	G	
G	G	-	G	T	A	G	A	G	
A	-	T	C	T	G	G	A	-	
-	-	A	G	T	G	G	A(T)	G	
-	-	-	-	T	G	G	A	G	

	R	G	T	R	T	G	G	A	G
A	2	0	1	2	0	1	0	7	0
G	4	5	0	4	0	7	8	1	7
T	0	0	3	0	8	0	0	0	0
C	0	0	0	1	0	0	0	0	0
	100	100	75	86	100	88	100	88	100



**Figure 16. Sequence data from multiple independent clones from the CG7938**

**binding site selection using EMSA.** Each table shows the consensus above with the full sequence information or a summary of that information below. Below: A visual summary of the relative abundance of each nucleotide at a given position in the sequence.

G	G	G	T	G	C	T/C	G/A
G	A	G	T	G	C	C	A
G	G	G	T	C	G	T	A
G	G	G	T	G	C	T	G
G	G	G	T	G	C	T	G
T	G	G	T	G	C	T	A
C	G	G	T	G	C	T	G
C	T	G	T	G	C	G	A
G	G	G	T	G [G]	C	T	G
G	G	G	T	A	C	G	A
G	G	G	T	A	C	C	G
C	G	G	T	G	G	T	G
C	G	G	C	G	C	T	G
G	G	G	A	G	C	C	G
G	G [C]	G	A	G	C	G	G
G	G	G	-	G	C	G	G
G	G	G	T	-	C	C	G
G	G	G	A	C	C	-	G

	G	G	G	T	G	C	T/C	G/A
A	0	1	0	3	2	0	0	5
T	1	1	0	12	0	0	8	0
C	4	0	0	1	1	15	4	0
G	12	15	17	0	13	2	4	12
%	70	90	100	70	75	90	70	100



**Figure 17. Sequence data from multiple independent clones from the CG17958**

**binding site selection using EMSA.** Each table shows the consensus above with the full sequence information or a summary of that information below. Below: A visual summary of the relative abundance of each nucleotide at a given position in the sequence.

C	A	G	C	G	C	A	G	T	G	G	G	C	C	C	C	A	C	
C	A	G	C	G	C	A	G	T	G	G	G	C	C	C	C	A	C	
C	A	G	C	G	C	A	G	T	G	G	G	C	C	C	C	A	C	
C	A	C	C	G	C	A	A	T	G	G	G	C	C	A	C	A	C	
C	A	G	C	G	C	A	G	T	G	G	G	C	C	C	C	A	C	
				G	C	A	-	T	C	G	G	C	C	C	C	C	C	
C	A	T	T	G	C	A	-	T	C	G	G	C	C	C	C	C	C	
C	A	-	C	A	C	A	-	T	G	A	G	C	C	C	G	G	T	
C	A	T	T	G	C	A	-	T	G	G	G	A	C	G	A	T	C	
C	A	T	T	G	C	A	-	T	G	G	G	T	A	C	G	T	G	
Consensus sequence																		
	C	A	G	C	G	C	A	G/A	T	G	G	G	C	C	C	C	A	C
A	0	8	0	0	0	0	9	1	0	0	1	0	1	1	1	1	4	0
T	0	0	3	3	0	0	0	0	9	0	0	0	1	0	0	0	2	1
C	8	0	1	5	1	9	0	0	0	2	8	9	7	8	7	6	2	7
G	0	0	3	0	8	0	0	3	0	7	0	0	0	0	1	2	1	1
	100	100	33	55	89	100	100	40	100	78	89	100	78	89	78	66	45	78

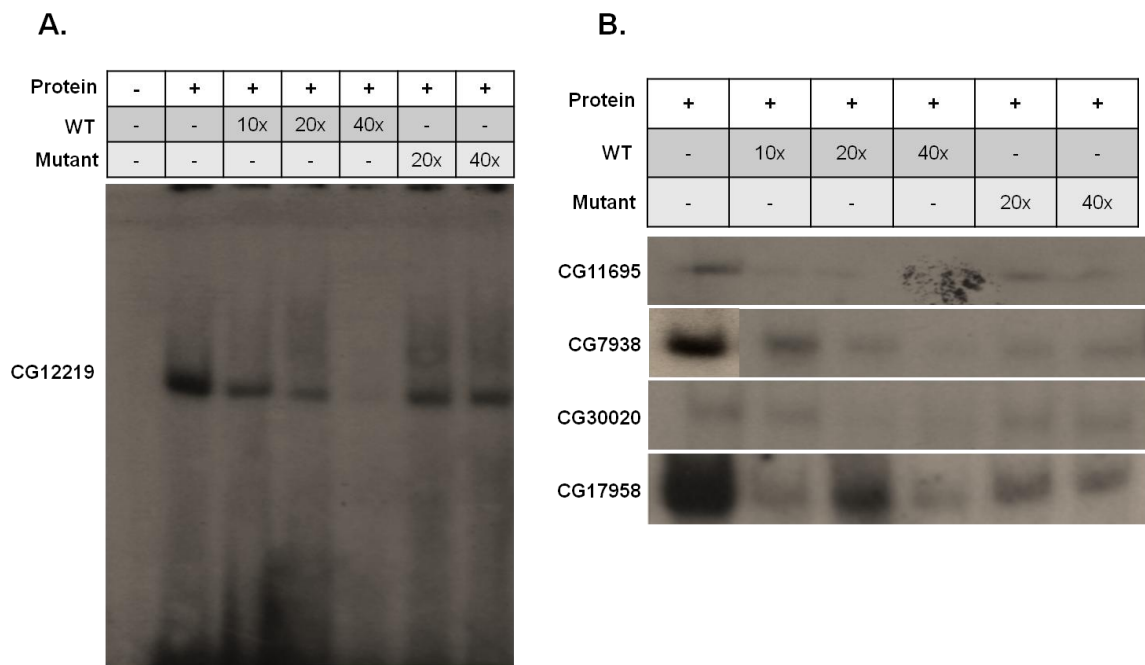


**Figure 18. Sequence data from multiple independent clones from the CG11695 binding site selection using EMSA.** Each table shows the consensus above with the full sequence information or a summary of that information below. Below: A visual summary of the relative abundance of each nucleotide at a given position in the sequence.

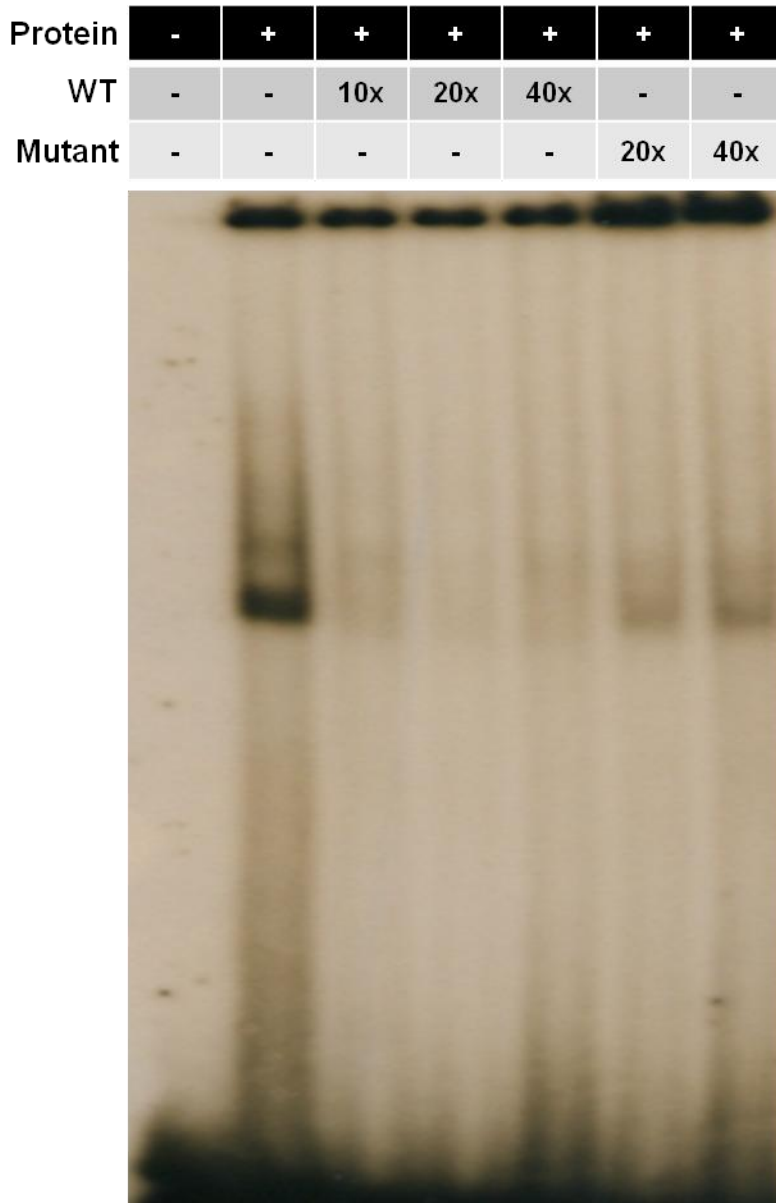
Consensus		N	N	C	A	N	N	T	G	N	N
		C	A	C	A	C	A	T	G	A	G
		C	A	C	A	A	G	T	G	G	A
		T	C	C	A	A	C	T	G	T	C
		C	C	C	A	C	C	T	G	T	C
		G	A	C	A	C	G	T	G	C	A
		T	G	C	A	C	A	T	G	G	G
		C	A	C	A	T	A	T	G	T	A
		T	G	C	A	C	C	G	G	C	T
		C	G	C	A	G	T	G	G	G	C
		C	A	C	A	T	G	G	G	A	C
		A	T	C	A	C	C	G	G	C	A
		G	T	C	A	T	A	G	G	T	C
		C	A	C	G	A	C	T	G	T	C
		C	G	C	C	G	T	T	G	G	T
		G	G	C	G	T	T	T	G	G	G
		G	G	C	T	A	T	T	G	C	T
		T	G	C	G	C	A	T	G	T	T
	C	8	2	17	1	7	5	0	0	4	6
	A	1	6	0	12	4	5	0	0	2	4
	T	4	2	0	1	4	4	12	0	6	4
	G	4	7	0	3	2	3	5	17	5	3
Ebox		N	N	C	A	C	A/G	T	G	N	N
Ebox		N	N	C	A	N	N	T	G	N	N



**Figure 19. Binding site selection and competitions.** The identified consensus sequences and a mutant form modified at the most conserved nucleotide positions were used in a series of competitive binding EMSAs. In each case, the identified consensus efficiently bound the protein and was best competed (identified by the loss of complex) by the wild type sequence. Mutant forms were much less effective at dislodging the consensus sequence, indicating a specificity of binding at those conserved positions. The full gel for GS12219 (**A**), and the relevant portions of CG17958, CG7938, CG11695, and CG30020 (**B**) are shown. Gels edited for size.

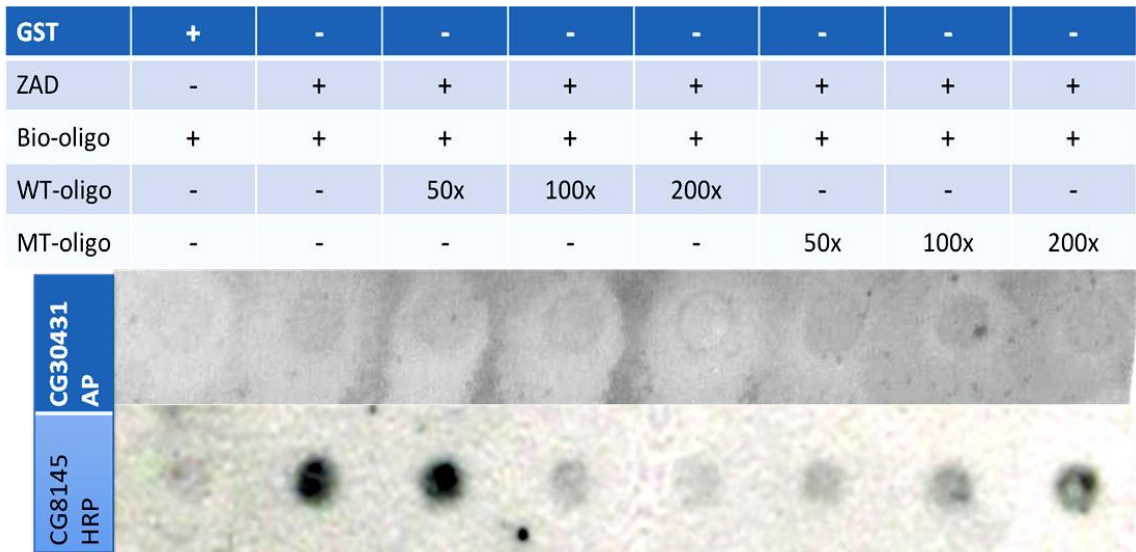


**Figure 20. Competition analysis of a protein selected under the modified binding site selection procedure.** The identified consensus efficiently bound the protein and was best competed (identified by the loss of complex) by the wild type sequence. Mutant forms were much less effective at dislodging the consensus sequence, indicating a specificity of binding at those conserved positions.

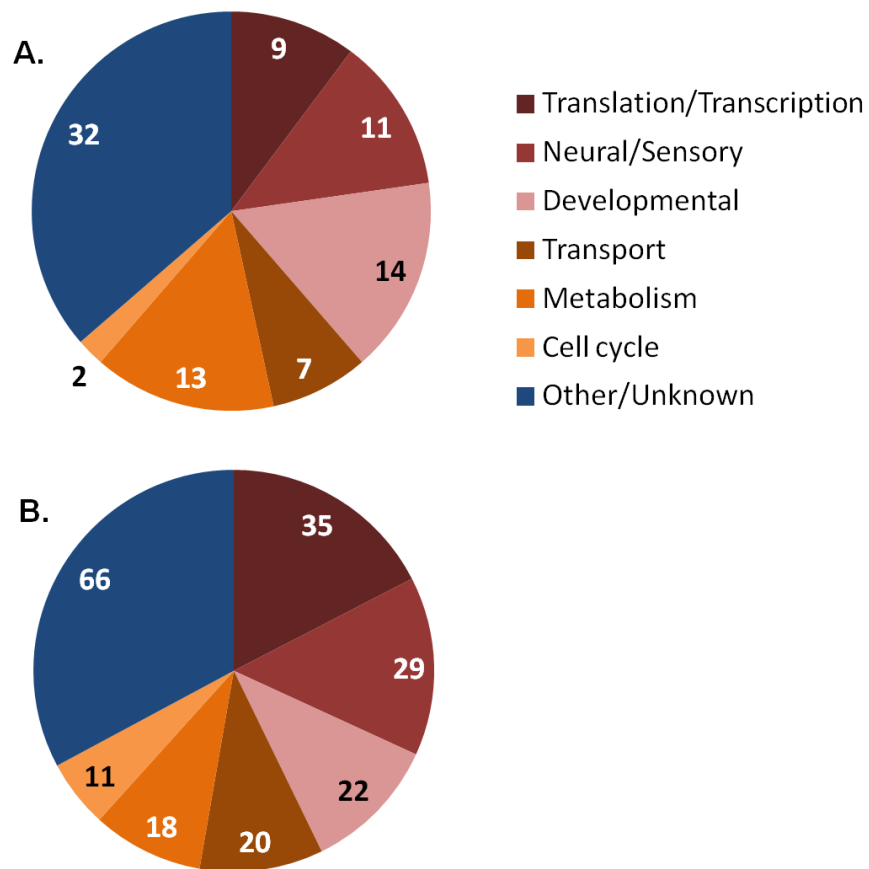


**Figure 21. Competition analysis developed using a modified dot blot analysis.**

Competitions of the biotinylated CG30431 and CG8145 consensus oligonucleotides against unlabeled wild type and scrambled oligonucleotides at 50, 100, and 200 times label concentration. Visualized by alkaline phosphatase (AP) and horse radish peroxidase conjugated to streptavidin.

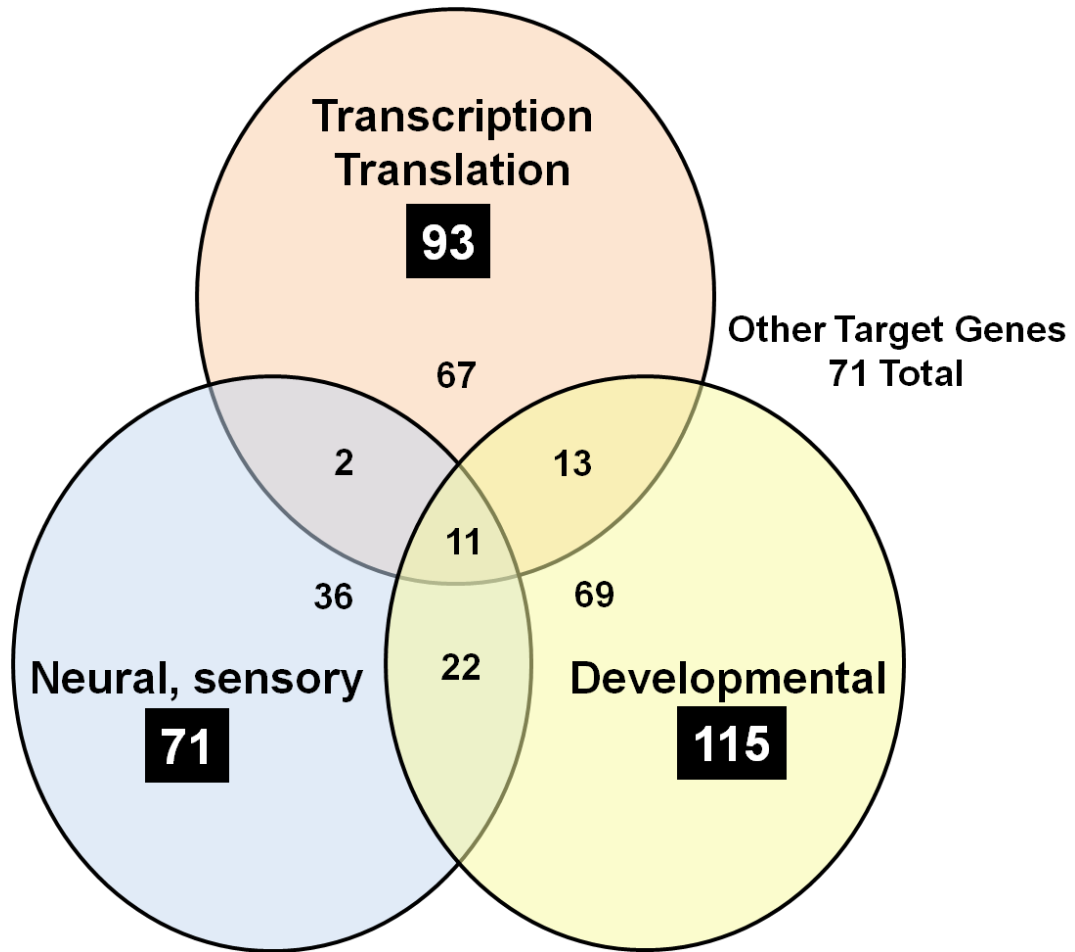


**Figure 22. The genes predicted as targets for CG18555 and CG7928.** The predicted target genes for CG18555 (A) and CG7928 (B) can be categorized into six main groups; transcription/translation, neural/sensory, developmental, transport, metabolism, and cell cycle control. The first three of these are known targets for the few previously characterized ZAD proteins. The last group contains primarily genes of unknown function and several genes unrelated to the other categories.

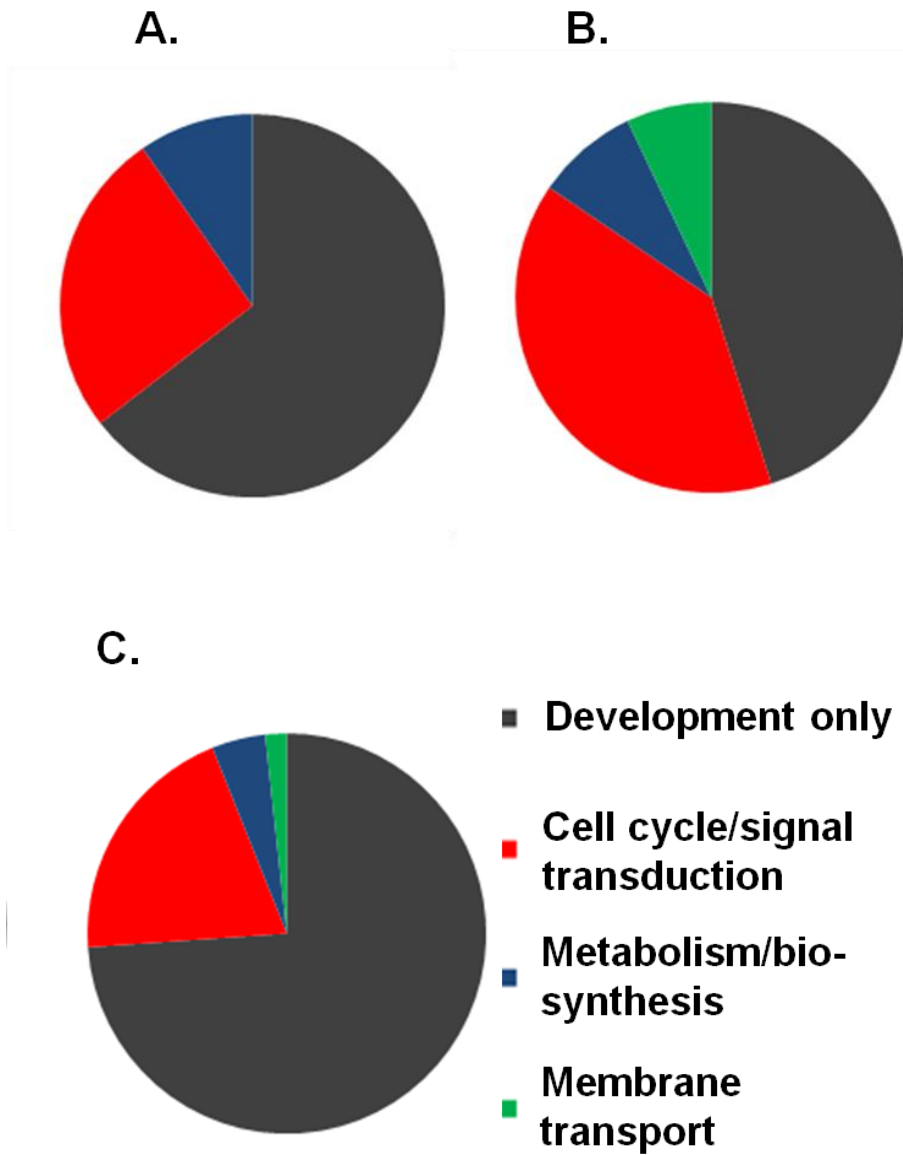




**Figure 23. Relative incidents of potential target genes with primary functions.** A set diagram showing the association between predicted ZAD target genes and multiple axis relevant to the previously reported ZAD functions.



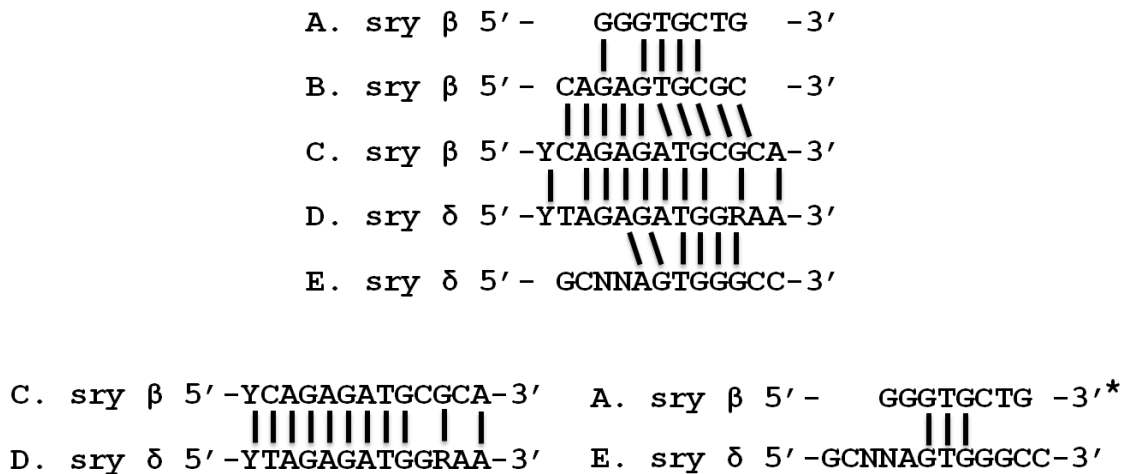
**Figure 24. An analysis of the prevalence of secondary gene functions.** Functions of those putative target genes with reported functions involving (A) Transcription and Translation, (B) Neural and Sensory, and (C) Development.



**Figure 25. Comparisons between five reported serendipity binding site sequences.**

**A-** Our EMSA derived consensus binding site for *sry*  $\beta$ , **B-** The reported in vivo binding site for *sry*  $\beta$  from Payre et al (1991), **C-** The nuclease protection assay derived consensus binding site for *sry*  $\beta$  from Payre et al (1991), **D-** The nuclease protection assay derived consensus binding site for *sry*  $\delta$  from Payre et al (1991), and **E-** Our EMSA derived consensus binding site for *sry*  $\delta$ . The binding site consensus sequence reported here for *sry*  $\beta$  is significantly similar (5 of 8 conserved nucleotides) to those previously reported, and the sequence reported here for *sry*  $\delta$  is also quite similar (6 of 10 well conserved nucleotides), the relative similarity between the two different members seen in each study is drastically different. Payre et al (1991) reported similarity in 10 of the 13 positions, and we are reporting only 3 of the 8/10 positions showing similarity.

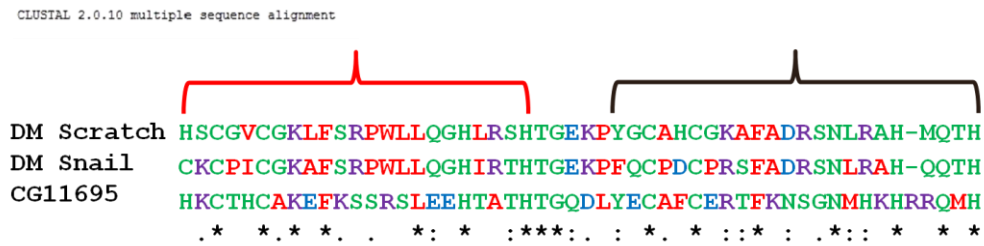
\* Alignment shifted from the possible 4/10 conserved position in order to include the 100% conserved T nucleotide that would otherwise be excluded.



**Figure 26. A ClustalW comparison between the conserved DNA binding domains in a selection of mammalian SNAG proteins.** Results show a strong homology between the second and third (Snail) and third and fourth zinc fingers (Smuc, Slug, Scratch) which are predicted to mediate the conserved DNA binding activity.

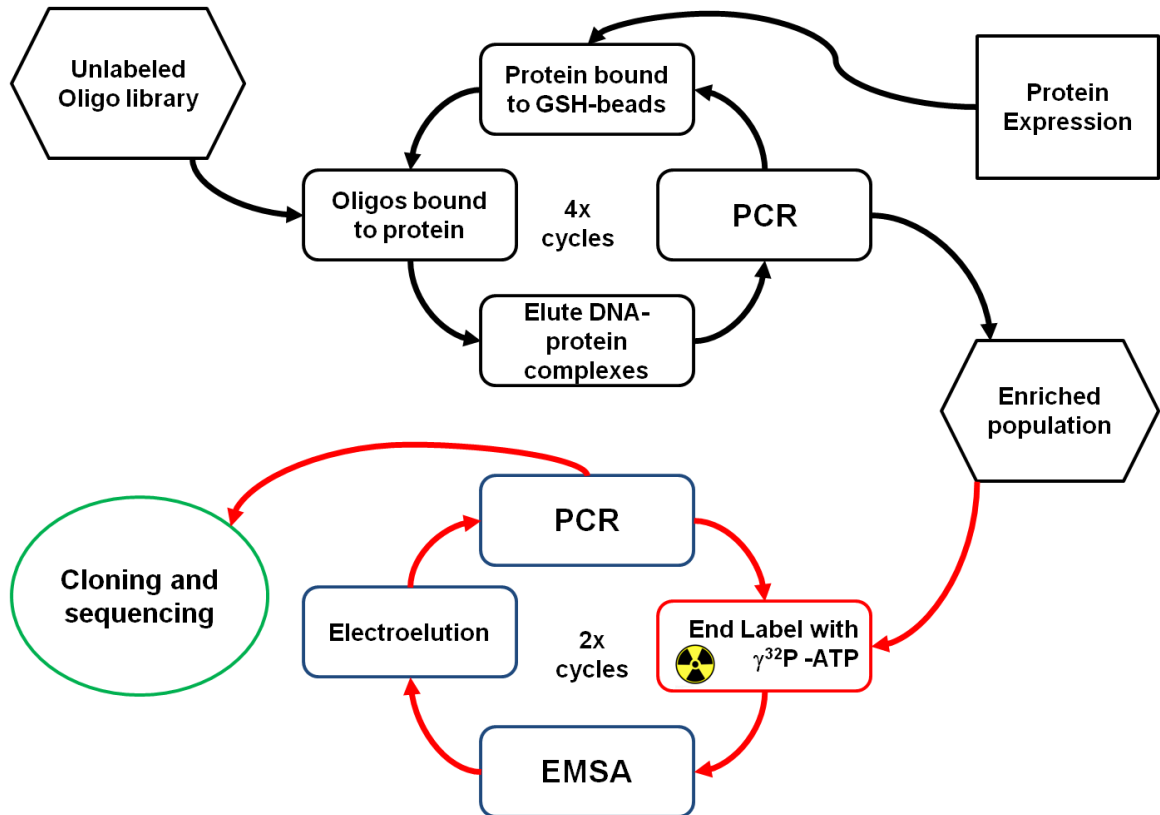


**Figure 27. A ClustalW comparison between the conserved DNA binding domains in the Drosophila SNAG proteins, Scratch and Snail and the zinc finger array from the ZAD protein CG11695. Results show a strong homology between the putative DNA binding regions. This homology is consistent with a similarity in sequence identified in the binding site selection.**

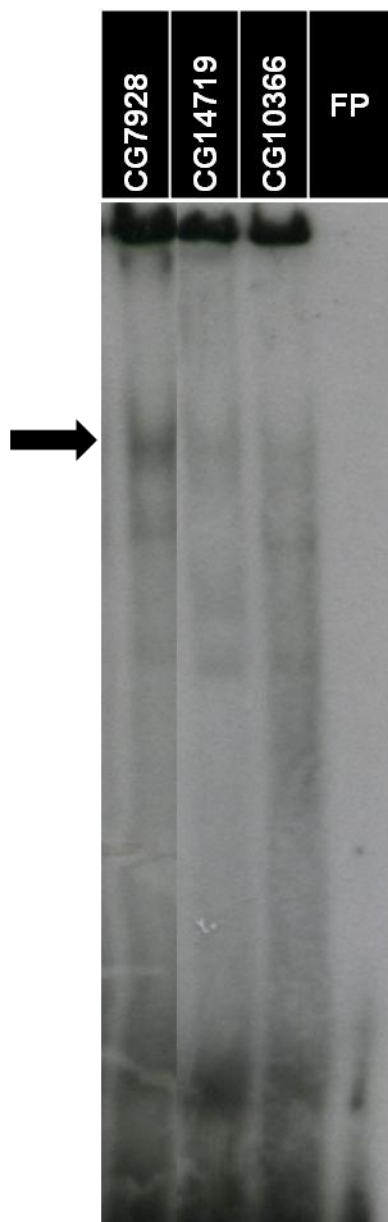


**Figure 28. A schematic representation of the coupled cold and hot binding protocol.**

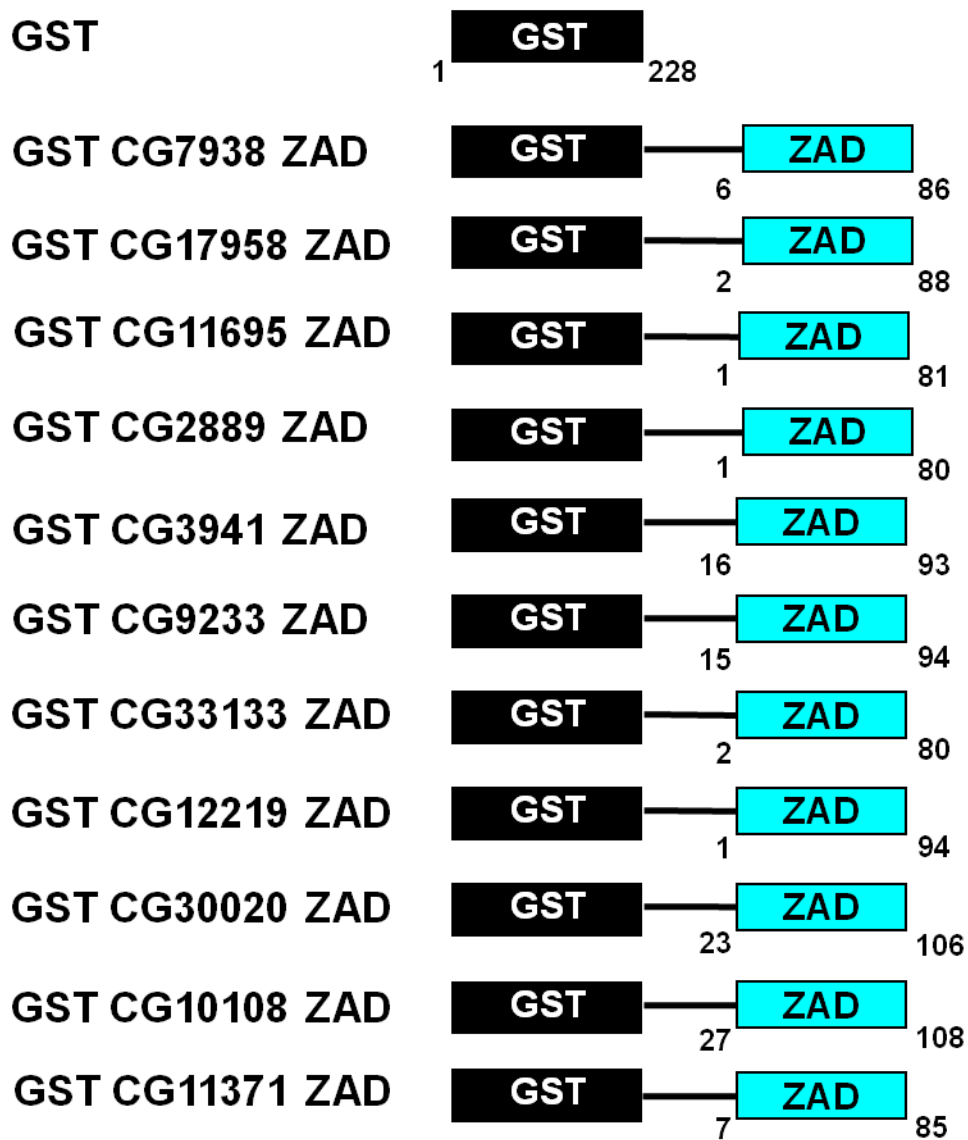
A random double stranded oligonucleotide library is selected first against proteins immobilized on GSH affinity beads and then in a traditional gel shift assay.



**Figure 29. Representative EMSA results from ZAD family members.** Complexes are shown from three members undergoing the first round of selection after the modified cold enrichment technique along with a free probe control.

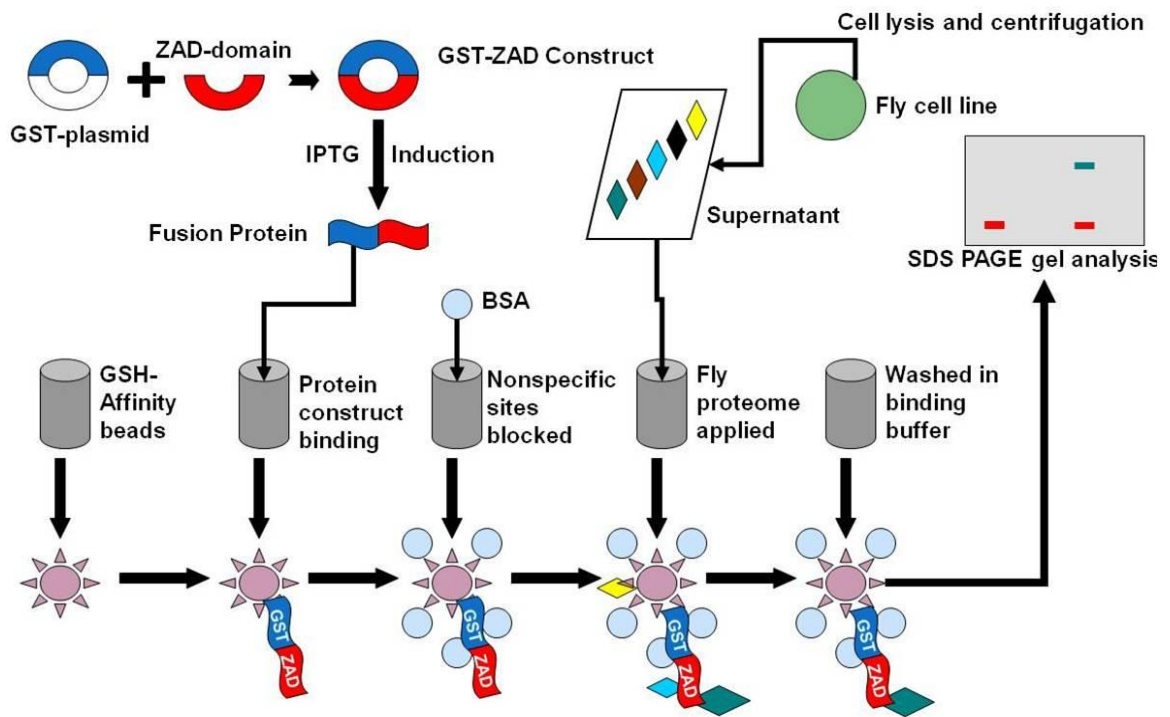


**Figure 30. Diagrammatic representations of the recombinant GST-ZAD fusion proteins.** DNA segments corresponding to the indicated amino acids were PCR amplified by using the full-length genes of ZAD-domain transcription factor family members as templates. Various ZAD domains were fused in frame with the Glutathione-S-Transferase tag to generate the respective recombinant fusion proteins.

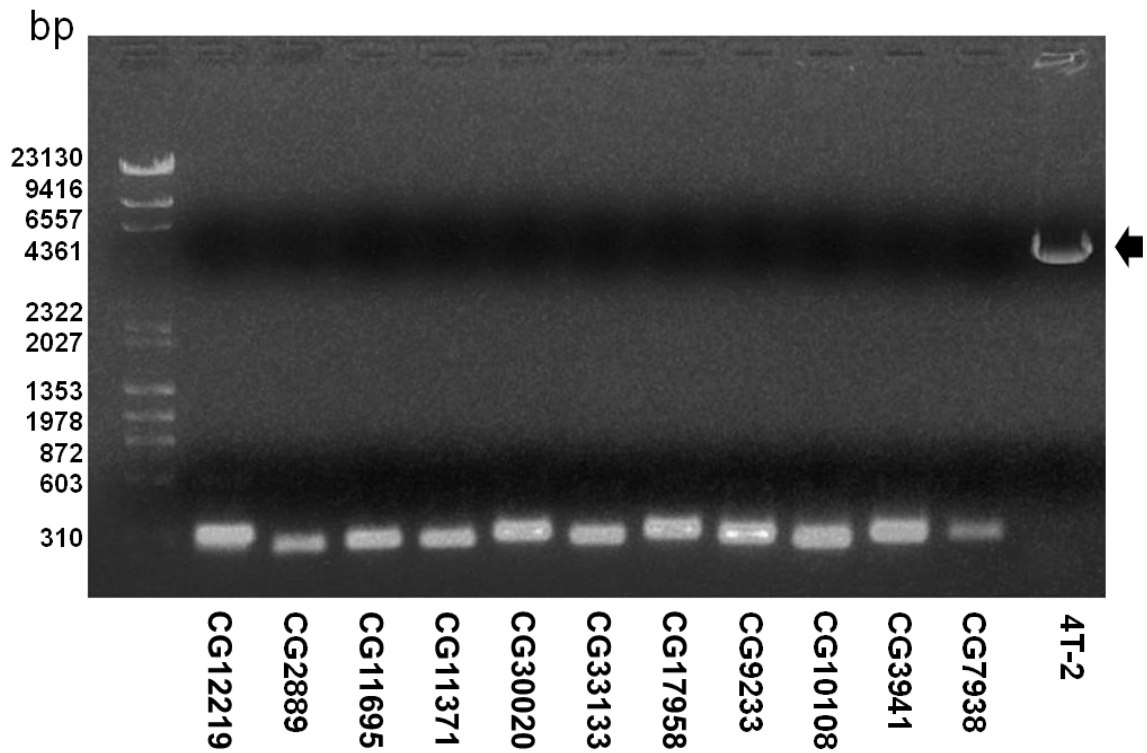




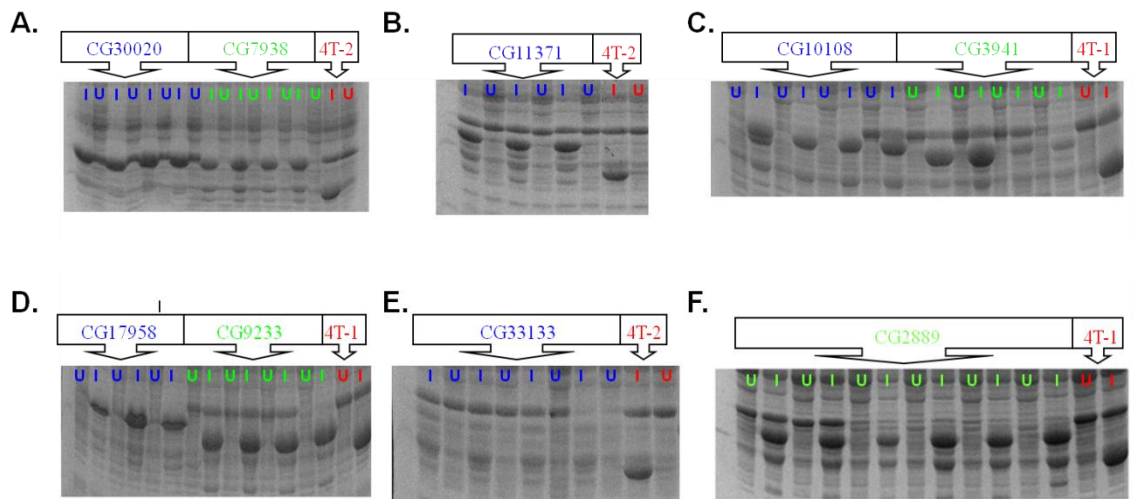
**Figure 31. Diagrammatic representation of GST-ZAD protein binding assay.** GST-ZAD fusion proteins were bound to GSH-affinity beads and incubated with fly cell proteomes extracted from S2 fly cell lines. The resulting protein aggregations were then washed in buffers containing sodium chloride concentrations up to 500mM to remove weakly bound proteins.



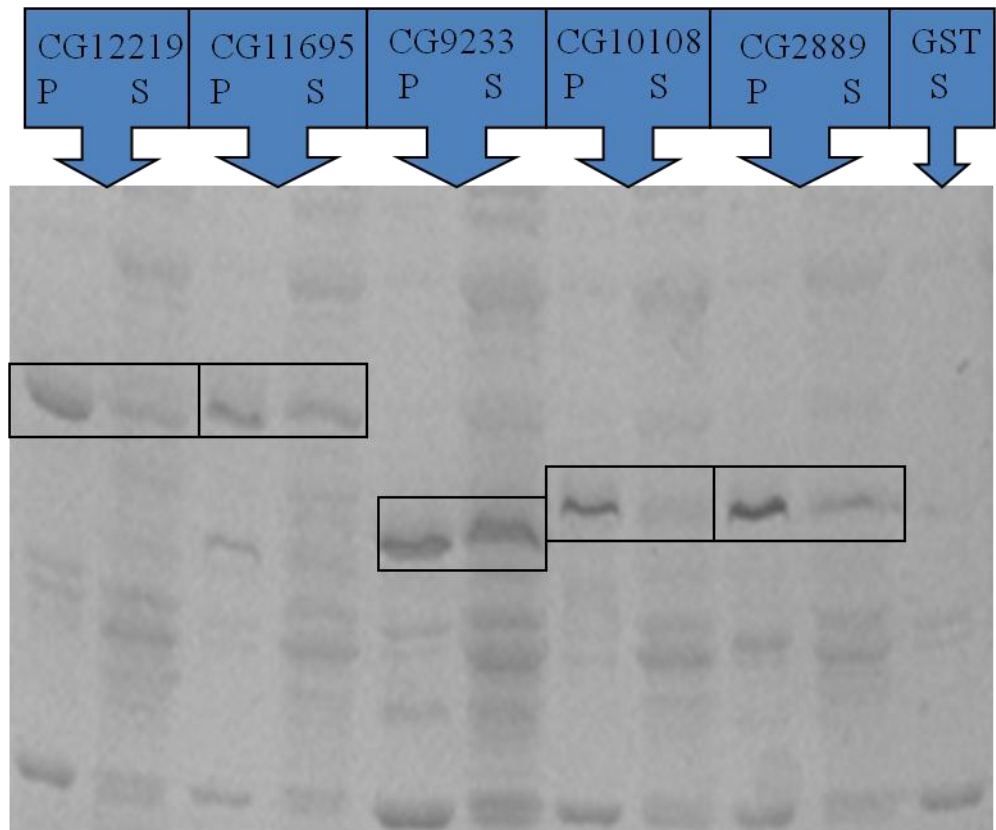
**Figure 32. ZAD domains were amplified by PCR from cDNA.** PCR results of each of the seven *Drosophila* genes using clones purchased from Open Biosystems with each primer pair utilized contained restriction sites. These sites were digested, along with the complementary sites in the pGEX 4T-1 and pGEX 4T-2 plasmid vector polylinker region for ligation and the construction of a GST-ZAD construct



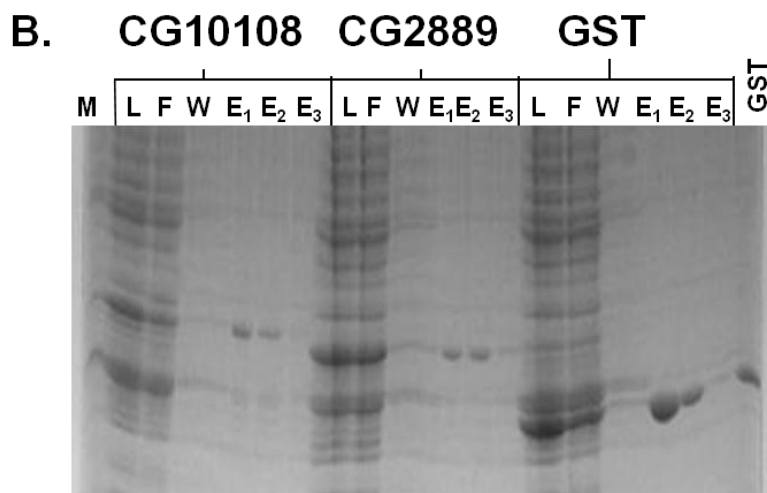
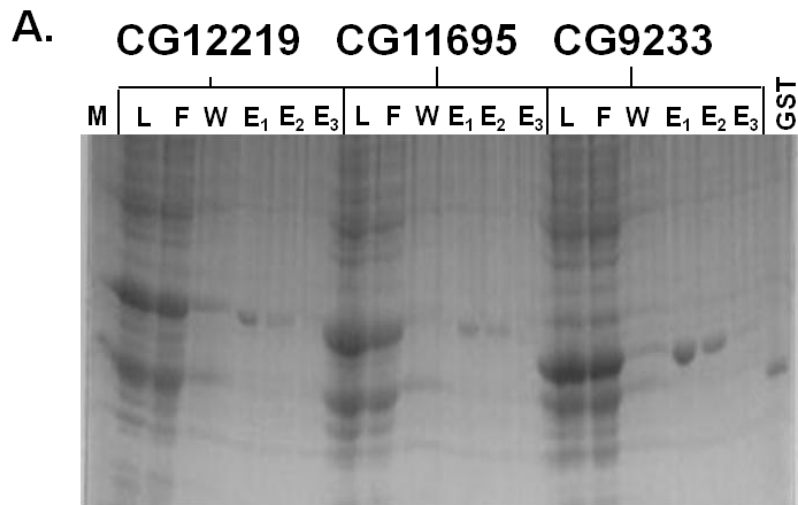
**Figure 33. GST-ZAD protein inductions.** A-F Multiple independent clones of each GST-ZAD domain construct protein were induced for protein production with IPTG treatment. Protein extracts recovered from induced(I) and uninduced(U) cultures were analyzed on SDS PAGE gel with broad range marker for size comparison and induced/uninduced empty pGEX family plasmid as a control.



**Figure 34. Fractionated GST-ZAD protein induction.** Induced cell lysate from induced GST-ZAD constructs were fractionated into soluble and insoluble portions and analyzed on SDS PAGE gels to identify clones producing sufficient soluble and more properly folded protein products.

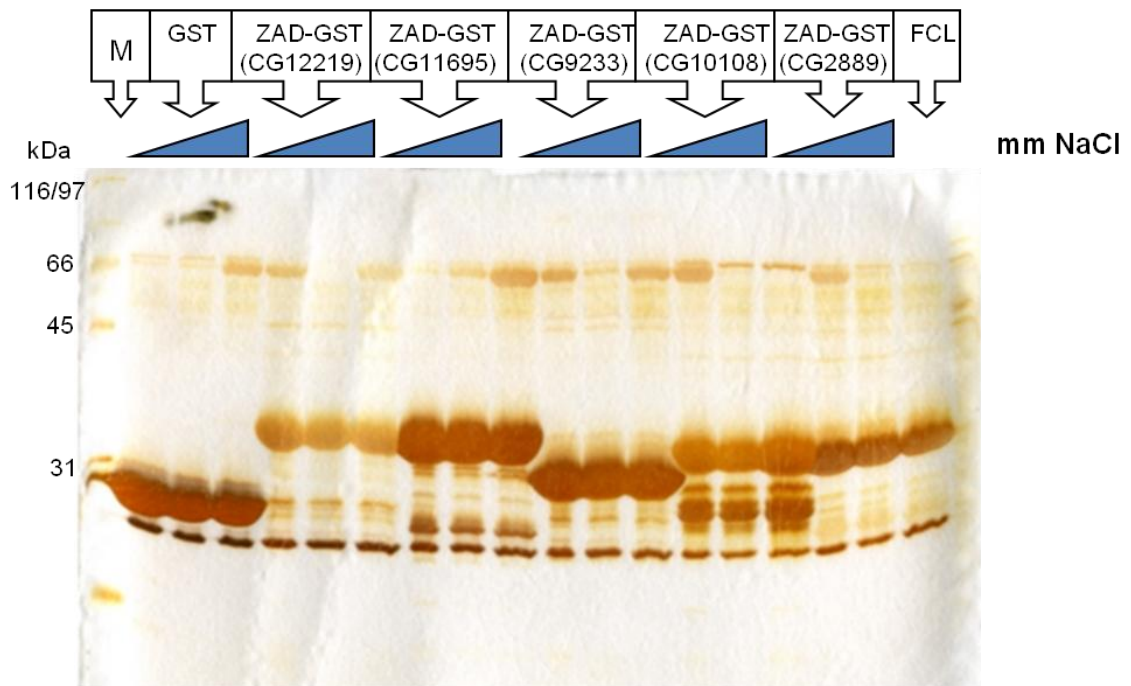


**Figure 35. The purification of GST-ZAD construct proteins expressed in *E. coli* cells visualized on a SDS PAGE gel.** Each gel (A and B) contains a molecular weight marker (M) and a control of purified GST protein. Each panel contains loaded sample (L), flowthrough (F), a PBS wash of the column (W) and three elutions with a Tris buffer containing reduced glutathione. Constructs for CG12219, CG11695, and CG9233 are shown in A. Constructs for CG10108, CG2889, and the GST control are shown in B.

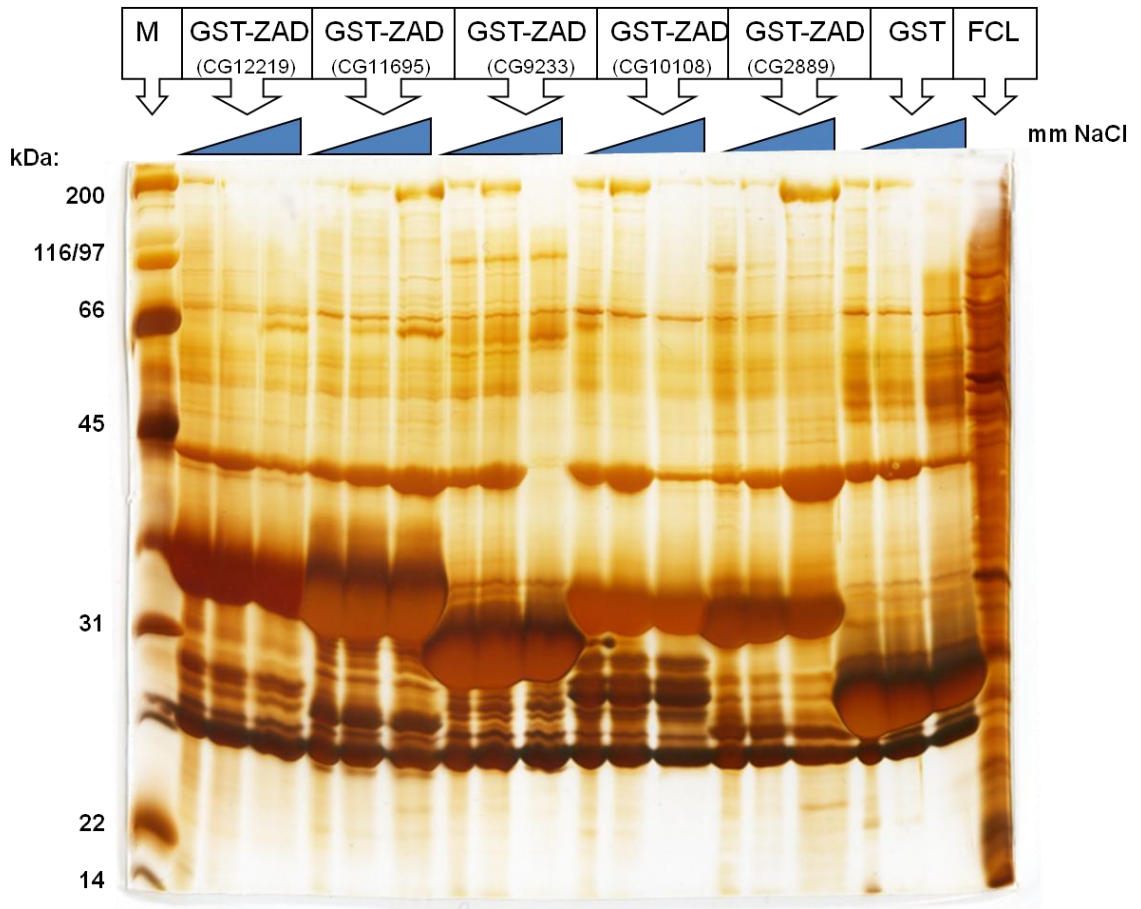


**Figure 36. GST pull down assay to identify putative ZAD-interacting proteins.**

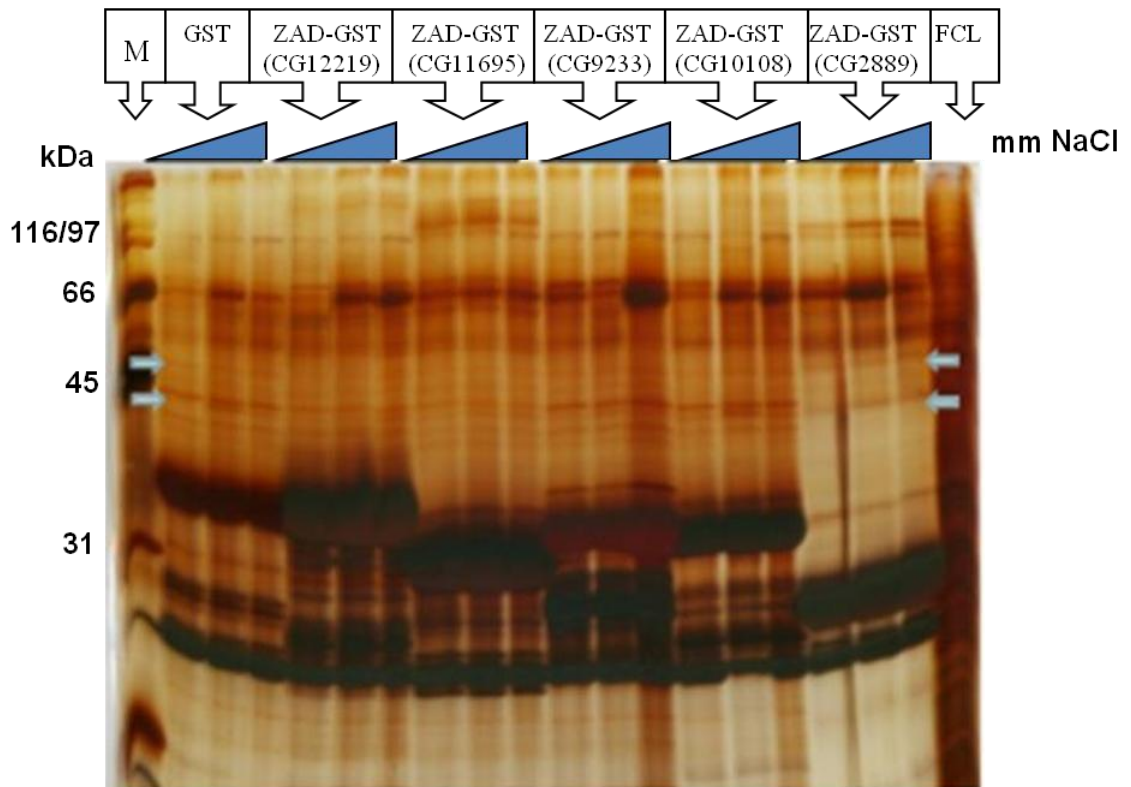
Associated proteins from the GST-ZAD pull down assay were visualized under our initial protocol on a 14% SDS PAGE gel. Each panel shows three separate bindings under increasing concentrations of NaCl, from 100 to 500 mM. Molecular weight marker (M) and the prebound fly cell lysate (FCL) are also shown.



**Figure 37. GST pull down assay.** In order to identify putative ZAD-interacting proteins the assay was visualized with a modified silver staining protocol and run parameters to improve visualizations in the 30-60kDa range.



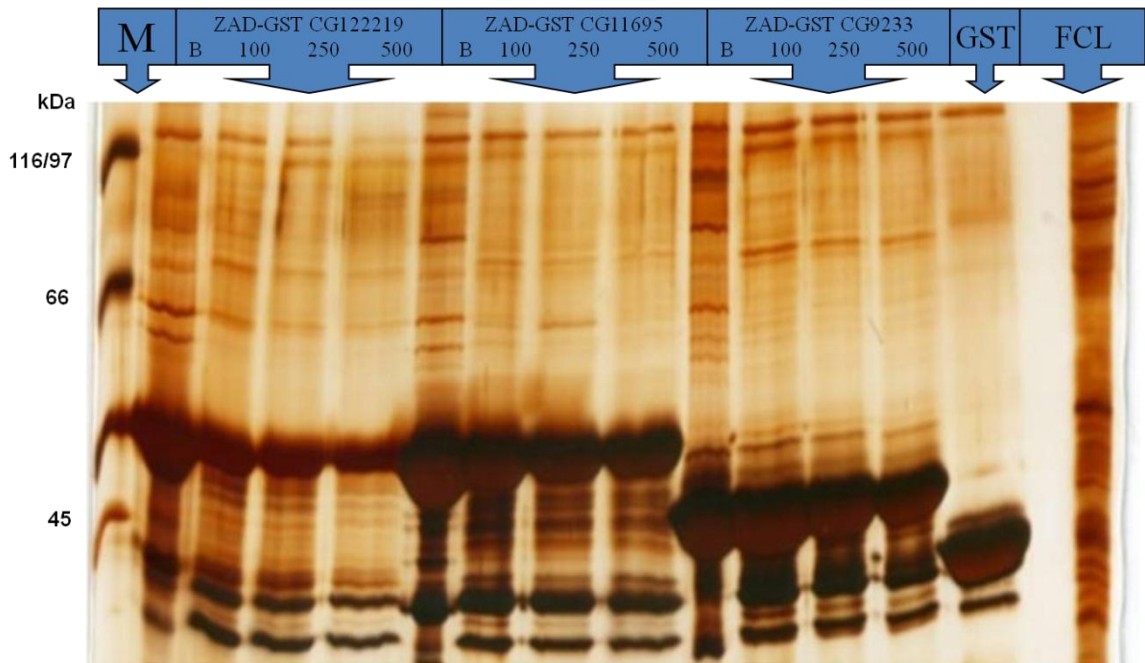
**Figure 38. A GST-pull down assay with the CG12219, CG11695, CG9233, CG10108, and CG2889 ZAD-GST constructs.** Associated proteins from the fly cell extracts retained on the columns were extensively washed with buffer solutions of increasing NaCl concentrations (100 mM, 250 mM, 500 mM) and then electrophoresed on a 12% SDS-polyacrylamide gel and silver stained with a second protocol with increased silver nitrate concentration and an extended incubation time with the substrate to visualize the proteins. A GST only induction (GST) was included as a control. Broad-range- marker (M) and Drosophila S2 cell lysate (FCL) were included for comparison. The region containing the most promising band representing a potential universal co-factor is indicated by arrows.



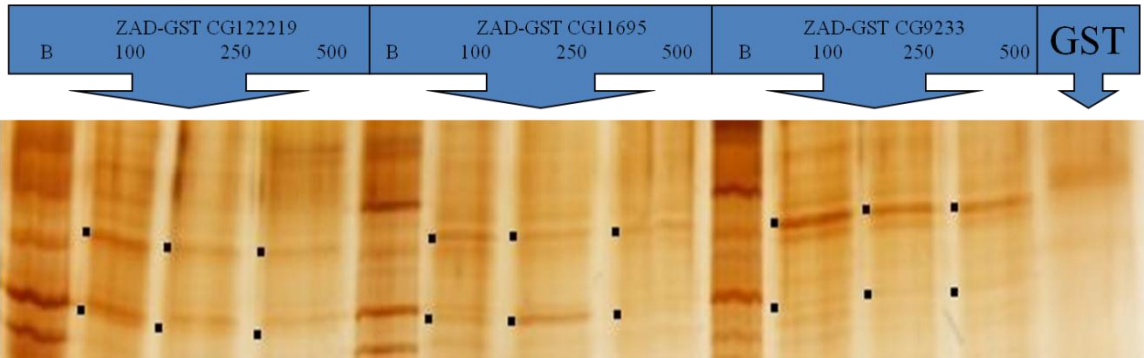


**Figure 39. Associated proteins from the fly cell nuclear extracts.** **A.** Proteins retained on the columns were extensively washed with buffer solutions of increasing NaCl concentrations (100mM, 250mM, 500mM) and then electrophoresed on a 14% SDS-PAGE. A GST only induction (GST) was included as a control. Broad-range- marker (M) and Drosophila Schneider Line 2 cell lysate (FCL) were included for comparison. An aliquot of each ZAD-GST construct bound to beads was retained and not combined with fly cell lysate (B). **B** An enlarged view of the region surrounding the 65kDa marker from the gel in Figure 3. Indicated bands represent bands that appear only in those treatments including both GST-ZAD constructs bound to beads and fly cell nuclear extract.

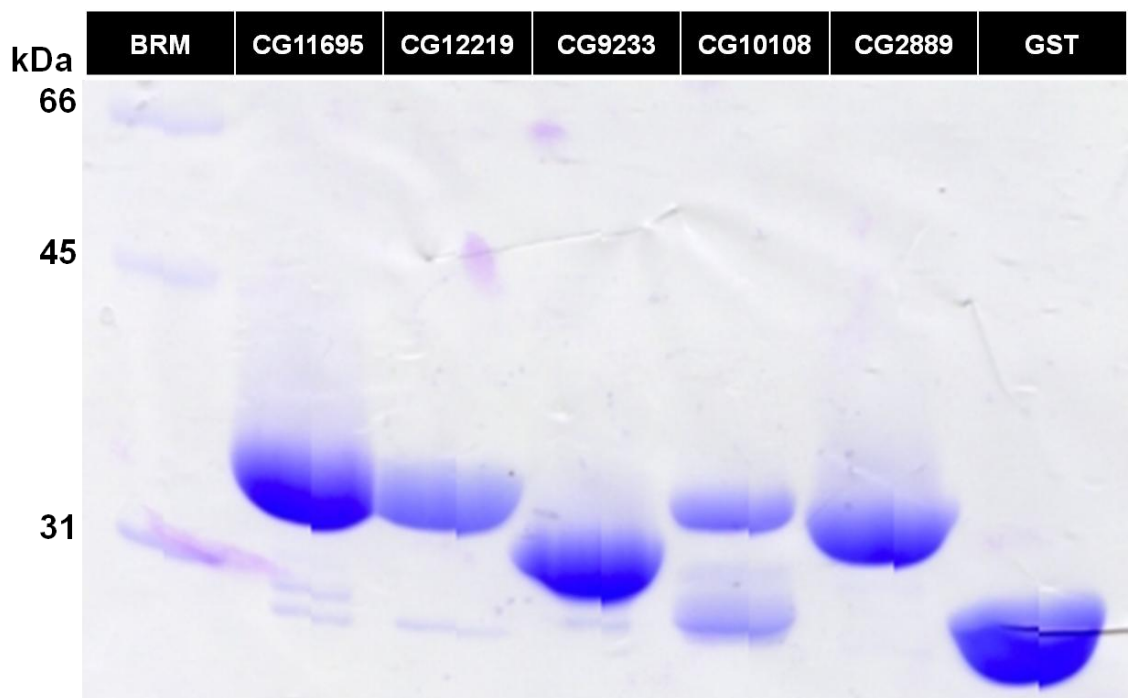
**A.**



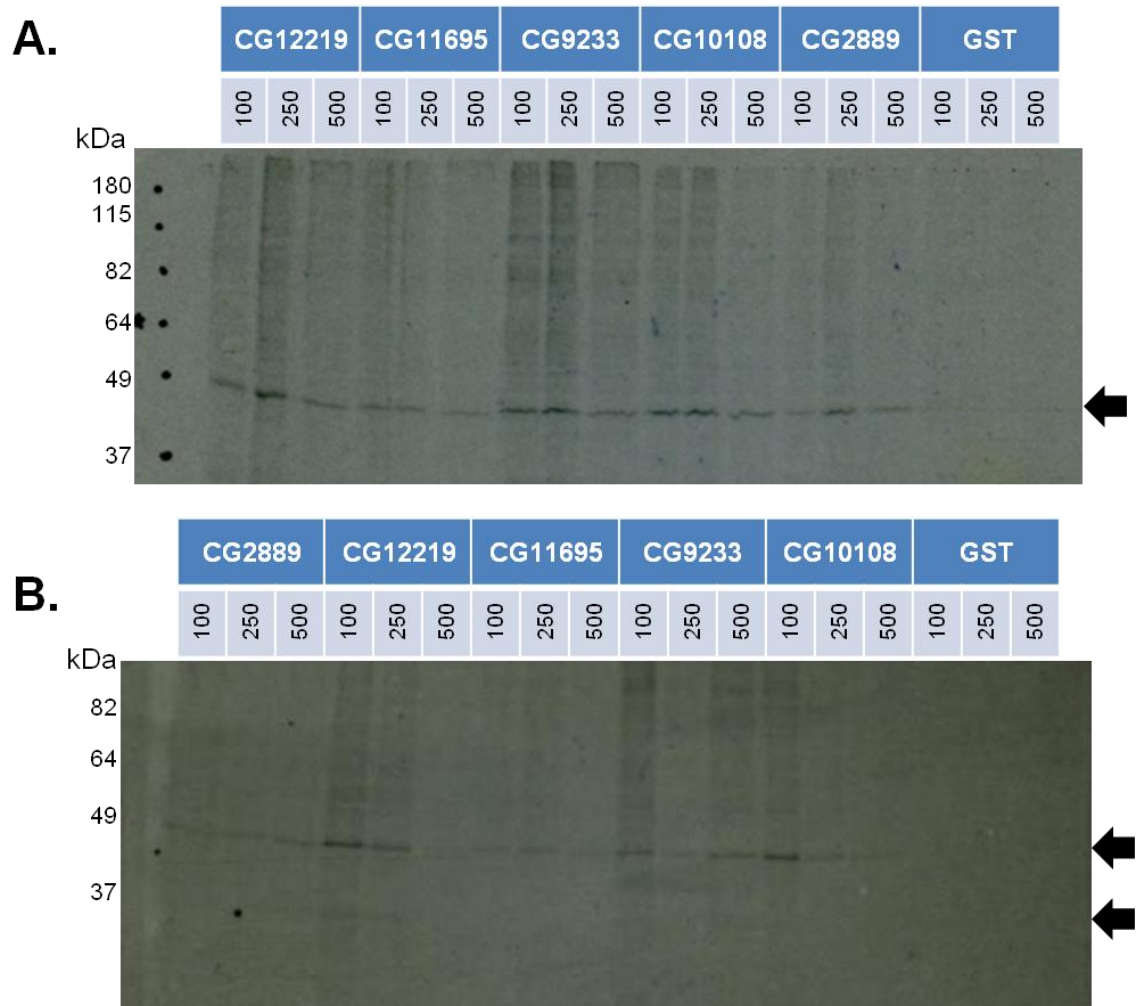
**B.**



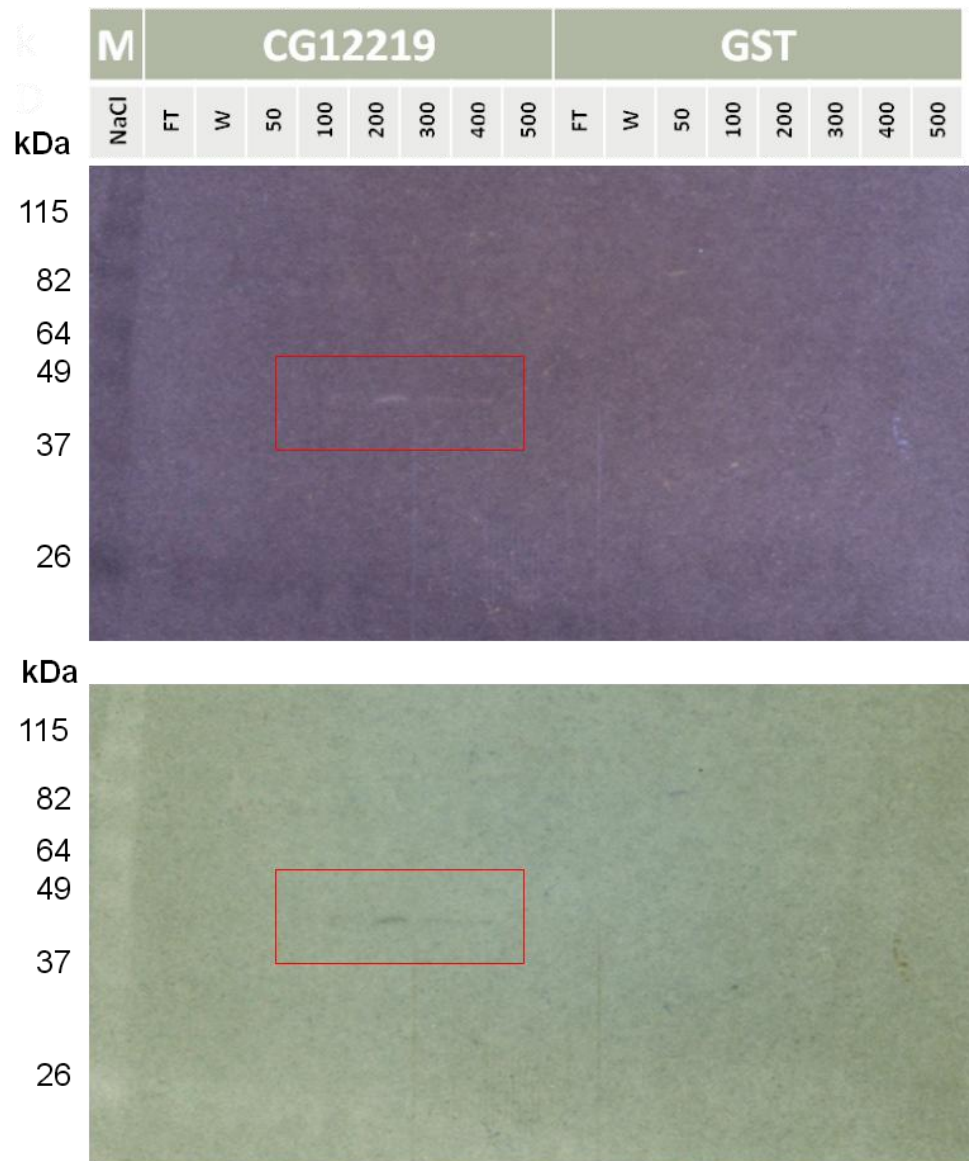
**Figure 40. Protein samples used in the pull-down assays.** Protein samples of equivalent volume and concentration used in the pull-down assays run on an SDS PAGE gel and visualized with Coomassie Blue stain for comparison of experimental and control binding protein concentrations.



**Figure 41. Protein binding pull-down assay repeated with  $^{35}\text{S}$ .** Protein binding pull-down assay was repeated with  $^{35}\text{S}$  labeled S2 cell lysate to reduce background originating from the prokaryotic protein expression vector. Putative ZAD interacting proteins (identified with arrows). Shown in multiple replicates, **A** and **B**.



**Figure 42. Diethylaminoethyl-agarose protein binding assay.** Diethylaminoethyl-agarose fractionated labeled S2 cell lysate was bound to GST-ZAD construct protein containing the ZAD domain from CG12219 and GST protein. A putative ZAD interacting partner in the 49 kDa size range was isolated in the 200mM to 300mM elutions.



**Table 1.** Archetypal ZAD family members selected for BSS analysis. 1-228 of each construct protein represents the GST affinity tag.

\*Additional ZnF domains located outside of the main region.

Protein ID	Native length	ZnF region	ZnFs	Construct length	Construct ZnF region
CG1792	372	196-333	5	381	229-381
CG2711	592	305-527	8	472	229-472
CG3485	690	214-320	4	350	229-350
CG4148	470	271-434	6	407	229-407
CG4730	392	181-346	6	408	229-408
CG4820	383	243-365	4	365	229-365
CG7357	430	207-346	5	383	229-383
CG7928	457	257-454	7	441	229-441
CG7938	356	171-342	6	427	229-427
CG8145	370	182-319	5	374	229-374
CG10267	388	223-359	5	380	229-380
CG10309	924	709-820	4	355	229-355
CG10321	835	625-772	5	391	236-391
CG10366	578	240-469	8	472	229-472
CG11695	544	266-502	8	492	229-492
CG12219	562	138-527	4	645	229-645
CG12391	587	416-525	4	353	229-353
CG14710	415	294-430	5	380	229-380
CG15436	346	126-316	8	434	229-434
CG17958	433	194-428	7	490	229-490
CG18555	690	144-335	6	435	229-435
CG30020	1309	140-995	9*	1095	229-1095
CG30431	418	232-403	6	415	229-415

**Table 2.** Primer sequences used to amplify ZnF domains from ZAD family members used in BSS analysis. Shown are the primers from the initial (**A**) and second (**B**) rounds.

A.			
CG7938-3' Sal	cacGTCGACctccgccgactctggcttc		
CG7938-5' Bam	cacGGATCCgaggaggagttcttcacc		
CG11695-3' Sal	gtgGTCGACcttctctgctgctgcagg		
CG11695-5' Bam	gtgGGATCCcacctgcacatgcacaacg		
CG12219-3' Sal	cacGTCGACcgagcgcgtggctcgcgatc		
CG12219-5' Bam	cacGGATCCaaagtggtagctccacaac		
CG17958-3' Sal	gtgCTCGAGtttcaagaccacgtcgtgat		
CG17958-5' Bam	gtgGGATCCgaggacaggccgacgaag		
CG30020-3' Sal	cacGTCGACggtgaagcgaatcgtgga		
CG30020-5' Bam	cacGGATCCgatgacgagcaatcgaagc		
B.			
CG1792-3' Sal	cacGTCGACcagtgccgectggggcgc	CG10267-3' Sal	cacGTCGACcatgetctcggccagc
CG1792-5' Bam	cacGGATCCgccctaaagcgggagcgcac	CG10267-5' Bam	cacGGATCCacaacatcggagcggcac
CG2711-3' Sal	gtgGTCGACgatgactgaacagcaacgg	CG10309-3' Sal	cacGTCGACcggagccgcgctttcc
CG2711-5' Bam	gtgGGATCCcgcgccgcgaccgctgg	CG10309-5' Bam	cacGGATCCcttgggagcgaagcgcc
CG3485-3' Sal	gtgGTCGACtattgacatttgaagggacg	CG10321-3' Xho	gtgCTCGAGgtctcgtgccactgatctg
CG3485-5' Bam	gtgGGATCCcacaaccgcgtccacacg	CG10321-5' Sma	gtgCCCGGGcagcaacctcgcctctgg
CG4148-3' Sal	gtgGTCGACggtttcaatcctctggagc	CG10366-3' Sal	cacGTCGACcgcgccttcttgagcagcc
CG4148-5' Bam	gtgGGATCCcatggtcccagagtctg	CG10366-5' Bam	cacGGATCCcggagggatgcagcagac
CG4730-3' Sal	cacGTCGACctccaggtcctccatctgc	CG12391-3' Sal	cacGTCGACgatgcgaacgctatcctcc
CG4730-5' Bam	cacGGATCCattgtgccacgcaatcggc	CG12391-5' Bam	cacGGATCCcagccctctcgaaggc
CG4820-3' Sal	cacGTCGACttttcttttctctcgcttc	CG14710-3' Sal	gtgGTCGACgctgggaatggtttgctccat
CG4820-5' Bam	cacGGATCCggtggaccgcaatggacc	CG14710-5' Bam	gtgGGATCCaacagccaaagtgcagg
CG7357-3' Sal	cacGTCGACcaccattaaccagcaagtggac	CG15436-3' Sal	gtgGTCGACcttgtgatctcgggatatg
CG7357-5' Bam	cacGGATCCaaacccaaggtcgcgctc	CG15436-5' Bam	gtgGGATCCgatgggaaaagcagcaag
CG7928-3' Sal	cacGTCGACaactggcgtgcagtcgtctc	CG18555-3' Sal	cacGTCGACagatatagccgcgttctc
CG7928-5' Bam	cacGGATCCaacgaaagctcggcgcaatac	CG18555-5' Bam	cacGGATCCcaatcgcacaagtgttcc
CG8145-3' Sal	cacGTCGACcgcagcaggtcagcggcttc	CG30431-3' Sal	cacGTCGACctcgcctctaactggggcc
CG8145-5' Bam	cacGGATCCctcgatcacgattacc	CG30431-5' Bam	cacGGATCCgcctctggtgcccgtgcac

**Table 3.** Consensus binding sequences identified for 23 ZAD family members.

<b>CG11695-5'-CRCACRTG-3'</b>	<b>CG7928-5'-YRCAGGG-3'</b>	<b>CG10321-5'-GGGGGTGGG-3'</b>
<b>CG17958-5'-RGTGTGGAG-3'</b>	<b>CG10267-5'-CCCATGGY-3'</b>	<b>CG15436-5'-CCCCTTGTRCCCC-3'</b>
<b>CG7938-5'-GGGTGCYR-3'</b>	<b>CG18555-5'-GCCACGRR-3'</b>	<b>CG10309-5'-GNGGGTGCNANYGTGGGG-3'</b>
<b>CG3485-5'-CATCCGTCTRR-3'</b>	<b>CG4820-5'-RCATGTGYYY-3'</b>	<b>CG12219-5'-CAxxGCARTGGGCCCCxC-3'</b>
<b>CG30020-5'-GGGCRCGG-3'</b>	<b>CG12391-5'-RCCACACGTCC-3'</b>	<b>CG2711-5'-TGRYCCYYCTG-3'</b>
<b>CG4148-5'-GATCCGTCTACCC-3'</b>	<b>CG8145-5'-CATTGTGG-3'</b>	<b>CG7357-5'-GCGGGTRAGGTGGCRGG-3'</b>
<b>CG10366-5'-RCGTGGGG-3'</b>	<b>CG4730-5'-TCACTRR-3'</b>	<b>CG14710-5'-GAGGAGTCATAG-3'</b>
<b>CG1792-5'-GTYGTGGC-3'</b>	<b>CG30431-5'-CRCTRRRCY-3'</b>	<b>CG4413</b>



**Table 4.** Oligonucleotide sequences used in competitive binding experiments for CG7938, CG17958, CG12219, CG30020, and CG11695.

CG7938 Consensus	5'	G G G C G G G T G C T G G A T	3'
CG7938 Consensus	3'	A T C C A G C A C C C G C C C	5'
CG7938 mutant	5'	G G G C G G C C G G T G G A T	3'
CG7938 mutant	3'	A T C C A C C G G C C G C C C	5'
CG11695 Consensus	5'	G G G C G C A C A T G T C G A T	3'
CG11695 Consensus	3'	A T C G A C A T G T G C G C C C	5'
CG11695 Mutant	5'	G G G C G A A C A G G T C G A T	3'
CG11695 Mutant	3'	A T C G A C C T G T T C G C C C	5'
CG12219 consensus	5'	G G G C G G T G T G G A G G A T	3'
CG12219 consensus	3'	A T C C T C C A C A C C G C C C	5'
CG12219 mutant	5'	G G G C G C C G C G T A C G A T	3'
CG12219 mutant	3'	A T C G T A C G C G G C G C C C	5'
CG17958 consensus	5'	C A G C G C A G T G G G C C C C A C	3'
CG17958 consensus	3'	G T G G G G C C C A C T G C G C T G	5'
CG17958 mutant	5'	A C G C T A C G G G G T C C C C A C	3'
CG17958 mutant	3'	G T G G G G A C C C C G T A G C G T	5'
CG30020 consensus	5'	G G G A G G G C A C G G G A T	3'
CG30020 consensus	3'	A T C C C G T G C C C T C C C	5'
CG30020 mutant	5'	G G G A T G G A A A G G G A T	3'
CG30020 mutant	3'	A T C C C T T T C C A T C C C	5'

**Table 5.** A quantification of predicted or known targets for ZAD proteins against homeobox containing early developmental genes. Predicted or known targets for ZAD proteins (top) against Homeobox and related genes (side).

	CG17958	CG7938	CG7928	CG10366	CG10309	CG8145	CG10267	CG3485	CG1792	CG30020	CG11695	CG12219
Bcd	1										1	
dfd		2										
antp		3				1					1	
eve		1	3	1							3	1
ftz	1	2	2							5	1	2
ac		3			1					5	2	
sc		1	3		1					1		
kr	1										2	
H		2										2
sxl		1								1	1	
aats-asp			1									
dally			1						1			
e2f2			2									
dap			1									
cg17508							1	1				
Ubx	4	5								5	17	8

**Table 6.** Primer sequences used to amplify ZAD domains from ZAD family members used in cofactor analysis.

Gene Name	Oligo Sequence	Enzyme
CG2889-3'	CAC CCC GGG CTC CTT CTT GAG TGA GCA C	Sma1
CG2889-5'	CAC GGA TCC ATG ATT TGT CGC CTT TGC C	BamH1
CG3941-3'	CAC GTC GAC ATA CTG GCG CAG GAG GGT C	Sal1
CG3941-5'	CAC GGA TCC AGG GTC TGC CGC TTC TGT C	BamH1
CG7938-3'	GTG GTC GAC TGC GTC CGC AAT CTG GCG	Sal1
CG7938-5'	GTG GGA TCC CCG TTT TGC TTC GTT TGC GG	BamH1
CG9233-3'	CAC GTC GAC CTG CTC CTG CAG GGA CTG TTC C	Sal1
CG9233-5'	CAC GAA TTC AGC ACT TGT CGC CTG TGC C	EcoR1
CG10108-3'	CAC GTC GAC ACT TTG GCT CTG GCT TTC GGC	Sal1
CG10108-5'	CAC GGA TCC CGC ACC TGC CTC ATC TGC	BamH1
CG11371-3'	CAC GTC GAC AAA GTC CCG AAA CAC GAG C	Sal1
CG11371-5'	CAC GGA TCC CAC TTG TGC CGC ATT TGC	BamH1
CG11695-3'	CAC GTC GAC CAG TTG CTG CAG GCC C	Sal1
CG11695-5'	CAC GGA TCC ATG ATA TGC CGC CTG TGC C	BamH1
CG12219-3'	CAC GTC GAC CTG GCT GCA CAG CGT CGT C	Sal1
CG12219-5'	CAC GGA TCC ATG ATC TGT CGC CTG TGT C	BamH1
CG17958-3'	GTG GTC GAC CTG GGT CGT CAG CCT CTT CTG G	Sal1
CG17958-5'	GTG GGA TCC GAT ACT TGC TTC TTC TGC GG	BamH1
CG30020-3'	CAC GTC GAC TTT CCG GCA CAG GAT CTC	Sal1
CG30020-5'	CAC GGA TCC TTG TCA TGT CGC TGC TGC	BamH1
CG33133-3'	CAC CTC GAG GGT GGT GGC GTA TAT CAC C	Xho1
CG33133-5'	CAC GAA TTC GAT ATC TGC CGC CTC TGT TTA C	EcoR1

## 6. BIBLIOGRAPHY

- Adryan, B., and Teichmann, S. A. The developmental expression dynamics of *Drosophila melanogaster* transcription factors. *Genome Biol* **11**, R40.
- Adryan, B., and Teichmann, S. A. (2006). FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics* **22**, 1532-3.
- Ayyanathan, K., Lechner, M. S., Bell, P., Maul, G. G., Schultz, D. C., Yamada, Y., Tanaka, K., Torigoe, K., and Rauscher, F. J., 3rd. (2003). Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. *Genes Dev* **17**, 1855-69.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Ewlinger, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002). The Pfam protein families database. *Nucleic Acids Res* **30**, 276-80.
- Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res* **32**, D138-41.
- Beerli, R. R., Dreier, B., and Barbas, C. F., 3rd. (2000). Positive and negative regulation of endogenous genes by designed transcription factors. *Proc Natl Acad Sci U S A* **97**, 1495-500.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. GenBank. *Nucleic Acids Res* **39**, D32-7.
- Brayer, K. J., Kulshreshtha, S., and Segal, D. J. (2008). The protein-binding potential of C2H2 zinc finger domains. *Cell Biochem Biophys* **51**, 9-19.
- Brayer, K. J., and Segal, D. J. (2008). Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem Biophys* **50**, 111-31.
- Breitkreutz, B. J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bahler, J., Wood, V., Dolinski, K., and Tyers, M. (2008). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**, D637-40.
- Brent, R., and Ptashne, M. (1985). A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* **43**, 729-36.
- Carninci, P., and Hayashizaki, Y. (2007). Noncoding RNA transcription beyond annotated genes. *Curr Opin Genet Dev* **17**, 139-44.

- Chen, B., Harms, E., Chu, T., Henrion, G., and Strickland, S. (2000). Completion of meiosis in *Drosophila* oocytes requires transcriptional control by grauzone, a new zinc finger protein. *Development* **127**, 1243-51.
- Choo, Y., Castellanos, A., Garcia-Hernandez, B., Sanchez-Garcia, I., and Klug, A. (1997). Promoter-specific activation of gene expression directed by bacteriophage-selected zinc fingers. *J Mol Biol* **273**, 525-32.
- Choo, Y., and Klug, A. (1994). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc Natl Acad Sci U S A* **91**, 11163-7.
- Chu, T., Henrion, G., Haegeli, V., and Strickland, S. (2001). Cortex, a *Drosophila* gene required to complete oocyte meiosis, is a member of the Cdc20/fizzy protein family. *Genesis* **29**, 141-52.
- Chung, H. R., Lohr, U., and Jackle, H. (2007). Lineage-specific expansion of the zinc finger associated domain ZAD. *Mol Biol Evol* **24**, 1934-43.
- Chung, H. R., Schafer, U., Jackle, H., and Bohm, S. (2002). Genomic expansion and clustering of ZAD-containing C2H2 zinc-finger genes in *Drosophila*. *EMBO Rep* **3**, 1158-62.
- Collins, T., Stone, J. R., and Williams, A. J. (2001). All in the family: the BTB/POZ, KRAB, and SCAN domains. *Mol Cell Biol* **21**, 3609-15.
- Consortium, I. H. G. S. (2004). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-45.
- Crozatier, M., Kongsuwan, K., Ferrer, P., Merriam, J. R., Lengyel, J. A., and Vincent, A. (1992). Single amino acid exchanges in separate domains of the *Drosophila* serendipity delta zinc finger protein cause embryonic and sex biased lethality. *Genetics* **131**, 905-16.
- Curwen, V., Eyraas, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M., and Clamp, M. (2004). The Ensembl automatic gene annotation system. *Genome Res* **14**, 942-50.
- Desbarats, L., Gaubatz, S., and Eilers, M. (1996). Discrimination between different E-box-binding proteins at an endogenous target gene of c-myc. *Genes Dev* **10**, 447-60.
- Desjarlais, J. R., and Berg, J. M. (1992). Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc Natl Acad Sci U S A* **89**, 7345-9.
- Dreier, B., Beerli, R. R., Segal, D. J., Flippin, J. D., and Barbas, C. F., 3rd. (2001). Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. *J Biol Chem* **276**, 29466-78.
- Drysdale, R. (2008). FlyBase : a database for the *Drosophila* research community. *Methods Mol Biol* **420**, 45-59.
- Duan, J., Xia, Q., Cheng, D., Zha, X., Zhao, P., and Xiang, Z. (2008). Species-specific expansion of C2H2 zinc-finger genes and their expression profiles in silkworm, *Bombyx mori*. *Insect Biochem Mol Biol* **38**, 1121-9.

- Fairall, L., Schwabe, J. W., Chapman, L., Finch, J. T., and Rhodes, D. (1993). The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* **366**, 483-7.
- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucleic Acids Res* **30**, 235-8.
- Fu, F., Sander, J. D., Maeder, M., Thibodeau-Beganny, S., Joung, J. K., Dobbs, D., Miller, L., and Voytas, D. F. (2009). Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays. *Nucleic Acids Res* **37**, D279-83.
- Gaszner, M., Vazquez, J., and Schedl, P. (1999). The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction. *Genes Dev* **13**, 2098-107.
- Gibert, J. M., Marcellini, S., David, J. R., Schlotterer, C., and Simpson, P. (2005). A major bristle QTL from a selected population of *Drosophila* uncovers the zinc-finger transcription factor *poils-au-dos*, a repressor of *achaete-scute*. *Dev Biol* **288**, 194-205.
- Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carroll, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., Jr., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727-36.
- Gogos, J. A., Hsu, T., Bolton, J., and Kafatos, F. C. (1992). Sequence discrimination by alternatively spliced isoforms of a DNA binding zinc finger domain. *Science* **257**, 1951-5.
- Greisman, H. A., and Pabo, C. O. (1997). A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* **275**, 657-61.
- Haerty, W., Artieri, C., Khezri, N., Singh, R. S., and Gupta, B. P. (2008). Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution. *BMC Genomics* **9**, 399.
- Hamilton, A. T., Huntley, S., Kim, J., Branscomb, E., and Stubbs, L. (2003). Lineage-specific expansion of KRAB zinc-finger transcription factor genes: implications for the evolution of vertebrate regulatory networks. *Cold Spring Harb Symp Quant Biol* **68**, 131-40.
- Harms, E., Chu, T., Henrion, G., and Strickland, S. (2000). The only function of *Grauzone* required for *Drosophila* oocyte meiosis is transcriptional activation of the *cortex* gene. *Genetics* **155**, 1831-9.
- Huntley, S., Baggott, D. M., Hamilton, A. T., Tran-Gyamfi, M., Yang, S., Kim, J., Gordon, L., Branscomb, E., and Stubbs, L. (2006). A comprehensive catalog of

- human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res* **16**, 669-77.
- Isalan, M., Klug, A., and Choo, Y. (1998). Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry* **37**, 12026-33.
- Jauch, R., Bourenkov, G. P., Chung, H. R., Urlaub, H., Reidt, U., Jackle, H., and Wahl, M. C. (2003). The zinc finger-associated domain of the Drosophila transcription factor grauzone is a novel zinc-coordinating protein-protein interaction module. *Structure* **11**, 1393-402.
- Klug, A. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu Rev Biochem* **79**, 213-31.
- Klug, A. The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. *Q Rev Biophys* **43**, 1-21.
- Klug, A., and Rhodes, D. (1987). Zinc fingers: a novel protein fold for nucleic acid recognition. *Cold Spring Harb Symp Quant Biol* **52**, 473-82.
- Krishna, S. S., Majumdar, I., and Grishin, N. V. (2003). Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res* **31**, 532-50.
- Krystel, J., Anderson, D., and Ayyanathan, K. (2009). A Database of Zinc-finger Associated Domain Containing Zinc Finger Proteins in Drosophila Melanogaster.
- Liu, P. Q., Rebar, E. J., Zhang, L., Liu, Q., Jamieson, A. C., Liang, Y., Qi, H., Li, P. X., Chen, B., Mendel, M. C., Zhong, X., Lee, Y. L., Eisenberg, S. P., Spratt, S. K., Case, C. C., and Wolffe, A. P. (2001). Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. Activation of vascular endothelial growth factor A. *J Biol Chem* **276**, 11323-34.
- Lodish, H., Berk, A., Matsudaira, P., Kaiser, C., Krieger, M., Scott, M., Zipursky, A., Darnell, J. (2004). Molecular Cell Biology.
- Mackay, J. P., and Crossley, M. (1998). Zinc fingers are sticking together. *Trends Biochem Sci* **23**, 1-4.
- Malik, S., Huang, C. F., and Schmidt, J. (1995). The role of the CANNTG promoter element (E box) and the myocyte-enhancer-binding-factor-2 (MEF-2) site in the transcriptional regulation of the chick myogenin gene. *Eur J Biochem* **230**, 88-96.
- McCarty, A. S., Kleiger, G., Eisenberg, D., and Smale, S. T. (2003). Selective dimerization of a C2H2 zinc finger subfamily. *Mol Cell* **11**, 459-70.
- Miller, J., McLachlan, A. D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *Embo J* **4**, 1609-14.
- Mingot, J. M., Vega, S., Maestro, B., Sanz, J. M., and Nieto, M. A. (2009). Characterization of Snail nuclear import pathways as representatives of C2H2 zinc finger transcription factors. *J Cell Sci* **122**, 1452-60.
- Morse, R. H. (2007). Transcription factor access to promoter elements. *J Cell Biochem* **102**, 560-70.
- Naidu, P. S., Ludolph, D. C., To, R. Q., Hinterberger, T. J., and Konieczny, S. F. (1995). Myogenin and MEF2 function synergistically to activate the MRF4 promoter during myogenesis. *Mol Cell Biol* **15**, 2707-18.
- Payre, F., Buono, P., Vanzo, N., and Vincent, A. (1997). Two types of zinc fingers are required for dimerization of the serendipity delta transcriptional activator. *Mol Cell Biol* **17**, 3137-45.

- Payre, F., Crozatier, M., and Vincent, A. (1994). Direct control of transcription of the *Drosophila* morphogen bicoid by the serendipity delta zinc finger protein, as revealed by in vivo analysis of a finger swap. *Genes Dev* **8**, 2718-28.
- Payre, F., and Vincent, A. (1991). Genomic targets of the serendipity beta and delta zinc finger proteins and their respective DNA recognition sites. *Embo J* **10**, 2533-41.
- Peng, H., Zheng, L., Lee, W. H., Rux, J. J., and Rauscher, F. J., 3rd. (2002). A common DNA-binding site for SZF1 and the BRCA1-associated zinc finger protein, ZBRK1. *Cancer Res* **62**, 3773-81.
- Reiss DJ, Mobley HL. Determination of target sequence bound by PapX, repressor of bacterial motility, in flhD promoter using systematic evolution of ligands by exponential enrichment (SELEX) and high throughput sequencing. *J Biol Chem* 2011;286:44726–44738.
- Rosenfeld, R., and Margalit, H. (1993). Zinc fingers: conserved properties that can distinguish between spurious and actual DNA-binding motifs. *J Biomol Struct Dyn* **11**, 557-70.
- Roussigne, M., Kossida, S., Lavigne, A. C., Clouaire, T., Ecochard, V., Glories, A., Amalric, F., and Girard, J. P. (2003). The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem Sci* **28**, 66-9.
- Ryan, R. F., Schultz, D. C., Ayyanathan, K., Singh, P. B., Friedman, J. R., Fredericks, W. J., and Rauscher, F. J., 3rd. (1999). KAP-1 corepressor protein interacts and colocalizes with heterochromatic and euchromatic HP1 proteins: a potential role for Kruppel-associated box-zinc finger proteins in heterochromatin-mediated gene silencing. *Mol Cell Biol* **19**, 4366-78.
- Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Edwards, C., Fan, J. B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matise, T. C., McKusick, K. B., Morissette, J., Mungall, A., Muselet, D., Nusbaum, H. C., Page, D. C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, C., She, X., Silva, J., Slonim, D. K., Soderlund, C., Sun, W. L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M. D., Auffray, C., Walter, N. A., Brandon, R., Dehejia, A., Goodfellow, P. N., Houlgatte, R., Hudson, J. R., Jr., Ide, S. E., Iorio, K. R., Lee, W. Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M. H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J. C., Sikela, J. M., Beckmann, J. S., Weissenbach, J., Myers, R. M., Cox, D. R., James, M. R., et al. (1996). A gene map of the human genome. *Science* **274**, 540-6.
- Sekido, R., Murai, K., Funahashi, J., Kamachi, Y., Fujisawa-Sehara, A., Nabeshima, Y., and Kondoh, H. (1994). The delta-crystallin enhancer-binding protein delta EF1 is a repressor of E2-box-mediated gene activation. *Mol Cell Biol* **14**, 5692-700.



- Swirnoff, A. H., and Milbrandt, J. (1995). DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol Cell Biol* **15**, 2275-87.
- Thiesen, H. J., and Bach, C. (1990). Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucleic Acids Res* **18**, 3203-9.
- Thorvaldsen, J. L., Sewell, A. K., McCowen, C. L., and Winge, D. R. (1993). Regulation of metallothionein genes by the ACE1 and AMT1 transcription factors. *J Biol Chem* **268**, 12512-8.
- Wang J, Lu J, Gu G, Liu Y. In vitro DNA-binding profile of transcription factors: methods and new insights. *J Endocrinol* 2011;210:15–27 . Witte, M. M., and Dickson, R. C. (1990). The C6 zinc finger and adjacent amino acids determine DNA-binding specificity and affinity in the yeast activator proteins LAC9 and PPR1. *Mol Cell Biol* **10**, 5128-37.
- Wu, H., Yang, W. P., and Barbas, C. F., 3rd. (1995). Building zinc fingers by selection: toward a therapeutic application. *Proc Natl Acad Sci U S A* **92**, 344-8.