CHARACTERIZATION OF SNAG-ZINC FINGER PROTEIN (ZFP) TRANSCRIPTION FACTORS

by

Cindy Chung-Yue Chiang

A Thesis Submitted to the Faculty of

The Charles E. Schmidt College of Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

Florida Atlantic University

Boca Raton, Florida

May 2009

CHARACTERIZATION OF SNAG-ZINC FINGER PROTEIN (ZFP) TRANSCRIPTION FACTORS

by

Cindy Chung-Yue Chiang

This thesis was prepared under the direction of the candidate's thesis advisor, Dr. Kasirajan Ayyanathan, Department of Biological Sciences, and has been approved by the members of her 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 Master of Science.

SUPERVISORY COMMITTEE:

Kasirajan Ayyanathan, Ph.D.
Thesis Advisor

Ramaswamy Narayanan, Ph.D.

James X. Hartmann, Ph.D.

Dr. Rodney K. Murphey, Ph.D.
Chair, Department of Biological Sciences

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

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

April 7, 2009
Date

ACKNOWLEDGEMENTS

The author wishes to express her thanks to her advisor Dr. Kasirajan Ayyanathan for his continued support and faith in me and my work as well as the time and effort put into this project, and to her committee members Dr. Ramaswamy Narayanan and Dr. James Hartmann for their time, knowledge, insight, materials, and ideas on how to better improve further her research. The author would also like to thank the other members of the Ayyanathan lab as well as the technicians at Florida Atlantic University for their assistance and support. Special thanks go to the author's family and friends, especially her parents who have continuously supported, encouraged, and allowed her to pursue this degree and further her education.

ABSTRACT

Author:	Cindy Chiang
Title:	Characterization of SNAG-Zinc Finger Protein (ZFP) Transcription Factors
Institution:	Florida Atlantic University
Thesis Advisor:	Dr. Kasirajan Ayyanathan
Degree:	Master of Science
Year:	2009

Transcriptional regulation is an important area of research due to the fact that it leads to gene expression. Transcription factors associated with the regulation can either be activators or repressors of target genes, acting directly or with the aid of other factors. A majority of transcriptional repressors are zinc finger proteins (ZFPs) which bind to specific DNA sequences. The Snail/Gfi (SNAG) domain family, with members such as Slug, Smuc, Snail, and Scratch, are transcriptional repressors shown to play a role in various diseases such as cancer. The SNAG transcription factors contain a conserved SNAG repression domain and DNA binding domain zinc fingers. The specific DNA sequences to which each SNAG-ZFP binds, as well as a general consensus -TGCACCTGTCCGA, have been determined. Also, putative protein-protein

interactions in which the Slug domain participates has been identified via binding assays. All these results contribute to better understanding of SNAG-ZFP functions.

CHARACTERIZATION OF SNAG-ZINC FINGER PROTEIN (ZFP)

TRANSCRIPTION FACTORS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Transcription Factors and Transcriptional Regulation

Transcription, the synthesis of RNA transcript from a DNA template, is an extremely important event in all aspects of biology. It leads to the expression of genes via an mRNA intermediate. Because of this, transcriptional regulation by transcription factors is an important field that requires in-depth research. By determining the mechanisms and interactions that cause genes to be expressed or repressed in a spatio-temporal fashion, a better understanding of its effects in a biological system may be reached. This includes obtaining a greater knowledge of how genes can affect development or diseases like cancer.

Eukaryotic genomes contain several thousand genes. These genes are transcribed with the help of regulatory genes known as transcription factors that can activate or silence the target genes when necessary. The human genome is comprised of approximately 30,000 genes, of which approximately 3,000 code for transcription factors that function in the regulation of the expression of the remaining 27,000 genes (Consortium, 2004; Lodish, 2004).

Transcriptional regulation mainly occurs at the level of control of transcription initiation of the target gene and is controlled by transcription factors, which function as either activators or repressors (Jackson et al., 2008; Okkema and Krause, 2005).

Transcriptional repressors may function by directly interacting with the components of the basal transcription machinery, remodeling chromatin in an ATP-dependent manner, or site-specific modifications of the histones (Berger, 2007; Tsukiyama et al., 1995; Tsukiyama and Wu, 1995). Binding of the transcriptional repressors to a particular DNA sequence can inhibit transcription in many ways. First, it could block the function of an activator termed active repression (Hanna-Rose and Hansen, 1996). The other mechanisms involve chromatin compaction leading to the loss of formation of the polymerase II holoenzyme complex at the promoter, which is needed to transcribe the DNA into RNA (Fondell et al., 1996; Goldmark et al., 2000). The repressors can inhibit the formation of the complex directly or by recruiting a co-factor to repress indirectly (Ayyanathan et al., 2003; Ayyanathan et al., 2007; Hanna-Rose and Hansen, 1996). In any case, transcription will be disrupted. In case of transcriptional repressors that are directly associated with causing a disease (for example involvement of Snail transcription factor in cancer metastasis), if this repression is blocked, the progression of a disease could ultimately be inhibited (Batlle et al., 2000; Cano et al., 2000).

C_2H_2 Zinc Finger Proteins

Majority of transcriptional repressors are zinc finger proteins which contain sequence-specific DNA binding zinc finger motifs. The binding of the zinc fingers to the upstream regulatory region in a DNA sequence-specific manner can selectively inhibit the transcription of the target gene. About one-third of the ~3,000 regulatory genes code for Cys_2His_2 (C_2H_2) zinc finger proteins (Huntley et al., 2006), named so for the specific zinc chelating domain they possess which consists of a pair of cysteine

residues in the beta helix and a pair of histidine residues in the alpha helix. These zinc fingers function as DNA binding domains that allow for the interaction of the DNA's major groove with the divalent cation (Miller et al., 1985). The zinc ion provides stability to the binding domain as the absence of it would cause the domain to unfold.

The C_2H_2 family of transcriptional repressors is classified into sub-families such as SNAG, BTB/POZ, (**B**road Complex, **T**ramtrack, **B**ric-a-brac/**Po**x virus **Z**inc finger) and KRAB (**Kr**uppel **A**ssociated **B**ox) based on the repressor domains they contain (Collins et al., 2001). There are around 40 members in the SNAG domain family, the most studied and well-known of which are Snail, Slug, and Gfi-1 (Nieto, 2002). Many transcription factors, regulatory proteins, and other proteins that interact with DNA contain these protein domains. The zinc fingers become a structural motif when they interact with the DNA, creating a binding domain.

SNAG (Snail/Gfi-1) Domain Zinc Finger Proteins (ZFPs)

The name SNAG comes from Snail/Gfi-1, two of the founding members of the family. The Snail superfamily of zinc finger transcription factors has been found to consist of two independent families, Snail and Scratch (Manzanares et al., 2001). The transcriptional repressors in this family all contain a SNAG domain, which was first discovered in Gfi-1 (Growth Factor Independence-1) (Grimes et al., 1996; Zweidler-Mckay et al., 1996), and is a highly conserved 21-amino acid sequence that mediates transcriptional repression. Each of these transcription factors has the SNAG domain at the N-terminus and a variable number of C_2H_2 zinc fingers (four to six in number) towards the C-terminus. The sequence is quite similar between numerous members of

SNAG-domain transcription factors in various organisms. There is another well-conserved domain whose function is unknown that occurs just prior to the zinc finger tandem repeats in Slug and Scratch.

Although the function of SNAG domain transcriptional repressors has been studied, the molecular mechanisms by which they act remained largely unknown, recent research has increased the information in this area. Ayyanathan et al. first determined that the SNAG domain recruits and functions with a corepressor called Ajuba, a multiple LIM domain protein. The Ajuba protein shuttles between the cytoplasm and the nucleus, and assembles repression complexes at target promoters (Ayyanathan et al., 2007). This observation has been extended in further studies by Hou et al. which showed that protein arginine methyltransferase 5 (PRMT5) is an effector recruited by Ajuba to mediate Snail-dependent repression (Hou et al., 2008).

Snail, Slug, and Scratch are examples of transcription factors that contain the SNAG domain and have roles in development and disease (Nieto, 2002). In *Drosophila,* mesoderm differentiation and development of the nervous system require a functional SNAG domain (Hemavathy et al., 2004). It has been discovered that SNAG-ZFPs play a role in human disease like cancer. For example, Snail ZFPs target E-cadherin promoters, which will result in the loss of E-cadherin, an important gene in cell adhesion, and will ultimately lead to highly malignant and invasive human tumor progression (Batlle et al., 2000). E-cadherin plays a role in epithelial-mesenchymal transition, and its repression by Snail has been established in both murine and human cell lines, tumors (Cano et al., 2000), and in early development of mouse and *Drosophila* (Nieto, 2002). SLUG, which is expressed in chick and *Xenopus* embryo

EMT regions, have been found to also be an E-cadherin repressor (Laux et al., 2004). Slug and Snail have also been shown to be highly expressed in invasive human breast tumors (Blanco et al., 2002), acting as repressors of Occludin and Claudin-1, integral membrane proteins present in tight junctions (Martinez-Estrada et al., 2006). Finding other genes repressed by these transcription factors is important for characterizing this group of repressors and potentially contributing to the knowledge of development and disease profiles.

DNA Binding Sites and E-box

As previously stated, binding of zinc fingers to a DNA sequence creates a motif during the interaction that makes the structure more stable. These motifs direct specific binding to DNA (Nagai et al., 1988). The specific binding is a good determinant as to what the protein can bind to.

A similar experiment was done and found that a consensus sequence with the highly conserved core CAGGTG was found for the *Drosophila* transcription factor SNAIL (Mauhin et al., 1993). They also found surrounding nucleotides to make up a longer consensus flanking the core E-box, which is also a goal of this research. Other binding sites have been identified for zinc finger proteins such as IA-1 (Breslin et al., 2003; Breslin et al., 2002) and ZBRK1 (Peng et al., 2002).

Transcription factors such as Snail and Scratch bind to E-box elements of E-cadherin's proximal promoter site (Cano et al., 2000; Giroldi et al., 1997). This E-box can regulate basic helix-loop-helix transcription factors as does Smuc (Kataoka et al., 2000), and can itself be regulated by other transcription factors, as is the case in IA-1

(Breslin et al., 2003). The E-box element found within the IA-1 promoter can be regulated by NeuroD1 and E47 heterodimers, both transcription factors critical to start transcription of the IA-1 gene (Breslin et al., 2003).

There are other E-box binding transcription factor families that have different functions from that of the SNAG family. The binding is known in various human muscle development transcription factors with basic helix-loop-helix motif structures such as the human muscle regulatory proteins Myf3-Myf6 (Braun and Arnold, 1991), MyoD, a muscle myogenic determination gene (Davis et al., 1990), and myogenin, a muscle-specific factor that participates in muscle differentiation (Brennan and Olson, 1990). Separate studies have shown that the regulatory proteins Myf3-Myf6 bind to the DNA sequence GCAGGTG, while myogenin binds to a CANNTG E-box such as the two found within the enhancer of muscle creatine kinase.

By identifying a longer consensus sequence for the SNAG-ZFPs or a sequence similar to E-box with nucleotides acting in conjunction with the region, more information can be derived on what genes are being regulated by these transcription factors, what else the E-box regulates, and ultimately the mechanisms behind it all. We will also be able to differentiate one family of E-box binding transcription factors from others through the actual consensus sequences found for the SNAG-ZFs. This will separate one group of E-box binding factors from other unrelated ones and prove significance between families of transcription factors based on their binding sites.

The main goals of this research are two-fold. The first is to identify sequence specific DNA binding sites of these SNAG transcription factors, further characterizing the way in which they function. By identifying the sequences to which the zinc fingers

bind, as well as discovering the interactions they are involved in, will further contribute to the information known about these transcriptional repressors, which could be potentially vital information to decoding the mechanisms by which these factors act in various diseases. The second goal is to unravel the novel functions associated with the conserved Slug and Scratch domains by identifying and characterizing interaction proteins.

Experimental Design and Data

The transcriptional regulation of genes is important in many aspects and can control the progression or development of a disease. Previous studies have shown the significance of SNAG domain transcriptional repressors in various diseases and the potential target genes involved.

The focus of this study is two-fold. The first goal is to decipher the specific DNA sequences that are recognized by each of the individual SNAG zinc finger proteins, and to ultimately find the common sequences, ranging in nucleotide sizes, that are identified for each transcriptional repressor. The comprehensive information obtained from this current research will add to the known information about SNAG-ZFPs and their target genes via unique DNA sequences not previously discovered or analyzed. The second aim is to determine interactions that occur between proteins with the Slug and Scratch SNAG-ZFPs by extending this project into human cell lines. Specifically, efforts will be made to identify novel protein partners that interact with Slug and Scratch domains represented in the diagrammatic representation of the structure of SNAG ZFPs presented in **Figure 1.** These aims will potentially allow for

the development of a novel model of repression involving a SNAG transcription factor, identify their target gene(s), and establish protein-protein interactions that occur.

IDENTIFICATION OF UNIQUE DNA-BINDING SPECIFICITIES FOR THE SNAG-TRANSCRIPTION FACTOR FAMILY MEMBERS

Bacterial Expression of GST-SNAG Fusion Proteins

In order to express these mammalian genes in a bacterial system, GST-SNAG constructs were made. The zinc finger constructs (given in amino acids position) made were as follows: Slug (121-268 aa), Smuc (152-292 aa), Snail long (144-264 aa), Snail short (154-259 aa), Scratch long (181-337 aa), Scratch short (191-327 aa), Gfi-1 (245-422 aa), Gfi-1B Hs (human) (153-330 aa), Gfi-1B Mm (mouse) (153-330 aa), insulinoma-associated 1 (IA-1) long (290-521 aa), and IA-1 short (300-508 aa). GST-fusion construction domains for Slug (84-127 aa) and Scratch that precede the zinc fingers were also produced. For Scratch, short (107-126 aa) and long (102-131aa) constructs were made, the short being just the zinc fingers alone, while the long were the zinc fingers with five additional amino acids on either end in order to facilitate proper folding. The expression constructs are illustrated in **Figure 2.**

Briefly, recombinant constructs were created by fusing the DNA fragments coding for the corresponding zinc finger motifs with a Glutathione-S-transferase tag in the pGEX-4T2 plasmid vector for bacterial expression in competent BL21 *Escherichia coli* cells. The ligations were transformed into *E. coli* DH5α cells and grown on LB agar plates with ampicillin. Colonies from these plates were grown in LB broth with

ampicillin for mini-plasmid DNA preps. The plasmids were checked for the correct insert size by restriction endonuclease digestion and gel electrophoresis. The positive recombinant plasmids were then transformed into the *E. coli* BL21 host for protein expression. The *E. coli* was grown in LB containing ampicillin and kanamycin and induced to express the proteins using 100 mM IPTG at 30°C. A small aliquot pre-induction was saved as the uninduced sample. A maxi-protein preparation was done on the 500 mL induced culture. The induction of protein expression was verified by visualization on a sodium dodecyl sulfate-polyacrylamide gel (SDS-PAGE) as illustrated in **Figure 3**.

Protein Purification

Once the expression level was found satisfactory, these GST-SNAG fusion proteins were purified from other *E. coli* proteins by using homogenized, soluble fractions. The appropriate conditions for each ZFP were determined. Due to their different natures, each protein behaved differently and had to be solubilized under different conditions. While most of the other proteins were expressed in soluble fraction to a considerable extent, the GST-GFI-1 proteins were more difficult and required urea for solubilization followed by renaturation.

The resuspended pellets were sonicated to disrupt the cell membrane and release the contents of the cell. Six to eight one-minute pulses were routinely carried out. The sonicated samples were first spun at 4°C at 21,000 g for 20 minutes, and then the supernatant was then spun at 4°C at 100,000 g for 30 minutes in an ultracentrifuge to obtain clear soluble fraction for chromatography. When needed, as in case of GST-

GFI-1 and GST-IA-1 proteins, the supernatant was collected and set aside for protein refolding, due to the fact that the proteins were no longer in their original conformation and must be refolded back to their appropriate shape. Refolding was done at 4°C and involved a refolding buffer (100 mM Tris Cl (pH 8.0), 20% glycerol, 10% sucrose) plus urea and water. The ultracentrifuged supernatant was added to the refolding buffer in a drop-wise manner as a stir bar continually mixed the sample into solution. The supernatant was $1/10^{th}$ the total volume of refolding buffer and sample combined. The refolded proteins were passed through GSH sepharose affinity chromatography column. The beads were washed with 5 mL of 1x PBS and collected. Next, the bound proteins were eluted off the column three times in 300 μL of elution buffer, consisting of 10 mM glutathione (GSH), 0.1% Triton X-100, and 50 mM Tris Cl (pH 8.0). Aliquots of all the samples were analyzed on SDS-PAGE gels to confirm the presence of and determine the amount of proteins present in each fraction. **Figure 4** illustrates the analysis of purification profiles of the GST-SNAG ZFs.

When significant amounts of purified, homogeneous proteins were obtained, all elutions were consolidated for dialysis to concentrate the protein and remove the GSH present in the elutions. These samples were loaded on an SDS-PAGE gel to determine the purity of all the recombinant proteins. This gel is shown in **Figure 5**. Protein estimation using the Bradford reagent assay was done to assess the concentration of each protein so that the appropriate amounts that would result in equal binding reactions could be used.

Zinc-Finger Array Binding Site Interactions

The purified proteins were tested using **E**lectrophoretic **M**obility **S**hift **A**ssay (EMSA) for their ability to bind a known DNA recognition site from the E-cadherin promoter. Radiolabeled E-box promoter (E-cadherin's promoter) sequence was bound to the purified recombinant GST-ZFPs to determine if binding would take place as expected. The results as presented in **Figure 6**, showed that Slug, Smuc, Snail, and Scratch had expected interactions as indicated by bands that shifted, but the GFI-1 samples did not. This corresponds to the literature that E-cadherin is repressed by the ZFPs tested except for the GFIs (Laux et al., 2004; Nieto, 2002), which does not bind to E-box motifs.

Next, DNA binding site selection experiments were carried out using recombinant zinc finger proteins and a randomized oligonucleotide library, essentially as described (Peng et al., 2002). The oligonucleotide (18N) library contained flanking sequences at the 5' end composed of a BamHI site and 3' end composed of an EcoRI site. The full sequence: 5'-agacGGATCCattgca-NNNNNNNNNNNNNNNNN-ctgtccGAATTCgga-3' (total length 49 nucleotides), with the restriction sites being those nucleotides which are underlined. Two oligonucleotides complementary to these flanking sequences were used to amplify the oligonucleotide library. The amplified oligonucleotide library was end-labeled with T4 polynucleotide kinase and ^{32}P for 1 hour at 37°C and was then bound to the proteins for 15 minutes at room temperature. The binding samples were subjected to EMSA at 4°C, transferred onto a sheet of Whatman filter paper and dried using a vacuum pump. The dried gel was subjected to

autoradiography in an -80°C freezer. The resulting autoradiograph is presented in **Figure 7**.

The gel bands corresponding to the signals were excised from the filter paper and placed in elution buffer comprised of 0.5 mM ammonium acetate (NH_4OAc) and 1 mM EDTA in a 37°C water bath overnight to elute the DNA-protein complexes. The purified DNA was subjected to polymerase chain reaction (PCR) to amplify the molecules present. PCR conditions were 94°C for 30 seconds, 48°C for 1 minute, then 72°C for 20 seconds, for 40 cycles. These DNA samples were visualized on a DNA-PAGE gel which revealed bands around the expected size of 49 base pairs. A portion of the amplified PCR product was radiolabeled and bound to fresh proteins, in order to enrich the sequences collected. Totally four cycles were carried out to obtain authentic binding sites. The binding site selections and their subsequent amplified products of the second, third, and final can be seen in **Figures 8, 9, and 10,** respectively. Finally, the enriched PCR products (putative binding sites) were subjected to proteinase K digestion, purified, and electrophoresed on a DNA-PAGE.

Cloning of DNA-Protein Binding Site Interactions

The gel slices containing the resultant PCR product bands were excised from the DNA-PAGE gel shown in **Figure 10 (panel B)**, and subjected to electroelutions in dialyis bag. This was done by applying 45 V for 2.5 hours and then reversing the polarity for 30 seconds at 100 V. These electro-eluted samples were extensively purified and used for cloning.

The plasmid vector used was pUC18, which was double digested along with the samples to ensure efficiency. The initial digestions with each respective endonuclease can be seen in **Figure 11 (panel A)**. The follow-up digestion with the alternate restriction enzyme is seen in **Figure 11 (panel B)**. The samples were first digested with EcoRI for 1 hour at 37°C, and then with BamHI for an additional hour at 37°C. The plasmid vector and insert DNAs were ligated and transformed into DH5α cells. Two LB/amp plates were used to grow up colonies of each sample, one at 50 μL, the other at 200 μL, to ensure proper growth of transformed cells. These plates were incubated at 37°C overnight. Numerous colonies were obtained, and at least twenty colonies from each cloning sample were taken for inoculation in 5mL LB/amp media. These were then used for mini-plasmid preps, which were analyzed on DNA-PAGE gels to identify positive recombinant clones. A sample of the results is presented in **Figure 12**. These mini-plasmid preps were further processed by using Calf Intestinal Alkaline Phosphatase (CIP) to remove the 5' phosphate group from the DNA to prevent self-ligation. The samples were then purified by PCI/CI and used for the subsequent procedure in which binding of these samples was carried out with cognate protein (to monitor specific binding) and unrelated protein (to monitor non-specific binding).

Binding and Competition EMSAs

With the CIP/PCI/CI purified samples, further gel shifts were performed to confirm binding of these isolated DNA sequences to the SNAG-ZFP that was used to obtain the sequences. The DNA was kinased using ^{32}P for 1 hour at 37°C and was then bound to the proteins for 15 minutes at room temperature, similar to the previous

procedures presented. The purified protein of each SNAG-ZF used in this experiment was normalized to a concentration of 300 ng. After running the gel shifts, exposing, and developing the film, results such as those presented in **Figure 13** were obtained. Bands indicating binding between DNA and protein were seen in each case. Additional studies were carried out, looking at the interaction between the DNA and some other unrelated protein. The protein used was Serendipity locus protein beta (CG7938), which is found in *Drosophila melanogaster*. By having this competing protein bind to the DNA, specificity of binding is assessed. Results of these EMSAs can be seen in **Figure 13**. A regular binding with the right protein matched up to its corresponding DNA and then a binding of CG7938 along with the obtained DNA was done for each protein, providing a set of autoradiographs for analysis of each SNAG-ZFP.

DNA Sequencing

For DNA sequencing, the ZFPs processed were Slug, Smuc, Snail long and short, and Scratch long and short. The number of binding site-selected positive clones processed and sequenced ranged from 20-26 in number for each of the above mentioned ZFPs.

The DNA templates for sequencing were prepared using the mini-plasmid prep of positive recombinant clones and digesting them with RNase for 1 hour. PCI/CI cleaning was performed after digestion. These samples were run on an agarose gel to confirm the integrity and purity of DNA. A picture of one such gel can be seen in **Figure 14**. Aliquots of RNAse-digested mini-plasmid preps were compiled into a 96-well microtiter plate to be sequenced. For those that showed strong bands, 2.5 μL of the

sample and 2.5 μL of TE buffer were used. For the rest of the samples, 5 μL was added to the wells.

The microtiter plates were sent (for cycle sequencing) to ICBR Genomics Core (University of Florida, Gainesville) in order to obtain the results of the binding site sequences which were then sent back. Each set of sequences was analyzed together to determine if there were any consensus sequences similar between the sequences. The sequences were identified from the 5' BamHI restriction site GGATCCATTGCA and the 3' EcoRI restriction site CTGTCCGAATC. The resultant sequences in between these restriction sites were collected and compared.

Thorough visual analysis was done to determine the best fit consensus sequence found in the majority, if not all, of the DNA sequences obtained. Nucleotides were similarly matched up between all of the sequences. Aberrant sequences such as those provided when empty vector was present or when partial sequences were obtained were unused. On the whole, most of the sequences were used in the analysis. Tables indicating the sequences obtained are located in **Tables 1-6**. Also provided is a diagrammatic representation of the percentages of nucleotide occurrence of each based on the consensus obtained.

Discussion

The results of the aforementioned procedures led to the final autoradiographs, DNA sequencing comparisons, and consensus sequences for each SNAG-ZFP. These binding sites of each protein was determined and visualized on the autoradiographs which show the binding results of the DNA.

Binding of the 18N DNA with the cognate protein resulted in some obvious shifts seen on the Smuc protein and DNA autoradiographs exhibited in **Figure 13 (Panels A and C)**. These figures correspond with the gels run with the competing, nonspecific CG7938 protein seen in **Figure 13 (Panels B and D)**. As expected, where there were once bands where the Smuc DNA and Smuc protein bound to each other, specificity of the binding site denies the interaction of the Smuc DNA with CG7938 protein. The strong bands that have shifted to a lower position on the gel correspond to the size of the appropriate control.

The known SNAG-binding E-box consensus sequence of CANNTG was found in at least one 18N sequence of each set of sequences and in at least nine full-length sequences of each set (due to the sharing of nucleotides in the BamHI and EcoRI cut sites). Several of the sequences contained no binding site and the result was the empty vector sequence GGATCCCCGGCTACCAAGCTCGAATTC. Others identified no sequence or empty vector.

In order to derive a consensus sequence for each SNAG-ZFP, the full 18N sequence as well are surrounding nucleotides were thoroughly analyzed and aligned for the best fit match of sequences. Adenine (A) and guanine (G), both purines were considered interchangeable nucleotides as were the pyrimadines cytosine (C) and thymine (T). The percentage of each nucleotide being selected at a specific position was calculated. A significant match was taken to be a base that occurred in nearly half of the sequences chosen. In all but one of the positions of the nucleotides in the consensus sequences obtained, a definite, common nucleotide was used which indicates there is a

great percentage that these sequences are specific because the same nucleotide occurs at the same position.

The analysis of sequences started with the SNAG-ZFP clones, which each protein varied in terms of clones obtained so the number of sequences varied for each set. There were a number of replicates within each SNAG-ZFP set, and even between long and short constructs like that of Snail, for example. These replicates, which came in pairs and sometimes triplets, indicated that these particular sequences are being positively selected by the ZFP more often than some of the other binding sites.

In addition to the multiple pairs or triplets of identical sequences, there were also sequences that matched those sets with minor nucleotide changes, with one to just a few substitutes. Of note is the fact that between the long and short constructs of Snail and of Scratch, matches of identical sequences were made. This is expected, of course, since the constructs are nearly the same, with the exception of an additional ten amino acids, but it is good evidence that the binding was selective since the provided results were reproduced between two sets of proteins.

Repeated sequences within the sets of sequences indicate that various clones of the same ZFP are identifying the same sequence and binding to it therefore, it may be important to take into consideration as a true binding site of the transcriptional repressor. Not only did the sequences show good match between the long and short constructs of Snail and long and short constructs of Scratch, as mentioned previously, but the consensus sequences were well conserved between all the proteins analyzed.

It would be expected that a common sequence occurs across the various SNAG-ZFPs, since they belong to the same family, however, there could be variations between

the proteins that are due to specificity of the zinc fingers. Using the known SNAG consensus sequence E-box (CANNTG) as a core to build upon with flanking nucleotides, these variations in the surrounding nucleotides were sought as well as a more concise representation of the N nucleotides within the E-box.

An initial identification of the E-box within the 49-N oligonucleotide sequences obtained provided the foundation to determining other common nucleotides. First, only exact E-box hits were used, and then the remaining sequences were analyzed to find variations of the CANNTG sequence. This includes extra or missing nucleotides, or a different nucleotide in one position of the sequence. These sequences were aligned based on the position of the E-box, and the flanking nucleotides were added around the core to determine if there was any common pattern. Tables containing these alignments are shown in **Tables 1-6.** The breakdown of nucleotides at each position is calculated below the tables to illustrate the percentages of each. Any nucleotide with a significant appearance was taken as part of the consensus which is further represented, just below their corresponding tables, as a sequence with nucleotides sized according to their likelihood of showing in a binding site. Nucleotides were added to either side of the E-box, which served as a base for the consensus, until no similarity continued.

Upon analysis of the various proteins, a conclusion was made as to the consensus of each one. The E-box is the core of each consensus, and all six show very similar patterns, however, there are sites of variation between them, mainly with the number of nucleotides involved in the consensus on either side of the E-box. The consensus for Slug was found to be TGCACCTGYCCGA (where Y is a pyrimidine), and the Smuc consensus sequence was TGCACCTGTCCGA. The long construct of

Snail was GCACCTGTCCGA, and the short construct of Snail was CACCTGYCCG. Finally the consensus for the long construct of Scratch was CATTGCACCTGTCCGA, and the short construct of Scratch was GCACCTGTCCGA. A table comparing and aligning all of the sequences between SNAG-ZFPs can be seen in **Table 7.**

Noteworthy are the consensus sequences between constructs of the same SNAG-ZFP. The difference between the long and short constructs are additional amino acids on either side of the protein so these extra amino acids may contribute to binding DNA differently than their shorter partners. Between the long and short constructs of Snail, the consensus for the long protein was found to bind an extra G prior to the E-box and an additional A at the end of the consensus. The remaining portion of the sequence is similar except for the position directly following the E-box. Both can be T, but Snail short can also be a C at the position, meaning that a pyrimidine holds the spot. Finally for Scratch, the consensus for the long construct is four nucleotides longer than that of the short construct. The remaining portions of the sequences are identical. In both cases, the short construct would also be bound by the long construct so that good evidence that the sequences are true binding sites. It also proves that the additional amino acids in the long constructs help the protein bind better to the same consensus as the rest of the SNAG-ZFPs. In the case of Scratch long, the consensus was extended even further prior to the general consensus.

The consensus sequences still showed that the E-box is an excellent binding site for the SNAG-ZFPs, but analysis of these results show that there may be additional nucleotides that contribute to the binding of the zinc fingers. Some may be more specific than others, and these may be crucial nucleotides necessary for binding.

A longer binding site consensus indicates more specific binding of DNA, which will be useful in determining what genes are regulated. The odds of a particular nucleotide occurring at a certain position in a DNA sequence is $^1/_4$. The longer the sequence becomes, and the longer the consensus becomes, the less likely it is to appear by random chance. For example, the chances that the sequence CACCTG appears in some DNA strand is $(^1/_4)^6$ which is equivalent to one in 4096. The more specific the consensus is, the more likely particular genes will be found more easily that contain the sequence. In this research, the consensus between all of the SNAG-ZFPs was 13 nucleotides long, meaning the chances of this exact sequence occurring is one in 67,108,864. This great specificity increases the odds of finding some genes that are specific for these transcriptional repressors.

Another important factor to consider in the binding of DNA by these SNAG-ZFPs is the number of zinc fingers within the protein. Since there is a highly conserved CACCTG sequence found across all SNAG-ZFP consensus sequences, the individual zinc fingers that bind were analyzed. Slug, Smuc, and Scratch all have five zinc fingers while Snail has four. The number of zinc fingers plays a role in the length of the consensus sequence bound. Notice that Snail, with its four zinc fingers, yielded the shortest consensus, while the other SNAG-ZFPs had longer consensus sequences, signifying that the more zinc fingers present, the more DNA bound.

Additionally, ClustalW analysis, which can align amino acid sequences based on their best fit matches, shows that two of the zinc fingers of each protein are highly conserved (**Figure 15)**. The conserved sequences are those of zinc finger three of Slug, Smuc, and Scratch with zinc finger two of Snail, and zinc finger four of Slug, Smuc,

and Scratch with zinc finger three of Snail. There is also some conservation between the rest of the zinc fingers, but in a more moderate manner and not such a strong conservation as the sets just mentioned. These other zinc fingers play a role in binding the extra nucleotides that extend on either side of the E-box.

The highly conserved sequences indicate that the binding capability of these proteins is of similar function and would have similar binding sites. Because zinc fingers bind three nucleotides each, this reveals that these two zinc fingers play a role in binding the CACCTG sequence. In order to confirm this, site directed mutagenesis of these zinc fingers will be performed. By mutating the highly conserved amino acids in these highly conserved zinc fingers, the binding to E-box is expected to be eliminated. This would prove that these two zinc fingers of each SNAG-ZFP plays the role in binding E-box.

Additional information derived from these sequences is that these consensus sequences are specific to the SNAG-domain family of transcriptional repressors. Although there are other E-box binding families, of those researched, none fit the general consensus from this research. For example, of the helix-loop-helix structured general muscle differentiation proteins discussed in the introduction, none of the known binding sites match the consensus found for the SNAG-ZFPs. Myf3-Myf6 are found to bind to GCAGGTG, which is characteristic for the binding site of that particular group of proteins. Myogenin is found to bind to the enhancer of muscle creatine kinase (MCK) which contains two separate E-boxes. The first is of a CATGTG nature, (also different from the CACCTG consensus found for SNAG-ZFPs), while the second is CACCTG. Although the second E-box within the MCK enhancer is the same of our

identified consensus, upon analysis of the extended regions bound on either side of the E-box, there is no indication that these binding sites are similar. This is an advantage of having the extended consensus for the SNAG-ZFPs and also a clue that the binding sites found for the SNAG-ZFPs is specific solely for this family of transcriptional repressors.

Knowing the DNA binding sites of these SNAG-ZFPs via their zinc fingers is important as these transcriptional repressors are important in a number of cancers and diseases. There may be various sequence-specific binding sites to which they bind, and by breaking these interactions, the repressor is itself repressed. Further studies using the data found here will be done to expand the information known about these transcriptional repressors and how they interact and function.

IDENTIFICATION AND CHARACTERIZATION OF NOVEL SLUG AND SCRATCH DOMAIN INTERACTING PROTEINS

The second aim of this thesis research is to determine interactions that occur between cellular proteins and the SNAG-ZFPs by extending this project into human cell lines. Specifically, efforts will be made to identify novel protein partners that interact with Slug and Scratch domains represented in the diagrammatic representation of the structure of SNAG ZFPs presented in **Figure 1.** These Slug and Scratch domains are highly conserved among the same proteins of different organisms indicating that there is some strong evidence that it plays a role in some protein interaction. ClustalW analysis was done for each of the domains that indicates this strong conservation. The results are found in **Figure 16.**

Experimental Design and Data

This study was done in parallel with the DNA binding site identification studies mentioned above. Induced samples were obtained as previously described for the GST-SNAG-ZFPs. Results are shown in **Figure 17 (Panel A)**. In order to begin the protein association studies, GST-Slug and GST-Scratch proteins were prepared from 50mL cell pellets of Slug (84-127 aa), Scratch short(107-126 aa), and Scratch long (102-131 aa). PBS (1x) and lysozyme were used to break the cell wall, followed by the addition of phenylmethylsulphonylfluoride (PMSF) and a proteinase inhibitor cocktail consisting of

aprotinin, leupeptin, and pepstatin. Dithiothreitol (DTT) and sodium lauryl sarcosine (SLS)/ Sarkosyl were added followed by sonication to release the desired proteins as shown in **Figure 17 panel B**. **Figure 17 (Panels C-E)** demonstrates the purification schemes of the GST-Slug and Scratch domains from the previously shown SLS protein preparation. This reduces the number of the proteins present in the sample and a good enrichment of purified domain protein was observed. The GST- Slug and Scratch domain proteins behaved well when compared to the zinc fingers using the same protocol, as previously presented.

A cell line was sought in which Slug and Scratch had good expression, indicating that there is some protein-protein interaction occurring in which the domain proteins participate. Online microarray databases are available in which a profile of the protein of interest is generated, specifying the expression of the protein in multiple cell lines. Using this approach, HEK293T was identified and used for the following experiment.

A nuclear extract of 293T cells was prepared to use in the bindings of the SLUG and SCRT domains. This cell line is a derivative of the cell line HEK (human embryonic kidney) 293. This preparation entailed a procedure involving NLB, NEB, and the proteinase inhibitor cocktail. The domain proteins and nuclear extract were bound on G75 fine Sepharose beads and washed, spun, removed of supernatant by vacuuming, and this cycle was repeated with binding buffers of varying salt concentrations (100, 250, and 500 mM). The resultant products were boiled and loaded on a polyacrylamide gel along with nuclear extract. These gels underwent silver staining to show the proteins. These gels are presented in **Figure 18**.

Discussion

Analysis of the gels was between the unbound protein domains and the bound domains to 293 T nuclear extract. Any bands visualized in the binding wells of varying salt concentration washes but not in the unbound nuclear extract would be taken as significant finding. For the identification of interactions between the domains of Slug and Scratch and the 293 T cell line, the results were difficult to determine. Thorough visual analysis was done to identify the presence and absence of bands.

For the Slug domain, a single band met the conditions needed and was taken as a putative protein-protein interaction band. This is marked by a box in **Figure 18.** Other similar bands were sought for the remainder of the gel and for the Scratch domain gel. Unfortunately, for the Scratch domain samples, looking between each sample showed no largely prominent or distinct bands in one set of bands versus another. Due to the large number of proteins found in the cell line, it was difficult to discern one independent band from another, and proved even more challenging to decide whether one appeared in one lane versus the others.

Additional studies using this binding assay method may need to be done with different conditions in order to determine where any protein-protein interactions are occurring. This will provide more information on to what the Slug and Scratch domains are binding. The next step to confirm the putative band found in the Slug binding is to perform metabolic labeling. In this procedure, a cell line will be starved in a methionine- and cysteine-deficient medium then fed with ^{35}S-methionine in order to label the entire proteome. By repeating the same binding assay, a gel will indicate bands where protein-protein interactions can be visualized. The expected result is that on these

gels, a band will appear in the same region as our putative band found in this binding assay. These bands can then be further analyzed using mass spectrometry.

FURTHER STUDIES

Additional studies to be taken related to those done for this thesis include finishing the research to complete the characterization for the remaining SNAG-ZFPs. This includes determining proper conditions for getting soluble protein fractions from those not easily solubilized including GFI-1 and its counterparts. Once accomplished, the entire procedure to identify DNA-binding site sequences will be repeated. The IA-1 ZFPs have been processed and are ready to be sequenced. The final EMSAs will be done in conjunction with the GFIs.

The consensus sequences determined from the data obtained through these experiments will be used to perform site-directed mutagenesis in order to confirm that these are true binding sites. By mutating nucleotides in the consensus, it is expected that binding will not occur, yet for the consensus, binding is expected if it is an actual site where the zinc fingers bind. Additionally, as discussed previously, the site-directed mutagenesis of the amino acids in the zinc fingers expected to play the biggest part in CACCTG binding will be performed.

In order to identify the novel function of the Slug and Scratch domains, the results from these binding assays will be used in conjunction with the proposed metabolic labeling experiment. Another completely different approach to distinguishing protein-protein interactions is to perform yeast two-hybrid screening, which will be

done to identify interactions of SNAG-ZFPs with other proteins in human cell lines.

By completing the entire characterization of these proteins, it will be quite useful and informative to knowing how these repressors interact, what they bind to, and how they function. This information will be quite useful to further the picture of how these proteins play a role in particular diseases and what methods can be employed to stop these activities from occurring. On a broader scope, knowing as much as possible about these transcriptional repressors will contribute to further understanding of similar transcription factors.

FIGURES

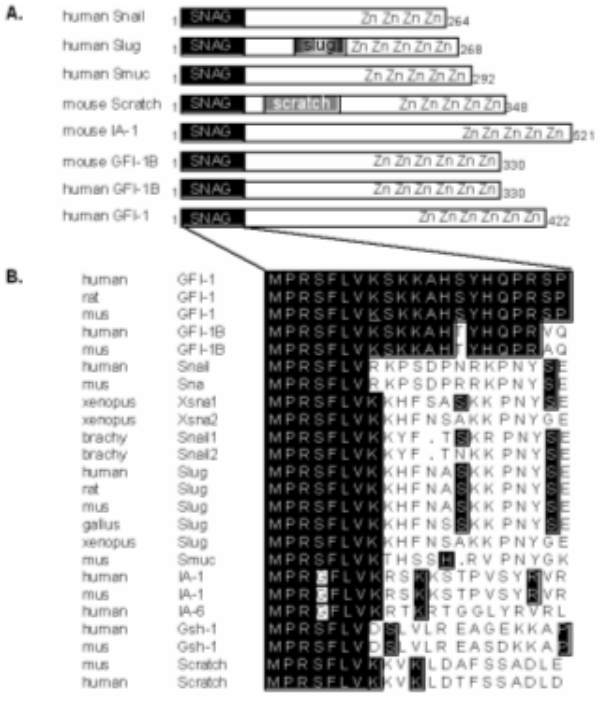

Figure 1. Alignment of SNAG Domain Transcription Factor Family members. SNAG domain transcription factors have a highly conserved 21-amino acid sequence at the N-terminal and a chain of C_2H_2 zinc finger repeats at the C-terminal, which are connected by a linker region. Slug and Scratch have domains in the linker region, but the function of these is unknown.

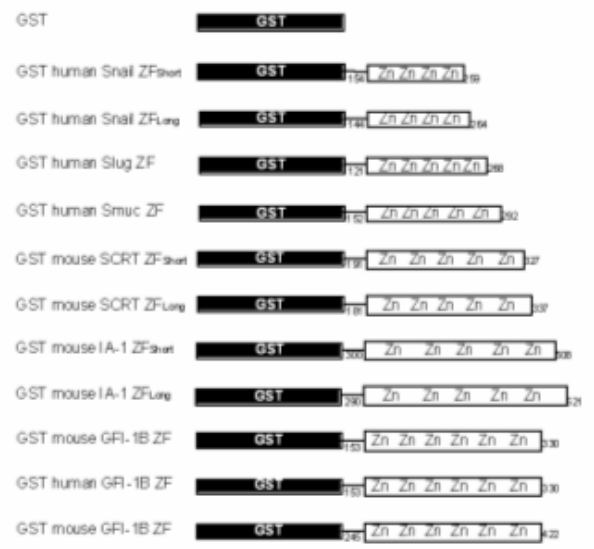

Figure 2. Diagrammatic representation of the GST-SNAG-ZF recombinant proteins. DNA segments corresponding to the indicated amino acids were PCR amplified by using the full-length genes of SNAG domain transcription factor family as templates. Zinc fingers were fused with Glutathione-S-transferase tag to generate recombinant fusion proteins.

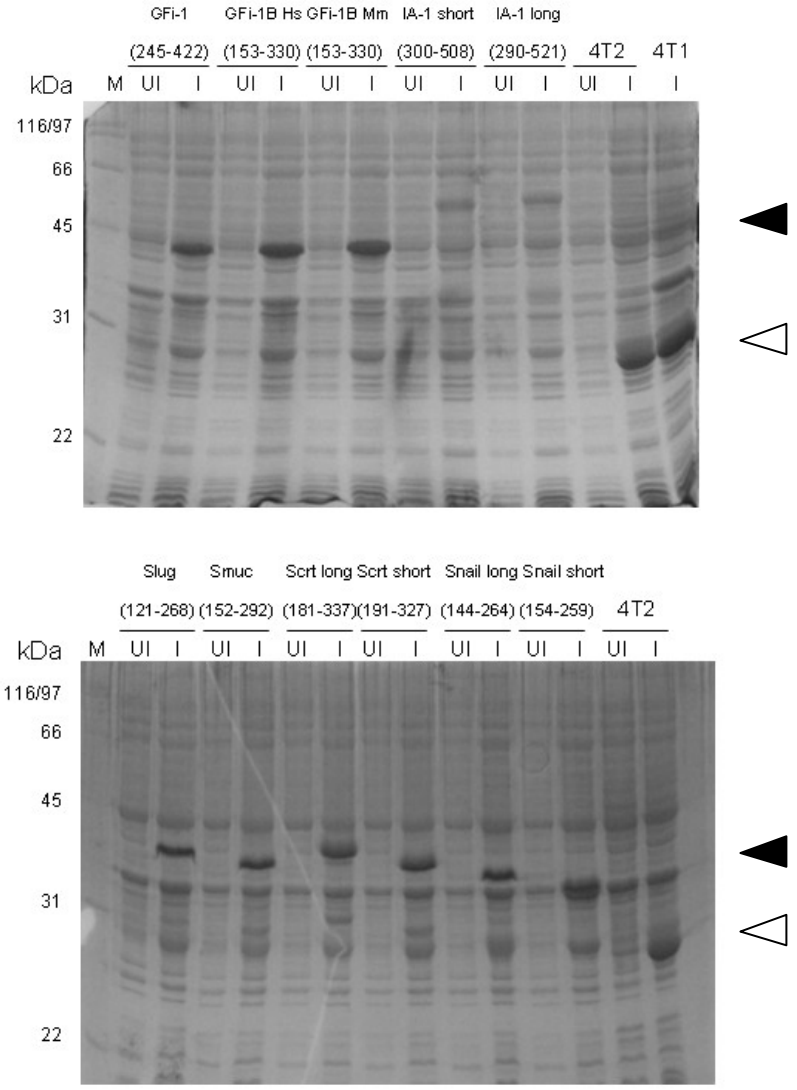

Figure 3. Induction demonstrating GST-SNAG ZF fusion proteins. Protein induction using positive GST-SNAG recombinant clones. GST-SNAG-ZF fusion proteins were induced (I) (filled arrowhead). The open arrowhead shows the GST protein. 4T2 and 4T1 are GST controls. Broad range marker (M) is also shown.

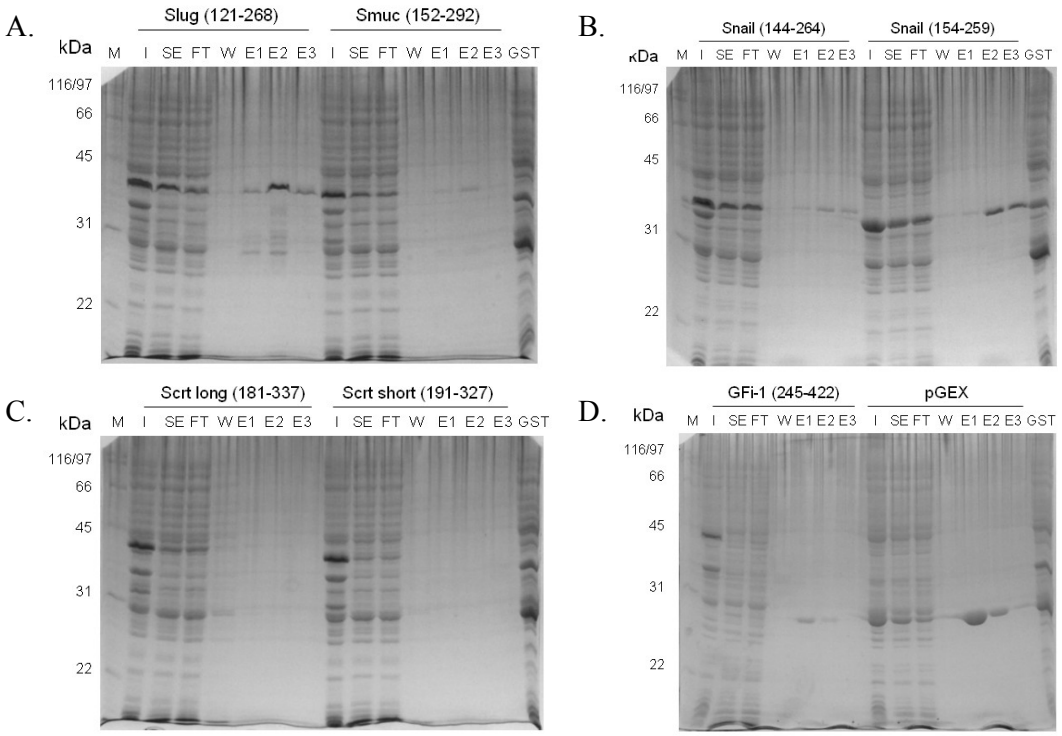

Figure 4. Solubilization and purification of GST-SNAG-ZFPs. GST-ZFPs expressed in *E. coli* BL21 cells were solubilized and purified via GSH-sepharose columns. Panels of induced cell culture (I), sonicated extract (SE), flowthrough of loaded material (FT), wash (W) and elutions (E1-E3) were run on SDS-PAGE gels and stained with Coomassie Blue. Broad range marker (M) and GST protein (GST) were also included as controls. (A) shows panels for Slug and Smuc, (B) shows the panels for long and short constructs of Snail, (C) shows the panels for long and short constructs of Scratch, and (D) shows the panels for GFi-1 and pGEX.

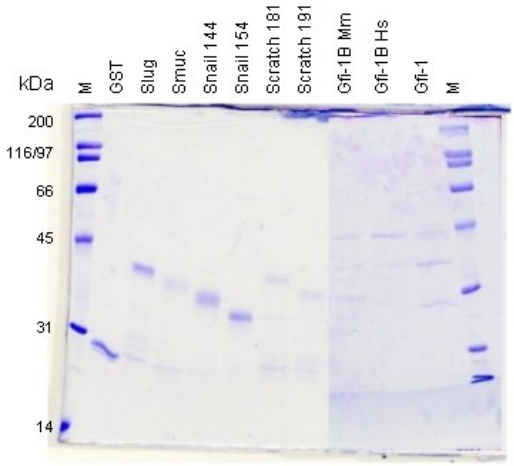

Figure 5. Purified GST-SNAG ZFPs used for binding site selection. Elution samples from column purification were further purified using electroelution. The purified proteins were used for bindings to E-cadherin and the randomized oligonucleotide library.

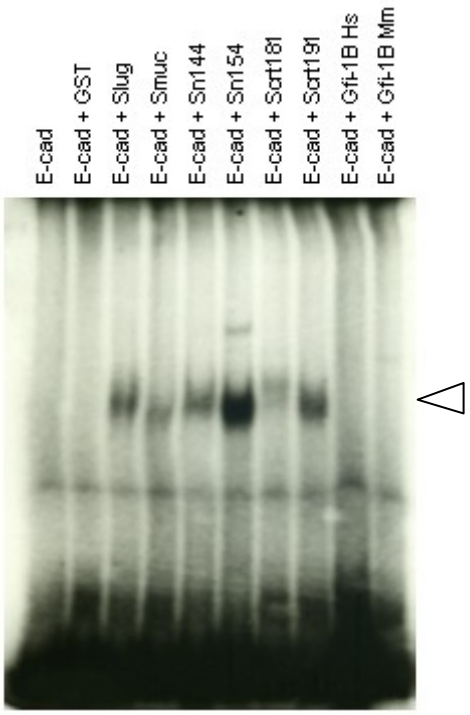

Figure 6 (left). Gel mobility shift assay of E-cadherin and GST-SNAG-ZFPs. The purified GST-SNAG-ZFPs were bound to E-cadherin promoter, a known binding site of Slug, Smuc, Snail, and Scratch. Bands of DNA-protein complexes were visualized for the proteins known to target E-cadherin, and showed no interaction with those that do not bind to the promoter. Bands are indicated by the open arrowhead.

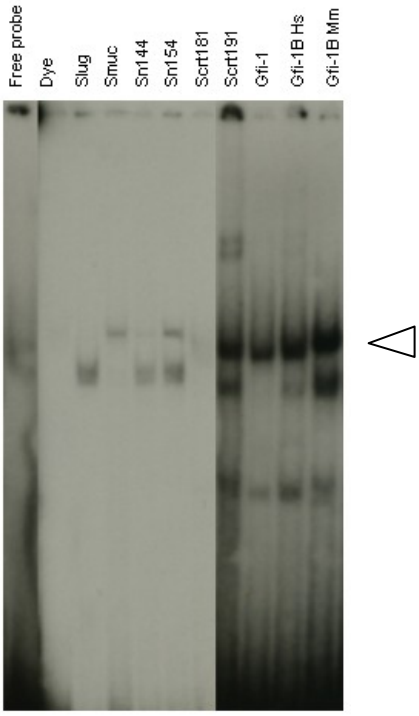

Figure 7 (left). EMSA of first binding site selection of SNAG-ZFPs with the randomized oligonucleotide library. GST-SNAG-ZFPs were bound to a randomized 49-N oligonucleotide library flanked with known PCR primers and restriction sites. Interactions between protein and DNA formed complexes that were retarded in the DNA-PAGE gel. The bands indicated by the open arrowhead were cut and eluted for subsequent PCR and gel shift assays.

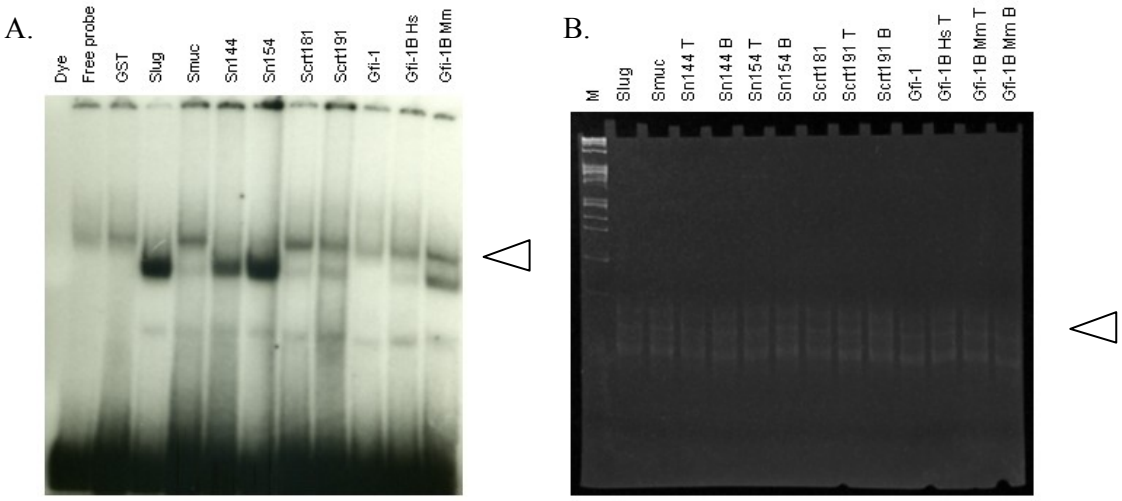

Figure 8. (A) Second binding site selection of SNAG-ZFPs. The enriched DNA binding site PCR products of the first binding site selection were cleaned and used for a second round of binding using the SNAG-ZFPs. The open arrowhead shows the area of bands used for subsequent PCR. For some products, top and bottom bands were used. **(B)** PCR products of previously bound and eluted enriched oligonucleotide library and SNAG-ZFPs run on a DNA-PAGE gel.

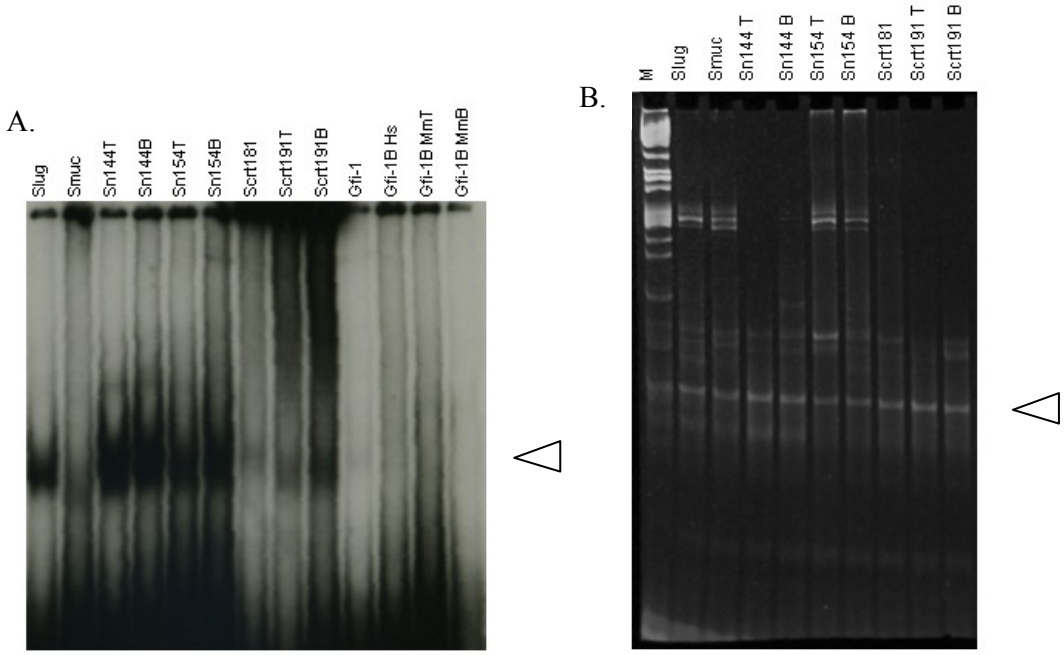

Figure 9. (A) Third binding site selection of SNAG-ZFPs with the enriched oligonucleotide library as shown by EMSA. Enriched PCR products from the second binding site selection procedure were used for binding SNAG-ZFPs. The open arrowhead shows the area of bands used for subsequent PCR. **(B)** PCR products of previously bound and eluted enriched oligonucleotide library and SNAG-ZFPs run on a DNA-PAGE gel and visualized with ethidium bromide staining and an ultraviolet illuminator.

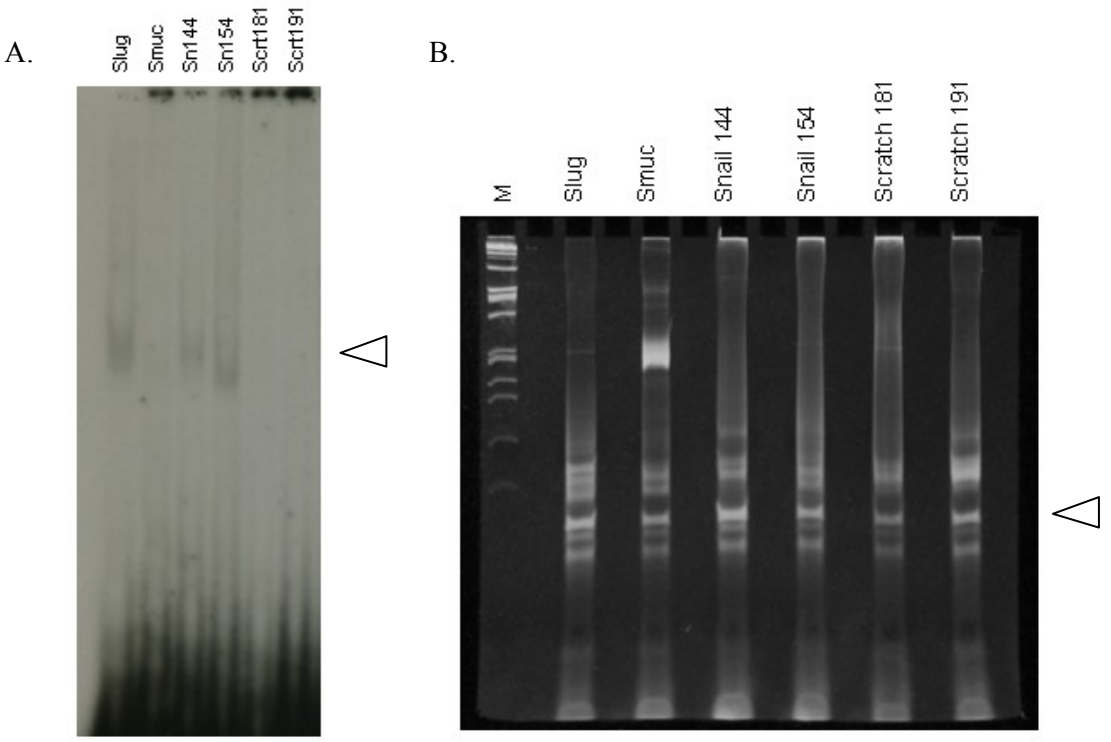

Figure 10. (A) Fourth and final binding site selection of SNAG-ZFPs with oligonucleotide library. EMSA of the binding between SNAG-ZFPs and enriched PCR product. **(B)** Enriched PCR products of previously bound SNAG-ZFPs and bound oligonucleotide library run on a DNA-PAGE gel and visualized with ethidium bromide staining and an ultraviolet illuminator.

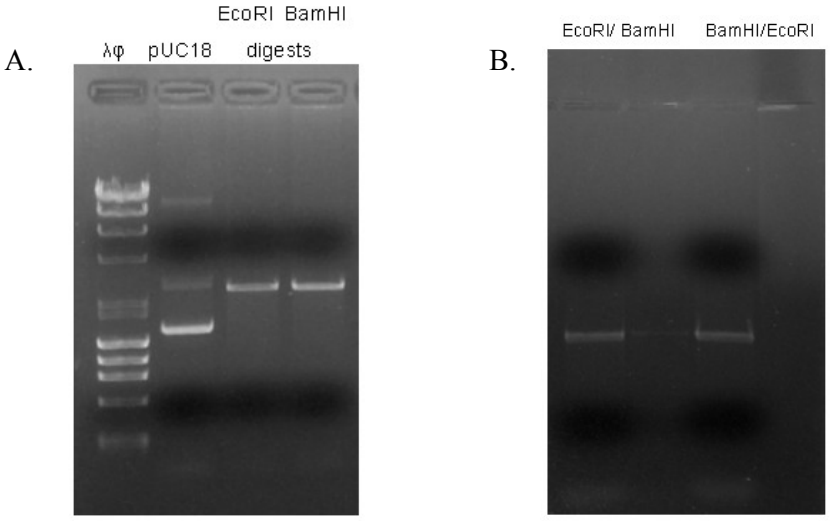

Figure 11. EcoRI/BamHI digestion of pUC18 vector. **(A)** Two pUC18 vector samples were first digested with restriction enzymes – one with EcoRI and the other with Bam HI. **(B)** The vectors were then digested with the restriction enzyme not used in the first digestion as part of the double digestion. This is done to ensure efficiency of the vector to be used for ligation.

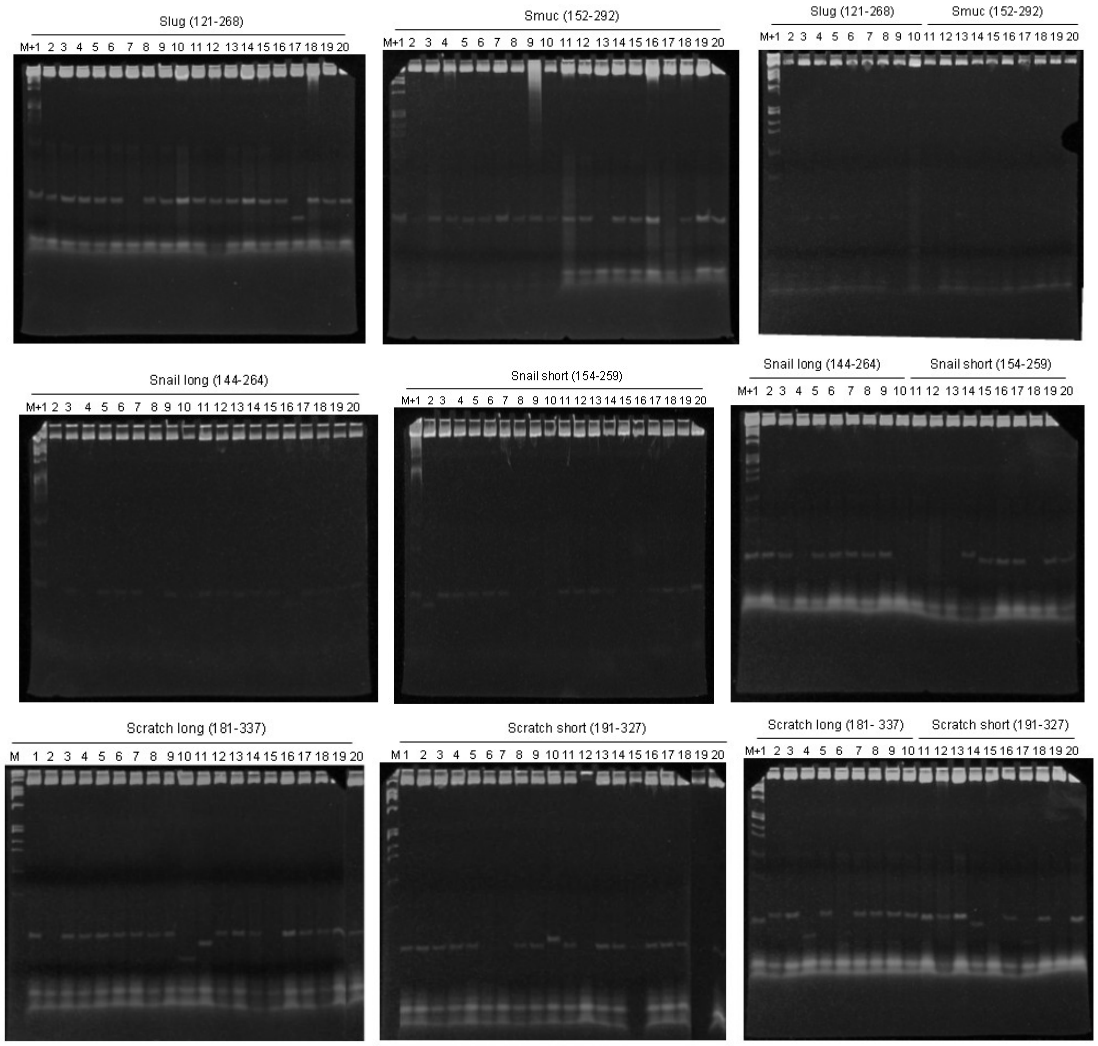

Figure 12. Positive recombinant clones. The electroeluted PCR products were pUC18-ligated and transformed into *E. coli* DH5α cells which were grown, mini-plasmid prepared, and run on a DNA-PAGE gel. Occurrence of a band indicates the presence of a positive clone.

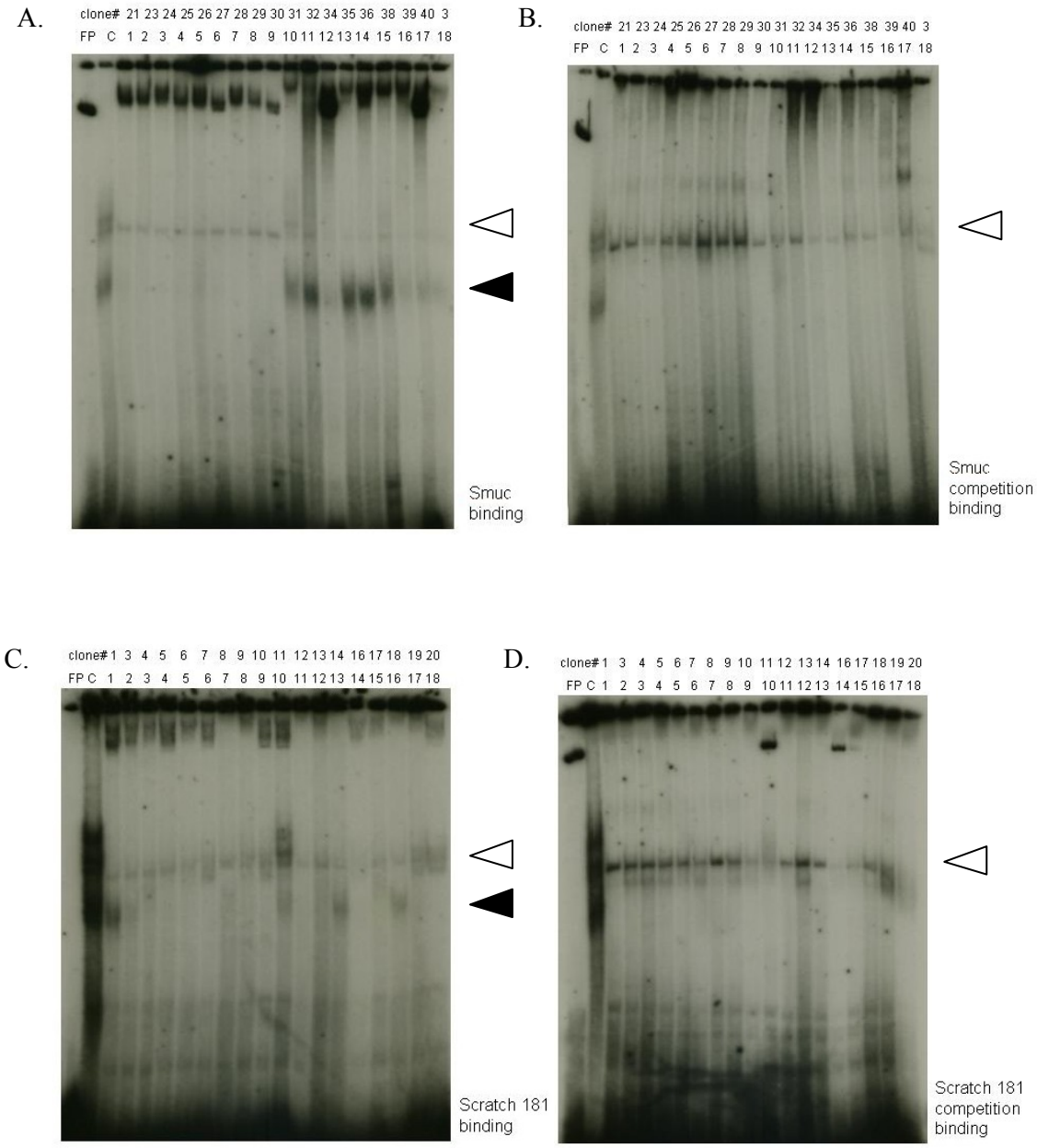

Figure 13. Binding of DNA to specific and nonspecific proteins. EMSAs of binding of the resultant positive clones with its cognate protein **(A)** and with a non-specific protein CG7938 **(B)**. Where a complex indicating binding between protein and DNA is seen in (A), the same is not seen for the competing protein autoradiograph (B). **(C)** Scratch long binding with its cognate protein shows similar pattern of bands that appear where an interaction is occurring. (D) The enriched Scratch long DNA obtained did not bind to the CG7938 protein, therefore no shifts corresponding to the control were seen. Filled arrowheads indicate non-specific binding and open arrowheads show the specific binding of DNA to protein.

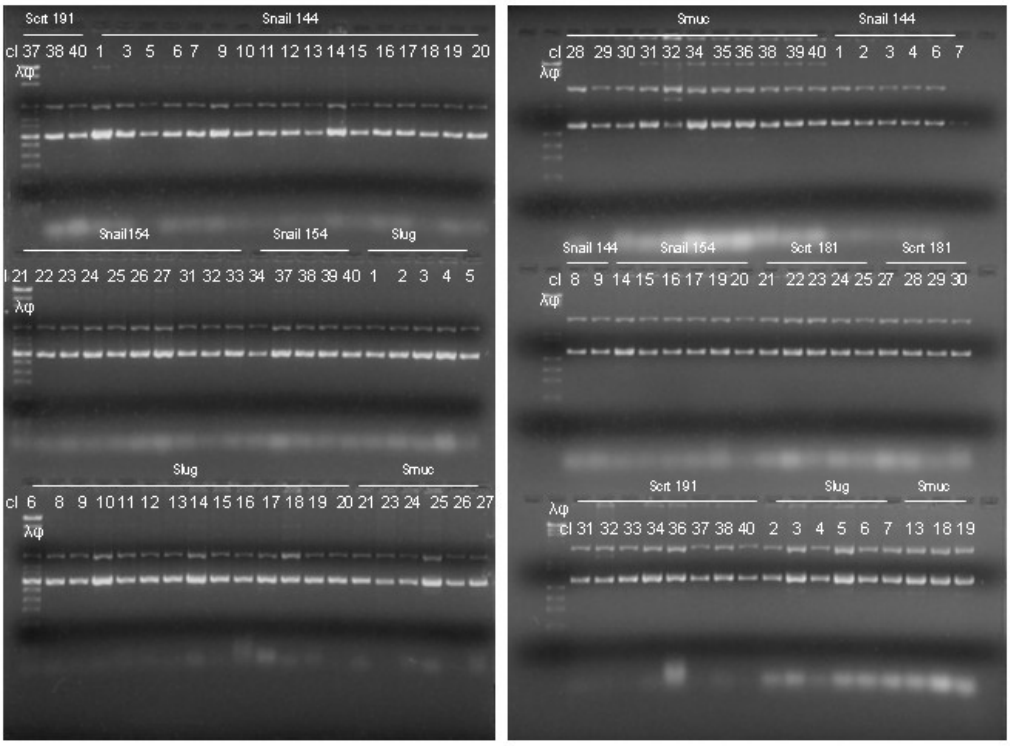

Figure 14. RNase and PCI cleaned recombinant clones. The mini-plasmid preps were RNase digested and purified in order to check the integrity and quantity of the DNA. Presented are the purified plasmids seen on agarose gels.

Slug	T	G	C	A	C	C	T	G	T/C	C	C	G	A
	T	A	C	A	C	C	T	G	T	C	C	G	A
	C	C	C	A	C	C	T	G	T	C	C	G	A
	T	G	C	A	C	C	T	G	T	C	C	G	A
	T	G	C	A	C	C	T	G	T	C	C	G	A
	T	G	C	A	G	G	T	G	G	G	G	G	G
	T	G	C	A	C	A	T	G	C	G	G	T	C
	T	G	C	A	G	G	T	G	G	A	C	G	G
	T	G	C	A	C	A	T	G	A	C	A	G	A
	T	G	C	A	C	A	T	G	C	A	C	A	C
	A	G	C	A	G	A	T	G	C	G	C	C	A
	T	G	C	A [A]	C	C	T	G	A	C	A	A	T
	T	G	C	A [A]	C	C	T	G	A	C	A	A	T
	T	G	C	A	C	C	C	G	A	G	G	A	G
	T	G	C	A	C	G	C	G	C	A	A	A	A
	T	G	C	A	C	C	A	G	G	C	G	G	C
	T	G	C	A	C	G	T [A]	G [G]	T	C	C	G	T
	T	G	C	A	C	C	G	G	G	G	G	G	T
	T	G	C	T	G	C [C]	T	G	T	C	C	G	A
	T	G	C	T	C	G [C]	T	G	T	C	C	G	A
	T	G	C	T	C	G [C]	T	G	T	C	C	G	A
	T	G	C	A	C	C	T	G	T/C	C	C	G	A
A	1	1	0	17	0	4	1	0	4	3	4	5	10
G	0	18	0	0	4	6	1	20	5	5	5	13	3
T	18	0	0	0	0	0	16	0	8	0	0	1	4
C	1	1	20	3	16	10	2	0	4	12	11	1	3
%	90	90	100	85	80	50	80	100	60	60	55	65	50

Table 1. Slug Consensus. The obtained consensus for Slug is TGCACCTGYCCGA as shown by the table of sequences from different clones and the percentage breakdown for each. The diagram below the table indicates the percentages as to the occurrence of a particular nucleotide at a certain position.

Smuc													
	T	G	C	A	C	C	T	G	T	C	C	G	A
	C	G	C	A	C	C	T	G	C	C	T	C	C
	T	G	C	A	G	G	T	G	C	C	C	G	A
	T	A	C	A	C	C	T	G	T	C	C	G	A
	A	C	C	A	C	C	T	G	T	C	C	G	A
	A	T	C	A	T	T	T	G	C	T	G	G	A
	T	A	C	A	C	C	T	G	T	C	C	G	A
	T	G	C	A	G	G	T	G	G	A	G	G	C
	T	G	C	A	G	G	T	G	G	A	C	G	T
	T	G	C	A	G	G	T	G	A	A	T	G	T
	T	G	C	A	C	A [C]	T	G	T	C	C	G	A
	T	G	C	A	C [C]	T	T	G	T	A	C	G	C
	T	G	C	A	A	T	A	G	T	T	A	A	C
	C	G	C	A	C	C	A	G	G	C	G	G	G
	T	**G**	**C**	**A**	**C**	**C**	**T**	**G**	**T**	**C**	**C**	**G**	**A**
A	2	2	0	**13**	1	1	2	0	1	4	1	1	**6**
G	0	**9**	0	0	4	4	0	**13**	3	0	3	**11**	1
T	**9**	1	0	0	1	2	**11**	0	**6**	2	2	0	2
C	2	1	**13**	0	**7**	**6**	0	0	3	**7**	**7**	1	4
%	**69**	**69**	**100**	**100**	**54**	**46**	**85**	**100**	**46**	**54**	**54**	**79**	**46**

Table 2. Smuc Consensus. The obtained consensus for Smuc is TGCACCTGTCCGA as shown by the table of sequences from different clones and the percentage breakdown for each. The diagram below the table indicates the percentages as to the occurrence of a particular nucleotide at a certain position.

Snail Long												
	G	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	G	G	T	G	T	C	T	G	T
	G	C	A	G	A	T	G	G	G	C	C	C
	C	C	A	C	C	T	G	T	C	C	G	A
	G	C	A	G	A	T	G	G	G	C	G	C
	G	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	C	C	T	G	T	C	C	G	A
	T	C	A	C	C	T	G	T	C	C	G	A
	G	C	A	C	G	T	G	A	T	G	A	A
	G	C	A	C	A	T	G	G	C	C	G	G
	G	C	A	C	A	T [T]	G	T	C	G	G	C
	G	C	A	T	G	G	G	A	C	G	G	T
	A	C	A	C	G	T	C	T	G	T	C	C
	G	C	A	A	G	T [C]	G	T	A	G	G	A
	G	C	A	A	C	G	G	T	A	C	G	C
	G	C	A	A	C	G	G	T	A	C	G	C
	G	C	A	C	C	T	G	T	C	C	G	A
A	1	0	**18**	3	4	0	0	2	3	0	1	**9**
G	**10**	0	0	3	5	3	**17**	3	3	4	**15**	1
T	1	0	0	1	0	**15**	0	**13**	1	2	0	2
C	6	**18**	0	**11**	**9**	0	1	0	**12**	**12**	2	6
%	**56**	**100**	**100**	**61**	**50**	**83**	**94**	**72**	**67**	**67**	**83**	**50**

Table 3. Snail Long Consensus. The obtained consensus for the long construct of Snail is GCACCTCCGA as shown by the table of sequences from different clones and the percentage breakdown for each. The diagram below the table indicates the percentages as to the occurrence of a particular nucleotide at a certain position.

Snail Short										
	C	A	C	C	T	G	T/C	C	C	G
	C	A	C	C	T	G	A	G	C	A
	C	A	C	C	T	G	T	C	C	G
	C	A	C	C	T	G	T	C	C	G
	C	A	C	C	T	G	T	C	C	G
	C	A	C	C	T	G	T	C	C	G
	C	A	A	G	T	G	C	C	C	G
	C	A	C	T	T	G	C	T	G	T
	C	A	G	G	T	G	C	T	G	T
	C	A	G	G	T	G	G	C	G	G
	C	C	G [G]	C	T	G	T	C	C	G
	C	G	G	G	T	G	T	G	G	G
	C	G	C	C	T	G	C	A	C	T
	C	A	G	A [G]	T	G	G	C	C	G
	C	A	A	G	T [C]	G	T	A	G	G
	C	A	-	C	T	G	C	A	C	C
	C	A	-	C	T	G	C	A	C	C
	C	A	C	C	T	G	T/C	C	C	G
A	0	**13**	2	1	0	0	1	4	0	1
G	0	2	5	5	0	**16**	2	2	5	**10**
T	0	0	0	1	**16**	0	**7**	2	0	3
C	**16**	1	**7**	**9**	0	0	**6**	**8**	**11**	2
%	**100**	**81**	**44**	**56**	**100**	**100**	**81**	**50**	**69**	**63**

Table 4. Snail Short Consensus. The obtained consensus for the short construct of Snail is CACCTGYCCG as shown by the table of sequences from different clones and the percentage breakdown for each. The diagram below the table indicates the percentages as to the occurrence of a particular nucleotide at a certain position.

Scratch Long																
	C	A	T	T	G	C	A	C	C	T	G	T	C	C	G	A
	T	T	G	C	A	C	A	C	A	T	G	A	C	G	G	A
	A	T	G	A	G	C	A	C	C	T	G	T	C	C	G	A
	G	C	A	C	C	C	A	C	C	T	G	A	G	C	A	G
	C	A	G	C	C	C	A	C	C	T	G	T	C	C	G	A
	C	A	T	T	G	C	A	C	A	T	G	T	T	G	A	T
	C	A	T	T	G	C	A	C	A	T	G	G	G	A	C	G
	C	A	T	T	G	C	A	C	C	T	G	T	C	C	C	G
	C	A	T	T	G	C	A	C	C	T	G	C	C	G	A	G
	A	G	C	C	A	C	A	C	C	T	G	T	C	C	G	A
	G	G	C	C	C	C	A	C	C	T	G	T	C	C	G	A
	C	G	G	G	G	C	A	A	C	T	G	T	C	C	G	A
	A	G	A	T	C	C	A	C	C	T	G	T	C	C	G	A
	G	G	T	G	C	C	A	C	C	T	G	T	C	C	G	A
	C	A	T	T	G	C	A	T	G	T	G	T	G	T	A	C
	C	A	T	T	G	C	A	T	A [G]	T	G	T	C	G	G	C
	C	C	G	C	C	C	A	T	G [C]	T	G	T	C	C	G	A
	C	A	G	A	C	C	A	C	A [C]	T	G	T	C	C	G	A
	C	A	T	T	G	C	A	C	T [G]	T	G	C	C	G	G	G
	C	A	T	T	G	C	A	C	A [G]	T	G [G]	T	G	G	G	C
	T	G	T	T	G	C	T	G	C	T	G	T	C	C	G	A
	C	A	T	T	G	C	A	C	C	T	G	T	C	C	G	A
A	3	**9**	2	2	2	0	**19**	1	6	0	0	2	0	1	4	**11**
G	3	6	6	2	**11**	0	0	1	2	0	**20**	1	4	6	**14**	5
T	2	2	**10**	**10**	0	0	1	3	1	**20**	0	**15**	1	1	0	1
C	**12**	2	2	6	7	**20**	0	**15**	**11**	0	0	2	**15**	**12**	2	3
%	**60**	**45**	**50**	**50**	55	**100**	**95**	**68**	55	**100**	**100**	75	75	**60**	**70**	55

Table 5. Scratch Long Consensus. The obtained consensus for the long construct of Scratch is CATTGCACCTGTCCGA as shown by the table of sequences from different clones and the percentage breakdown for each. The diagram below the table indicates the percentages as to the occurrence of a particular nucleotide at a certain position.

Scratch Short												
	G	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	C	C	T	G	T	C	C	G	A
	G	C	A	C	C	T	G	T	C	C	G	A
	G	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	C	C	T	G	T	C	C	G	A
	C	C	A	C	C	T	G	T	C	C	G	A
	G	C	A	C	G	T	G	G	T	C	C	C
	G	C	A	C	C	T	G	T	C	C	G	A
	G	C	A	C	G	T	G	T	T	G	C	G
	G	C	A	C	C	T	G	A	C	G	A	T
	G	C	A	G	G	T	G	G	G	T	A	C
	A	C	A	G	C	T	G	T	A	C	G	T
	T	C	A	C	G	T	G	T	T	G	C	T
	G	C	A	C	A	T [T]	G	C	A	A	G	T
	G	C	A	A [A]	C	T	G	T	A	C	G	C
	T	C	A [A]	C	C	T	G	T	C	C	G	A
	G	C	A	C [C]	A	T	G	C	A	A	C	A
	A	C	A	G [G]	C	T	G	T	C	G	C	T
	G	C	A	C	C	T	G	T	C	C	G	A
A	2	0	**19**	1	2	0	0	1	4	2	2	**10**
G	**10**	0	0	3	4	0	**19**	2	1	4	**12**	1
T	2	0	0	0	0	**19**	0	**14**	3	1	0	5
C	5	**19**	0	**15**	**13**	0	0	2	**11**	**12**	5	3
%	**53**	**100**	**100**	**79**	**68**	**100**	**100**	**74**	**58**	**63**	**63**	**53**

Table 6. Scratch Short Consensus. The obtained consensus for the long construct of Scratch is GCACCTGTCCGA as shown by the table of sequences from different clones and the percentage breakdown for each. The diagram below the table indicates the percentages as to the occurrence of a particular nucleotide at a certain position.

Slug					T	G	C	A	C	C	T	G	T/C	C	C	G	A
Smuc					T	G	C	A	C	C	T	G	T	C	C	G	A
Snail Long						G	C	A	C	C	T	G	T	C	C	G	A
Snail Short							C	A	C	C	T	G	T/C	C	C	G	
Scratch Long		C	A	T	T	G	C	A	C	C	T	G	T	C	C	G	A
Scratch Short						G	C	A	C	C	T	G	T	C	C	G	A

Table 7. SNAG-ZFP Consensus Sequences. The consensus sequences of the SNAG-ZFPs are aligned to reveal a common pattern of nucleotides that extends from the core E-box binding site and that this sequence of TGCACCTGTCCGA is conserved among the proteins.

Figure 15. ClustalW alignment of zinc fingers. Although conserved in most regions between zinc fingers of various SNAG-ZFPs, the most highly conserved are those seen in the third and fourth groupings depicted. Snail has four zinc fingers whereas Slug, Smuc, and Scratch have five. The second finger of Snail and the third fingers of Slug, Smuc, and Scratch show good conservation. Also highly conserved are the third zinc finger of Snail and the fourth fingers of Slug, Smuc, and Scratch. Asterisks denote perfect matches between those aligned, colons denote good matches, and periods denote some similarity.

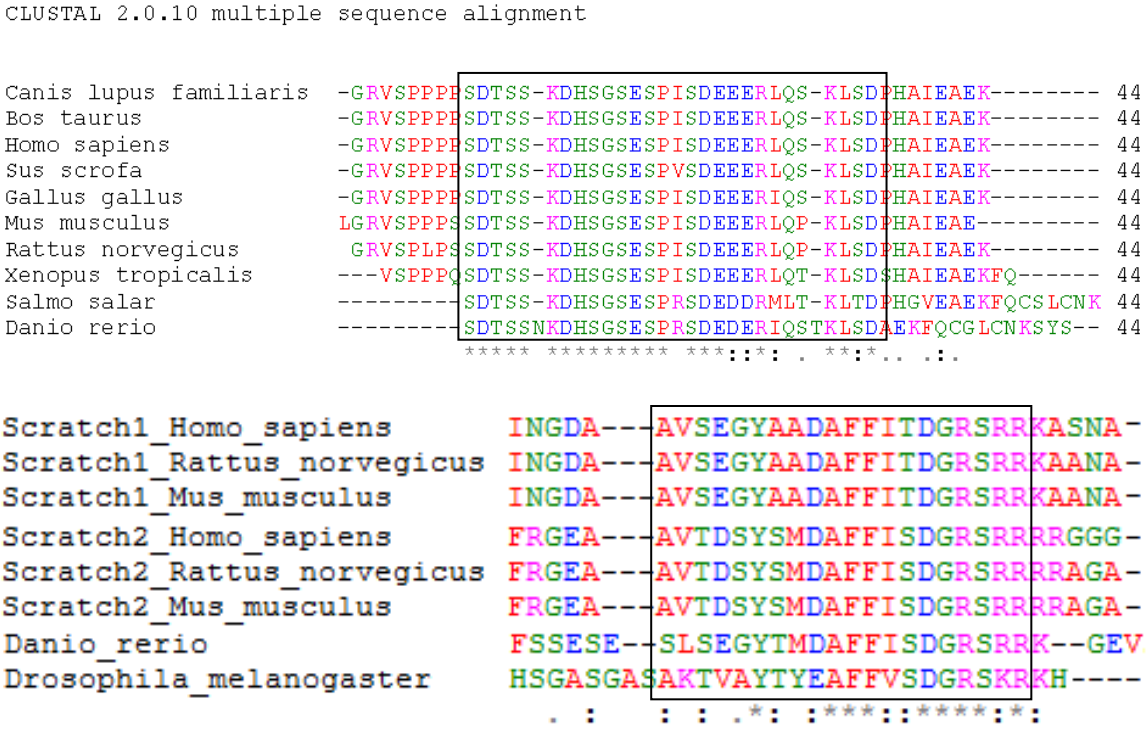

Figure 16. Alignment of Slug domains and Scratch domains. The Slug and Scratch domains are highly conserved among a host of organisms indicating some importance. They are found exclusively in the same proteins of the different organisms. The boxes around the amino acid sequences show the actual domain sequence, while the additional amino acids on either side are to show the flanking regions added in the constructs made for each domain.

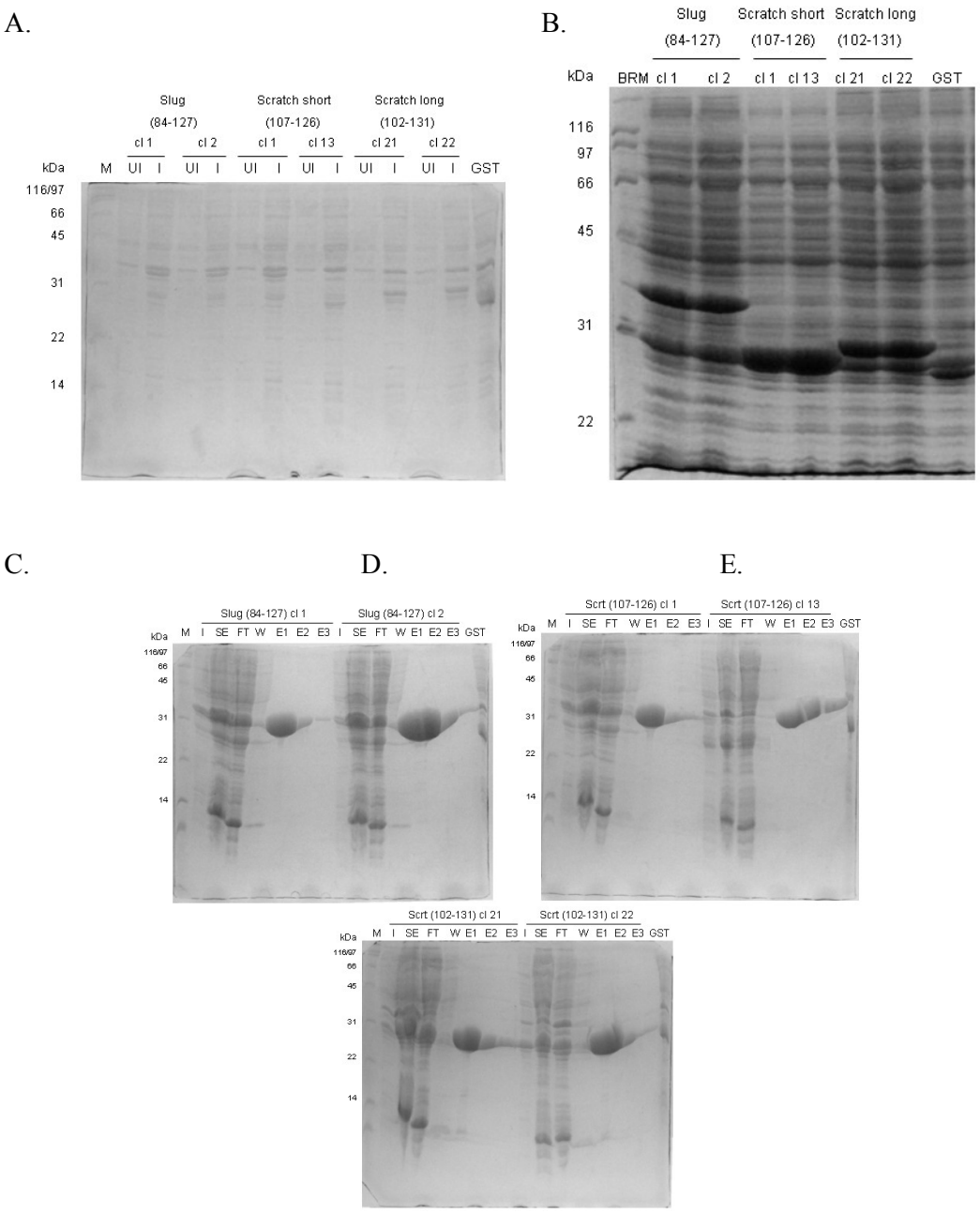

Figure 17. Slug and Scratch domains. **(A)** Protein induction demonstrating Slug and Scratch domain fusion proteins. GST-SNAG-ZF fusion proteins were induced (I) and run along their uninduced (UI) counterparts. Broad range marker (M) and GST control are also shown. **(B)** Purification of domain proteins. Sodium laurel sulfate was used to purify the domains of Slug and both constructs of Scratch. **(C-E)** Solubilization and purification of GST-SNAG-ZFPs. Slug and Scratch domain proteins expressed in *E. coli* BL21 cells were solubilized and purified via GSH-sepharose columns. Panels of induced cell culture (I), sonicated extract (SE), flowthrough of loaded material (FT), wash (W) and elutions (E1-E3) are shown Broad range marker (M) and GST protein (GST) were also included as controls. All samples were run on SDS-PAGE gels and stained with Coomassie Blue.

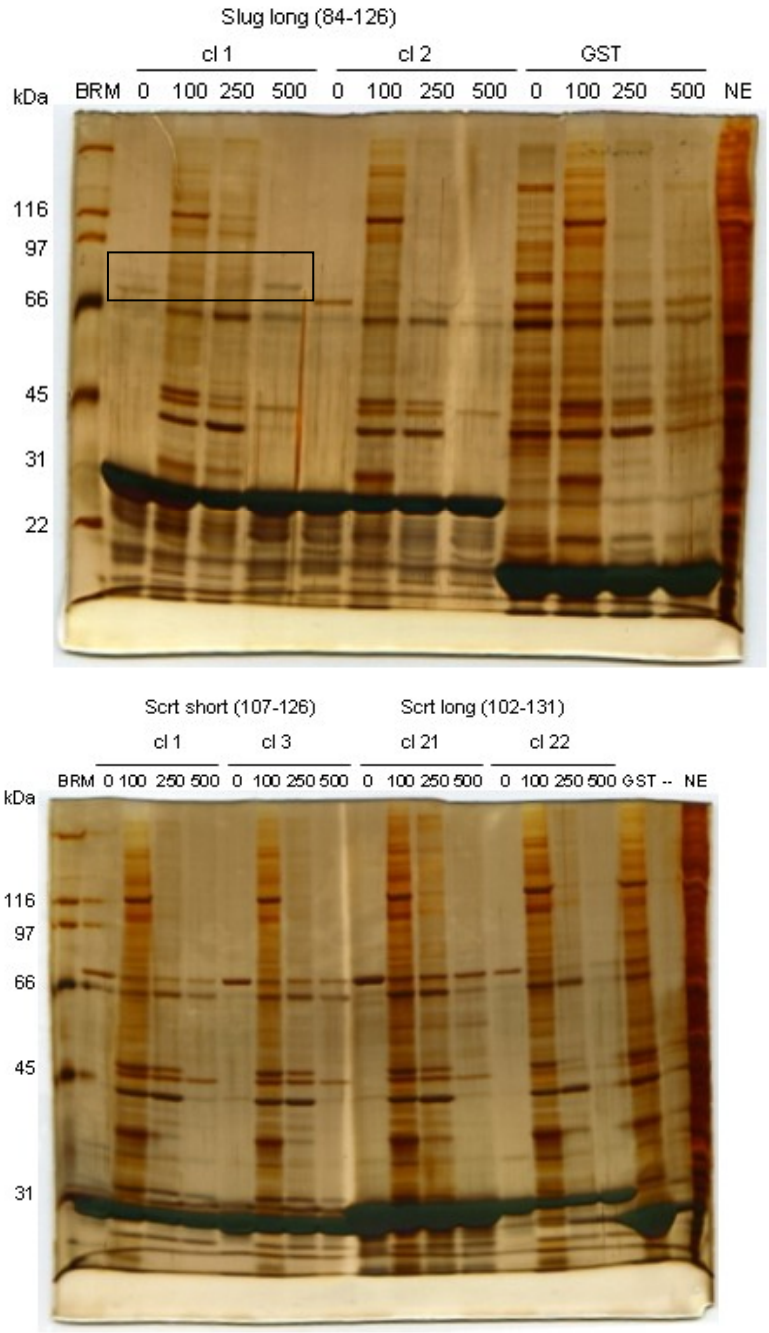

Figure 18. Binding assays of the Slug and Scratch domains. The Slug and Scratch domains were bound with the human embryonic kidney cell line 293T to identify protein-protein interactions. These binding assays were performed at various binding buffer concentrations and run on SDS-PAGE gels and silver stained. Any differentiating bands between bound nuclear extract and unbound nuclear extract would be significant. Further testing will be done to identify putative protein targets.

WORKS CITED

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.

Ayyanathan, K., Peng, H., Hou, Z., Fredericks, W. J., Goyal, R. K., Langer, E. M., Longmore, G. D., and Rauscher, F. J., 3rd. (2007). The Ajuba LIM domain protein is a corepressor for SNAG domain mediated repression and participates in nucleocytoplasmic Shuttling. *Cancer Res* 67, 9097-106.

Batlle, E., Sancho, E., Franci, C., Dominguez, D., Monfar, M., Baulida, J., and Garcia De Herreros, A. (2000). The transcription factor snail is a repressor of E-cadherin gene expression in epithelial tumour cells. *Nat Cell Biol* 2, 84-9.

Berger, S. L. (2007). The complex language of chromatin regulation during transcription. *Nature* 447, 407-12.

Blanco, M. J., Moreno-Bueno, G., Sarrio, D., Locascio, A., Cano, A., Palacios, J., and Nieto, M. A. (2002). Correlation of Snail expression with histological grade and lymph node status in breast carcinomas. *Oncogene* 21, 3241-6.

Braun, T., and Arnold, H. H. (1991). The four human muscle regulatory helix-loop-helix proteins Myf3-Myf6 exhibit similar hetero-dimerization and DNA binding properties. *Nucleic Acids Res* 19, 5645-51.

Brennan, T. J., and Olson, E. N. (1990). Myogenin resides in the nucleus and acquires high affinity for a conserved enhancer element on heterodimerization. *Genes Dev* 4, 582-95.

Breslin, M. B., Zhu, M., and Lan, M. S. (2003). NeuroD1/E47 regulates the E-box element of a novel zinc finger transcription factor, IA-1, in developing nervous system. *J Biol Chem* 278, 38991-7.

Breslin, M. B., Zhu, M., Notkins, A. L., and Lan, M. S. (2002). Neuroendocrine differentiation factor, IA-1, is a transcriptional repressor and contains a specific DNA-binding domain: identification of consensus IA-1 binding sequence. *Nucleic Acids Res* 30, 1038-45.

Cano, A., Perez-Moreno, M. A., Rodrigo, I., Locascio, A., Blanco, M. J., del Barrio, M. G., Portillo, F., and Nieto, M. A. (2000). The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. *Nat Cell Biol* 2, 76-83.

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.

Davis, R. L., Cheng, P. F., Lassar, A. B., and Weintraub, H. (1990). The MyoD DNA binding domain contains a recognition code for muscle-specific gene activation. *Cell* 60, 733-46.

Fondell, J. D., Brunel, F., Hisatake, K., and Roeder, R. G. (1996). Unliganded thyroid hormone receptor alpha can target TATA-binding protein for transcriptional repression. *Mol Cell Biol* 16, 281-7.

Giroldi, L. A., Bringuier, P. P., de Weijert, M., Jansen, C., van Bokhoven, A., and Schalken, J. A. (1997). Role of E boxes in the repression of E-cadherin expression. *Biochem Biophys Res Commun* 241, 453-8.

Goldmark, J. P., Fazzio, T. G., Estep, P. W., Church, G. M., and Tsukiyama, T. (2000). The Isw2 chromatin remodeling complex represses early meiotic genes upon recruitment by Ume6p. *Cell* 103, 423-33.

Grimes, H. L., Chan, T. O., Zweidler-McKay, P. A., Tong, B., and Tsichlis, P. N. (1996). The Gfi-1 proto-oncoprotein contains a novel transcriptional repressor domain, SNAG, and inhibits G1 arrest induced by interleukin-2 withdrawal. *Mol Cell Biol* 16, 6263-72.

Hanna-Rose, W., and Hansen, U. (1996). Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet* 12, 229-34.

Hemavathy, K., Hu, X., Ashraf, S. I., Small, S. J., and Ip, Y. T. (2004). The repressor function of snail is required for Drosophila gastrulation and is not replaceable by Escargot or Worniu. *Dev Biol* 269, 411-20.

Hou, Z., Peng, H., Ayyanathan, K., Yan, K. P., Langer, E. M., Longmore, G. D., and Rauscher, F. J., 3rd. (2008). The LIM protein AJUBA recruits protein arginine methyltransferase 5 to mediate SNAIL-dependent transcriptional repression. *Mol Cell Biol* 28, 3198-207.

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.

Jackson, K. A., Valentine, R. A., Coneyworth, L. J., Mathers, J. C., and Ford, D. (2008). Mechanisms of mammalian zinc-regulated gene expression. *Biochem Soc Trans* 36, 1262-6.

Kataoka, H., Murayama, T., Yokode, M., Mori, S., Sano, H., Ozaki, H., Yokota, Y., Nishikawa, S., and Kita, T. (2000). A novel snail-related transcription factor Smuc regulates basic helix-loop-helix transcription factor activities via specific E-box motifs. *Nucleic Acids Res* 28, 626-33.

Laux, H., Tomer, R., Mader, M. T., Smida, J., Budczies, J., Kappler, R., Hahn, H., Blochinger, M., Schnitzbauer, U., Eckardt-Schupp, F., Hofler, H., and Becker, K. F. (2004). Tumor-associated E-cadherin mutations do not induce Wnt target gene expression, but affect E-cadherin repressors. *Lab Invest* 84, 1372-86.

Lodish, H., Berk, A., Matsudaira, P., Kaiser, C., Krieger, M., Scott, M., Zipursky, A., Darnell, J. (2004). Molecular Cell Biology.

Manzanares, M., Locascio, A., and Nieto, M. A. (2001). The increasing complexity of the Snail gene superfamily in metazoan evolution. *Trends Genet* 17, 178-81.

Martinez-Estrada, O. M., Culleres, A., Soriano, F. X., Peinado, H., Bolos, V., Martinez,

F. O., Reina, M., Cano, A., Fabre, M., and Vilaro, S. (2006). The transcription factors Slug and Snail act as repressors of Claudin-1 expression in epithelial cells. *Biochem J* 394, 449-57.

Mauhin, V., Lutz, Y., Dennefeld, C., and Alberga, A. (1993). Definition of the DNA-binding site repertoire for the Drosophila transcription factor SNAIL. *Nucleic Acids Res* 21, 3951-7.

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.

Nagai, K., Nakaseko, Y., Nasmyth, K., and Rhodes, D. (1988). Zinc-finger motifs expressed in E. coli and folded in vitro direct specific binding to DNA. *Nature* 332, 284-6.

Nieto, M. A. (2002). The snail superfamily of zinc-finger transcription factors. *Nat Rev Mol Cell Biol* 3, 155-66.

Okkema, P. G., and Krause, M. (2005). Transcriptional regulation. *WormBook*, 1-40.

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.

Tsukiyama, T., Daniel, C., Tamkun, J., and Wu, C. (1995). ISWI, a member of the SWI2/SNF2 ATPase family, encodes the 140 kDa subunit of the nucleosome remodeling factor. *Cell* 83, 1021-6.

Tsukiyama, T., and Wu, C. (1995). Purification and properties of an ATP-dependent nucleosome remodeling factor. *Cell* 83, 1011-20.

Zweidler-Mckay, P. A., Grimes, H. L., Flubacher, M. M., and Tsichlis, P. N. (1996). Gfi-1 encodes a nuclear zinc finger protein that binds DNA and functions as a transcriptional repressor. *Mol Cell Biol* 16, 4024-34.