Developing a Model of *AFP* Transcriptional Regulation by *Afr2*, a Gene Implicated in Liver Cancer

By

Zachary T. Grimes

Thesis proposed to the graduate faculty of Middle Tennessee State University in partial fulfillment of the requirements for the Master of Science degree in Biology

Middle Tennessee State University

August 2017

Thesis Committee:

Dr. Rebecca L. Seipelt-Thiemann, Thesis Mentor and Committee Chair

Dr. Erin E. McClelland

Dr. Jason R. Jessen

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Rebecca Seipelt-Thiemann for not only the opportunity to work on this project, but also for her continued guidance and support throughout the course of the project and my time here. I would also like to thank Dr. Jason Jessen and Dr. Erin McClelland for agreeing to be part of my committee, and all of the duties and time that that entails. Thank you to Dr. Justin Miller for helping me to understand protein modeling programs and how to evaluate the output. I also extend special thanks to Dr. Martha Peterson and Dr. Brett Spear, from the University of Kentucky Department of Microbiology, Immunology & Molecular Genetics – their contributions helped make this project possible. I would also like to thank the entire Biology office staff for always showing their support, and offering a safe place to vent about life. I owe immense thanks to my wife, KristyAnn, for listening to me rant throughout this experience when my experiments were not working like they should have been, or I was just feeling down and she would offer advice and support, even though she did not always fully understand what I was talking about. I would like to thank my parents for encouraging me and believing in my abilities to do anything I set my mind to. Lastly, I would like to thank my fellow graduate students and friends for keeping me sane throughout the last two years, I couldn't have done it without you.

ABSTRACT

α-fetoprotein (AFP) is a protein that is active during liver development and hepatocellular differentiation that is under the transcriptional control of two regulators – *Afr1* and *Afr2*. Afr2 acts to transcriptionally reactivate *AFP* in liver regeneration and tumorigenesis. This observation led to AFP utilization as a diagnostic marker for hepatocarcinogenesis. The purpose of this study was to identify and clone *Afr2* candidate genes. To begin, the *AFP* promoter was analyzed for potential transcription factors. A genetic map of chromosome 2 was corrected and utilized to localize the candidate region more accurately. This region was analyzed for genes variant between two mouse strains (C3H/HeJ and C57BL/6) with opposite *AFP* reactivation phenotypes. Candidate genes were also identified from gene expression analyses from the same strains. These variant genes were then analyzed for interactions with potential *AFP* transcription factors and others within the candidate pool. This identified a candidate pathway containing *Ciao1*, *WT1*, and *Ywhae*.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

The Liver in Development and Disease

The liver is the largest internal organ and performs a wide variety of homeostatic functions, including the production of digestive enzymes that are important for macromolecule metabolism (NIH 2009), detoxification of harmful chemicals in the blood, removal of worn-out erythrocytes from the blood, storage of macromolecules and vitamins, and production of various hormones. One of the most distinguishing properties of the liver is it's potential for regeneration upon damage. The liver has the ability to regenerate up to 2/3 of the organ body if it is damaged or removed, while still maintaining homeostatic functions (Michalopolous 2007).

The liver's regenerative process mimics embryonic liver development, making it possible to study genes that are active during the developmental processes leading to competence. There are a multitude of genes that are implicated as having a role in both liver development and regeneration (Zaret and Grompe 2008, Shin and Monga 2013). One of the major proteins involved in the development of the fetal liver is α-fetoprotein (AFP). This protein seems to function not only in initial liver development *in utero*, but also in both the regenerative and tumorigenic processes of the liver (Tomasi 1977). In early development, AFP is present and active in the endoderm of the foregut at low levels. Once the foregut begins to differentiate under the action of the transcription factors *Foxa1* and *Foxa2*, the portion that is to become the liver begins to express higher levels of AFP. *AFP* gene expression is then transcriptionally silenced after the perinatal

period and repressed during the postnatal period due to a number of genes (*Prox1*, *Hex*, *Hlx*, *HNF4α*, *GATA6*, *HNF1β*, and *HNF6*), all of which have been shown to be transcriptional regulators of multiple hepatic genes (Spear et al 2006). However, *AFP* gene expression is transcriptionally reactivated during liver regeneration (Lin et al 2009), as well as in liver cancer (Spear 1999). Two regulators for *AFP* mRNA expression were initially found using mouse genetics: *Afr1* and *Afr2*. *Afr1* functions almost exclusively in the embryonic expression of *AFP*, while *Afr2* has been shown to control AFP levels during liver regeneration and tumorigenesis (Spear 1999). Further exploration identified *Zhx2* as the *Afr1* gene through the use of positional cloning (Peterson et al 2011). While some information has been gained about *Afr2*-mediated regulatory mechanisms, the molecular identity of *Afr2* remains unknown.

Afr2 Influences *AFP* Transcription

Jin et al (1998) identified the region of the *AFP* promoter that was required for *Afr2*-mediated regulation by constructing transgenes that deleted different portions of the *AFP* promoter and assaying for reactivation of *AFP* transcription. They found mice with a deletion between -1,010 and -838 bp produced less AFP mRNA present in the liver when regeneration was initiated by either CCl_4 intoxication or partial hepatectomy. This suggested that transcriptional regulation, rather than RNA instability, is the mechanism by which *AFP* mRNA levels are controlled.

Eukaryotic Transcription Mechanisms

Eukaryotic transcriptional control mechanisms are complex and diverse (reviewed by Lelli et al 2012) (Figure 1). Some mechanisms involve chemical changes made to the DNA, such as methylation or phosphorylation events which have the potential to alter

accessibility of regulatory regions, such as DNA being tightly bound by histones. Micro RNAs (miRNA) are also known to influence gene expression through RNA silencing post-transcription.

Specific DNA sequences, *cis*-acting sequences, that can influence transcription are promoters, enhancers, and repressors. Enhancers and repressors are bidirectional and can act over large distances. Promoters, however, are found immediately upstream of the gene's transcriptional start site. Gene-specific promoter elements act with *trans*-acting factors known as transcription factors in order to regulate transcription beyond the normal, basal level controlled by the basal transcription machinery. These transcription factors, when present, can act upon promoters either alone or by forming complexes in a larger network. In addition, binding of the same transcription factor to different sequences can alter the conformation of that transcription factor, potentially producing different transcriptional levels (Figure 1A). Regulation may take place simply due to the presence, absence, or expression level of a single transcription factor which could be due to the transcription factor being expressed in tissue-specific, developmental-specific, or other condition-specific ways. Transcription factors can also act via quaternary structures, such as dimeric and trimeric complexes (Figure 1B) with the same or different binding partners. Homodimers and heterodimers are also known to bind different DNA sequences, thereby regulating different genes (Kosugi and Ohashi 2002). One variation on transcription factor multimerization is the formation of enhanceosomes as a complex. Each transcription factor in the complex is sequence-specific, but the complex only binds as a single unit (Figure 1C). Finally, one transcription factor can also act as an anchor to recruit other non-DNA binding proteins that influence transcription (Figure 1D).

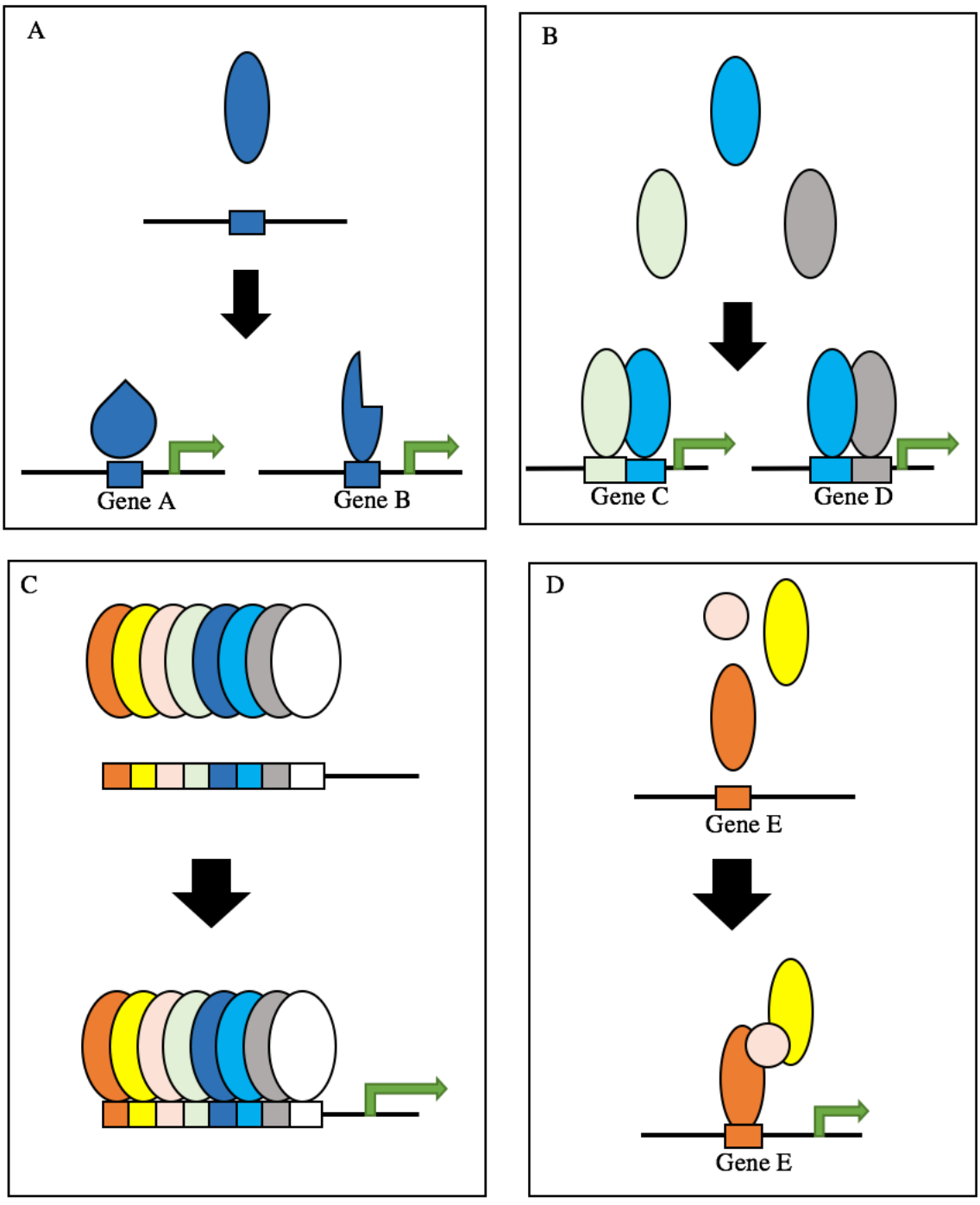

Figure 1. Selected Mechanisms of Eukaryotic Transcriptional Regulation
A) Conformational change in TF due to DNA allosteric effects. **B)** Common TF activating multiple genes through different binding partners. **C)** Multiple TFs forming an enhanceosome complex. **D)** Transcription is activated through a non-DNA binding complex recruited by initial seeding of a DNA binding transcription factor. Adapted from Lelli et al 2012.

Afr2 Identification Attempts

Classical mouse genetic experiments were used to map *Afr2* to the long arm of chromosome 2. Mouse strains C3H/HeJ and C57BL/6 both exhibit the expected repression of *AFP* postnatally. *AFP* mRNA levels are high in C3H/HeJ mice with liver damage. In contrast, upon liver damage and initiation of regeneration, *AFP* mRNA levels are low but detectable in C57BL/6 mice. The C57BL/6 (low AFP) strain is resistant to liver tumors, while the C3H/HeJ (high AFP) strain is more susceptible to liver tumors. In the mapping experiment, C3H/HeJ and C57BL/6 mice were mated and the heterozygous offspring were backcrossed to C57BL/6. The backcross offspring were genotyped for markers across the entire genome and their ability to reactivate *AFP* gene expression after liver damage was assayed. This data revealed that *Afr2* was located between the end of marker D2Mit398 and the beginning of D2Mit224, genomic locus range chr2:125,725,767-129,288,472. However, no additional attempts to identify Afr2 have been successful.

In order to develop a model of AFP transcriptional reactivation by *Afr2*, it was first necessary to investigate the identity of *Afr2*. To this aim, a thorough and systematic approach was developed to identify candidate genes which subsequently led to the development of a model of *AFP* transcriptional regulation involving three proteins: Ciao1, WT1, and Ywhae, which may act via the anchoring model (Figure 1D).

Research Goal and Strategies

The overarching goal for this thesis project was to identify *Afr2* gene candidates using bioinformatics analyses, existing gene expression data, computational analyses, and then to clone the candidates for later functional experiments.

The specific strategies used in this project were:

Strategy 1: Utilize existing strain variation to build a list of candidate genes based upon genetic variation of the target region on chromosome 2.

Strategy 2: Analyze existing microarray data and build a list of candidate genes based upon mRNA expression levels when comparing:

1) C3H/HeJ and C57BL/6 in liver damage, and

2) C3H/HeJ quiescent to regenerating liver.

Strategy 3: Perform combinatorial and computational biology analyses in order to narrow the pool of candidate genes.

Strategy 4: Clone the candidate gene(s) to be used in later functional experiments.

II. MATERIALS AND METHODS

BIOINFORMATICS ANALYSIS

To identify the *Afr2* gene, a bioinformatics approach was used to perform a thorough analysis of all possible candidate genes and the *AFP* promoter.

AFP Promoter Analysis

Sequence for the *AFP* promoter region linked to *Afr2* regulation (-1,010 to -838 bp upstream of the *AFP* initial start codon) was retrieved for both the C57BL/6 and C3H/HeJ strains from the University of California at Santa Cruz (UCSC) Genome Browser (Kent et al 2002, Karolchik et al 2004) using genome build GRCm38/mm10, and the Sanger Genome Evaluation Browser (gEVAL) (Chow et al 2016), respectively. The specific coordinates used for sequence retrieval were 5:90489727-90489899. The sequences were then aligned using ClustalOmega (Sievers et al 2011), and shaded using BOXSHADE to easily identify differences between the sequences. Both sequences were then analyzed for possible transcription factor binding sites using the DNA binding site algorithm PROMO (Messeguer et al 2002).

Linkage Map Correction

The linkage map generated by Jin and Feuerman (1997) was corrected by comparing the coordinates of published marker loci found in the Mouse Genome Informatics (MGI) database (Blake et al 2017) to the coordinates in the original study.

Predicted Function of Variant Genes

Both the MGI database and the Sanger Wellcome Trust Institute (Keane et al 2011, Yalcin et al 2011, respectively) were used to locate genes with variation in the

region from 125,725,472-129288472 on the mouse chromosome 2 including single nucleotide polymorphisms (SNPs), insertions/deletions (InDels), and structural variants (SVs). Gene Ontology (GO) terms for each of the genes was identified using Homologene within the National Center for Biotechnology Information (NCBI).

Reference sequences for genes without identified GO terms were retrieved from NCBI (Table 1), and were subsequently translated into amino acid sequences using ExPASy Translate (Artimo et al 2012). Amino acid sequences for genes which resulted in proteins >100 amino acids in length were then analyzed for protein domains and motifs using both SMART (Schultz et al 1998) and InterProScan (Jones et al 2014) tools to gain information regarding possible functions for those genes initially lacking any GO terms. Those genes with identified GO terms and/or predicted protein domains indicating transcription and/or RNA stability related functions were carried into the next phase of analysis. Any results that indicated possible miRNAs were analyzed using miRBase (Ambros et al 2003) to check for miRNA interactions with candidate genes.

Table 1. Reference Identification Numbers for Candidate Genes

Candidate Gene	NCBI Reference ID
Anapc1	NM_008569.2
Blvra	NM_026678.4
Chchd5	NM_025395.3
Ciao1	NC_000068.7
Cops2	NM_001285507.1
Gabpb1	NM_001271467.1
Il1a	NM_010554.4
Mrps5	NM_029963.2
Prom2	NM_138750.2
Slc27a2	NM_011978.2
Stard7	NM_139308.2
WT1	NM_144783.2
Ywhae	NM_009536.4
Zc3h8	NM_020594.2
Zfp661	NM_001111029.1

Candidate genes and corresponding reference sequences used in the developed analytical pipeline for genes without identified GO terms.

Microarray Data Mining

Unpublished microarray data from M. Peterson/B. Spear were evaluated to identify genes upregulated in liver regeneration. In order to produce the RNA used in the microarray experiment, mice (specifically C3H/HeJ and C57BL/6) were inoculated intraperitoneally with either 50 μL of mineral oil (MO) with 10% carbon tetrachloride (CCl_4), or 50 μL MO, as described by L. Morford (2007).

Pathway Analysis for Protein-protein Interactions

Two related, but distinct, methods were used to screen candidate proteins for previously known or predicted direct interactions with the potential *AFP* transcription factors. First, the candidate genes were searched by name in a pathways and interactions database, STRING, that uses published experimental evidence (Artimo et al 2012) to identify direct interactions between candidate proteins and transcription factors that were predicted to bind in the -1,010 to -838 region of the *AFP* promoter. Another database specifically for interactome analyses, MENTHA, was also used (Calderone et al 2013). This generates maximum likelihood interactome pathways between the multiple genes used as input using published experimental evidence.

Protein Modeling of the Candidate Pathway

Due to the lack of available murine protein structure models, SWISS-Model (Altimo et al 2012) was used to generate homology models for the candidates and putative interacting proteins identified within the appropriate region on chromosome 2 in both the variation screen and microarray data mining. Amino acid sequences were loaded into the SWISS-Model server which searches the Protein Data Bank (PDB) and other protein databases for similar sequences with available models. These model templates

were then either included or excluded from further analysis based upon sequence identity/similarity, which are evaluated 1-100, and total coverage of the target sequence. Candidate genes that were used in modeling are shown in Table 2 with the corresponding SWISS-Model template IDs. The candidate pathway was modeled in the structural modeling software CLC Drug Discovery Workbench (Qiagen) to examine possible interactions between the proteins in the pathway.

Table 2. Template Identification Numbers for Candidate Protein Models

Candidate Gene	SWISS-Model Template ID
Anapc1	5g05.1.A
Blvra	1lc3.1.A
Chchd5	2lql.1.A
Ciao1	3fm0.1.A
Il1a	2l5x.1.D
Mrps5	5aj4.3.A
Stard7	1ln1.1.A
WT1	2i13.1.C
Ywhae	2br9.1.A
Zfp661	5eh2.1.C

Selected models for each candidate gene on basis of sequence identity and total coverage ultimately used in protein model analysis.

MOLECULAR CLONING OF CANDIDATE GENES

RNA

Mouse RNA samples were kindly provided by Drs. M. Peterson and B. Spear and included RNA from C3H/HeJ and C57BL/6 mouse livers treated with CCl_4 or mineral oil, and samples from liver tumors. Before the isolation was performed, the RNA samples were checked for quality. The samples were fractionated on a 1.5% formaldehyde gel and viewed under UV light.

RT-PCR and PCR

Intact samples were used in reverse transcription polymerase chain reaction (RT-PCR) to produce a pool of cDNA using M-MLV Reverse Transcriptase as directed by the manufacturer (Invitrogen).

Primers for *Ciao1*, *WT1*, and *Ywhae* (Table 3) were designed either by using Primer3Plus (Untergasser et al 2007) or manually and checked for hairpin formation in OligoCalc (Kibbe 2007). Optimization of the primer sets was performed through gradient PCR using the Phusion Polymerase protocol provided by the manufacturer (Thermo Scientific), using annealing temperatures of 50.0, 51.5, 53.9, 57.5, 52.2, 66.0, 68.5, and 70.0 degrees Celcius.

The fragments were fractionated on a 1% agarose gel using the Log2 Ladder from New England Biolabs (NEB) as the standard for determining size.

Table 3. Primer Information for Candidate Genes

Gene Name	Primer Direction	Primer Sequence	Added Restriction Sites	Primer Design Method	Annealing Temperature (°C)	Expected Fragment Size
Ciao1	Forward	5′CACACGAATTCCATGAAAGATTCTCTGGTACT3′	BamHI	By hand, checked in OligoCalc	68.5	1,020 bp
	Reverse	3′GTGTGGGATCCTCAGAGACCTGCAGGCTGGTGAT5′	SalI			
WT1	Forward	5′CACACGTCGACCATGGGTTCCGACGTGCGGGA3′	XhoI		50.0	1,553 bp
	Reverse	3′GTGTGCTCGAGTCAAAGCGCCAGCTGGAGTT5′				
Ywhae	Forward	5′AGACGCTATCCGCTTCCAT3′	None	Primer3Plus	62.2	767 bp
	Reverse	3′TGGTTTCTCTTGTTGGCTTTT5′				

The primer information for each of the genes in the candidate pathway including the sequences of each primer, added restriction sites, the source of the primer sequence, annealing temperature found through the use of Gradient PCR, and the expected PCR fragment size.

Ligation and Transformation

PCR products were ligated individually into the pCR-Zero BluntII TOPO vector, as directed by the manufacturer (Invitrogen), though the initial ligation was incubated at room temperature for 60 minutes instead of the recommended 5 minutes. Ligations were used to transform One Shot® TOP10 chemically compotent *E.coli* cells. Transformed bacteria were grown on tryptic soy agar (TSA) containing kanamycin (100 μg/mL).

Transformant Screening

Kanamycin-resistant colonies were both patched onto TSA plates containing kanamycin, at the concentration above, and inoculated into 2 mL TSA broths plus kanamycin. Liquid cultures were grown in a 37 °C shaking incubator at 250 rpm for 16 hours. Plasmid DNA was then isolated according to Green and Sambrook (2012), and screened via restriction enzyme analysis. Initial digestion was done with EcoRI in order to determine which plasmids contained inserted fragments. Those colonies that contained inserts approximating the expected sizes were then digested with AvaII, which has multiple cut sites in the candidate genes and within the vector, as determined through the use of RestrictionMapper. Expected fragment sizes were calculated by hand using the AvaII restriction site locations (GG[A/T]CC) (Table 4).

Table 4. Expected Fragment Sizes from AvaII Restriction Digest.

Candidate Gene	Expected fragment sizes (bp)
pCR Zero BluntII TOPO + *Ciao1*	1807, 1539, 552, 424, 123, 55, 33, 6
pCR Zero BluntII TOPO + *WT1*	1602, 1567, 644, 424, 200, 168, 142, 112, 111, 55, 33, 6
pCR Zero BluntII TOPO + *Ywhae*	2148, 1621, 424, 55, 33, 6

III. RESULTS

AFP is a protein that is active in initial embryonic hepatocellular differentiation as well as liver regeneration in the adult and tumorigenesis. Elucidation of the mechanism by which transcriptional reactivation of *AFP* occurs has important implications for both organ regeneration and cancer. Two different regulators were genetically identified in embryonic and adult *AFP* RNA expression: *Afr1* and *Afr2*.

To begin a detailed study into the molecular identification of *Afr2*, an analysis of the previous mapping data was undertaken. The original map produced from backcrossing experiments (Jin and Feuerman 1997) was found to be in error based upon current loci coordinates for the genetic markers used in the mapping study (Table 5). The markers were identified and used to reconstruct the map (Figure 2) to reveal that *Afr2* was actually located between markers D2Mit 398 and D2Mit 208. Upon correction of the linkage map, the approximate size of the region of interest decreased from 9,554,863 bp to 3,562,705 bp.

Table 5. Genetic Markers on Chromosome 2 used to Determine the Potential Location of *Afr2*

Marker	Locus Coordinates on Chr2
D2Mit 92	71579623..71579831
D2Mit 185	105495600..105495807
D2Mit 444	118774657..118774901
D2Mit 398	125725565..125725767
D2Mit 304	128331760..128331925
D2Mit 208	135280630..135280893
D2Mit 224	129288472..129288585
D2Mit 258	130407596..130407742
D2Mit 280	146051794..146052073
D2Mit 281	148532836..148533044
D2Mit 423	148825388..148825577
D2Mit 451	155649632..155649793

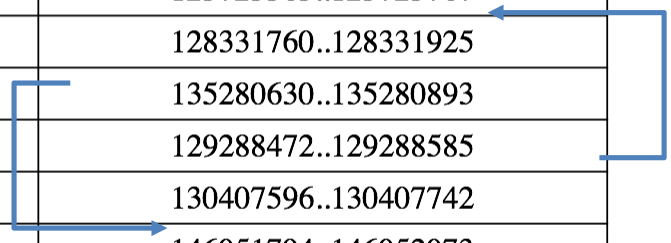

Table of marker coordinates listed in the order of the original linkage map from Jin and Feuerman 1997. Arrows indicate where markers were moved according to published coordinates.

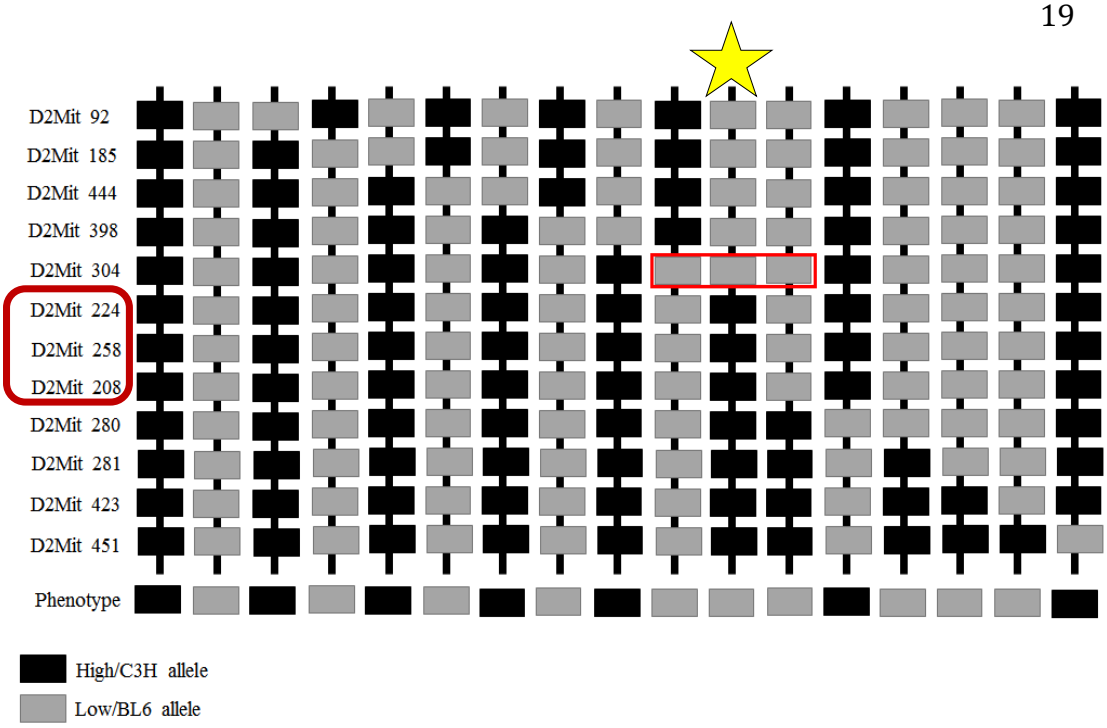

Figure 2. Corrected Linkage Map used to Identify the Region of Interest. Each column of boxes represents a recombinant chromosome. Markers are listed at the left. Boxes at the bottom represent the phenotype of the mouse carrying the chromosome. The red box at the left indicates the markers that were out of order in the original study, here showing them in the corrected order. The red box within the figure shows the crossover event that resulted in a change in phenotype, and by extension the location of *Afr2*. The star indicates the chromosome generated by the crossover event where the phenotype change occurred.(Corrected from Jin and Feuerman 1997).

Variation within this region and/or the *AFP* promoter between C57BL/6 and C3H/HeJ must therefore be responsible for the different *AFP* phenotype in these mouse strains. First, variation in the *Afr2*-specific region of the *AFP* promoter between the strains was investigated. The -1,010 bp to -838 bp region of the *AFP* promoter from both strains was aligned and revealed many differences (Figure 3). To gain better insight into which transcription factor might differentially bind each promoter, potential DNA binding sites were then analyzed for both promoters using Alggen PROMO. The resulting transcription factors were then searched within MENTHA, and only those with mouse homologs were retained, resulting in 75 possible transcription factors (Table 6).

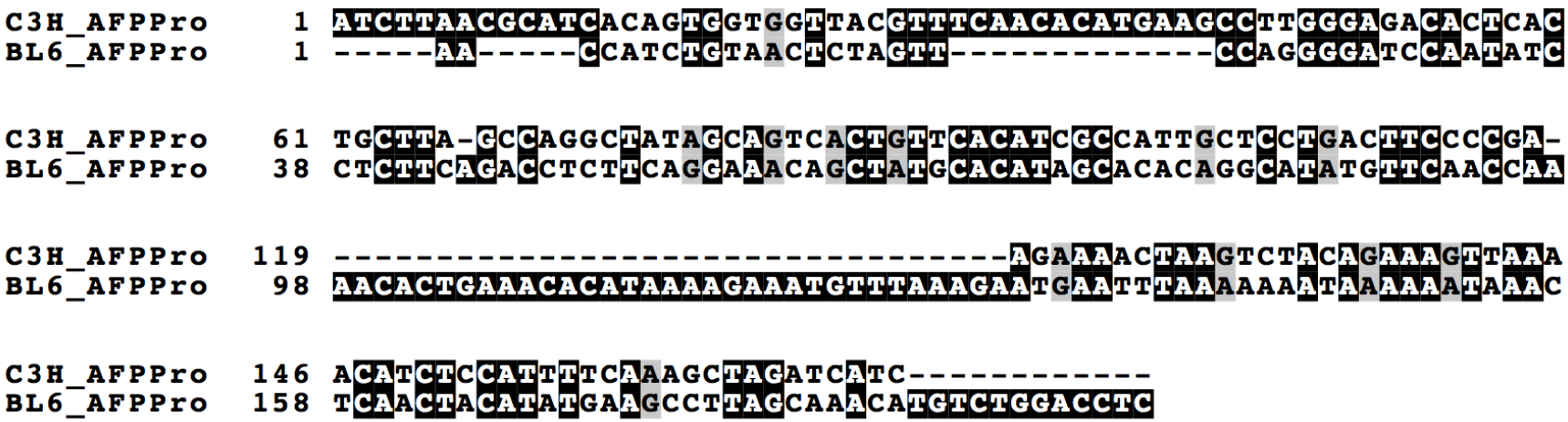

Figure 3. Alignment of the *AFP* Promoter in both Mouse Strains.
C3H/HeJ (above as C3H_AFPPro) and C57BL/6 (above as BL6_AFPPro).
Black boxes identify matches, grey indicates transition variants (A<>G, C<>T), white indicates insertions/deletions or transversion variants (purine<>pyrimidine). Dashes indicate no matching sequence.

Table 6. Potential *AFP* Transcription Factors Identified from *AFP* PROMO Analysis

Possible AFP Transcription Factors					
PROMO Gene ID	MENTHA Gene ID	PROMO Gene ID	MENTHA Gene ID	PROMO Gene ID	MENTHA Gene ID
AIRE		*GATA-2*		*Nxk2-2*	
AP-2	*AP2S1*	*Helios*	*IKZF2*	*p300*	*EP300*
AR		*HNF-1a*		*Pax-2*	
ARF1		*HNF-3α*	*FOXA1*	*Pax-5*	
ATF3		*HNF-3β*	*FOXA2*	*Pax-6*	
BTEB3	*KLF13*	*HNF-3γ*	*FOXA3*	*Pax-8*	
c-Ets-1	*ETS1*	*HOXA3*		*PKNOX1*	
c-Ets-2	*ETS2*	*HOXD10*		*RelA*	
c-Fos	*FOS*	*HOXD9*		*Smad3*	
c-Jun	*JUN*	*IRF-1*		*Smad4*	
c-Myb	*MYB*	*IRF-2*		*STAT1*	
c-Myc	*MYC*	*IRF-3*		*STAT3*	
Cdx-1	*CDX1*	*JunB*		*STAT4*	
CDX2		*JunD*		*STAT5A*	
COE1	*EBF1*	*LEF-1*		*STAT5B*	
DBP		*LIM-1*	*LHX1*	*STAT6*	
Deaf-1		*Lyf-1*	*IKZF1*	*TBP*	
E47	*TCF3*	*MAC1*	*PTEN*	*TCF-3*	
EBF	*EBF1*	*MED8*		*TGGCA-binding protein*	*NF1C*
Elf-1		*MitF*		*USF-1*	
Elk-1		*MSX1*		*USF2*	
FOXJ2		*MYOG*		*WT1*	
FOXO2		*NF-AT1*	*NFATC2*	*Zic1*	
FOXP3		*NF-AT2*	*NFATC1*	*Zic2*	
GATA-1		*Nkx2-1*		*Zic3*	

The transcription factors listed are those with mouse homologs found in the MENTHA Interactome Browser. Those with grey boxes indicate that the same name is used in both Alggen and MENTHA databases.

Genomic localization of the transcription factors showed that none of transcription factors were within the mapped region identified as the location of *Afr2*. Based on this analysis and the knowledge that transcription factors can act in a cooperative manner with non-DNA binding regulators, two approaches were undertaken. First was the mining of unpublished microarray data (Peterson, unpublished). Second was an investigation of variation within the mapped region of chromosome 2.

Eight genes within the designated region of chromosome 2 were found to be upregulated in C3H/HeJ CCl_4 compared to MO treated and five genes were upregulated in C3H/HeJ CCl_4 treated compared to C57BL/6 CCl_4 treated (Figure 4A; Table 7). The variation analysis using the Sanger database returned 26,622 SNPs, 4,919 InDels, and 157 Structural Variants. Ninety-two variant genes were in the region of interest. The variation analysis using the Mouse Genome Institute (MGI) returned 2,002 SNPs. Ninety-nine were in coding regions (Figure 4B). All genes from the MGI search were also found in the Sanger search, so the resulting lists were merged (Figure 4B; Table 8). In total, this represented the initial candidate gene pool.

All candidate genes were then searched for either transcriptional or RNA stability-related functions using an ontology search at NCBI. Genes with known function related to RNA level regulation (either transcriptional or RNA stability), along with those whose function was unknown were retained within the candidate pool (Figure 4C; Tables 7 and 8)

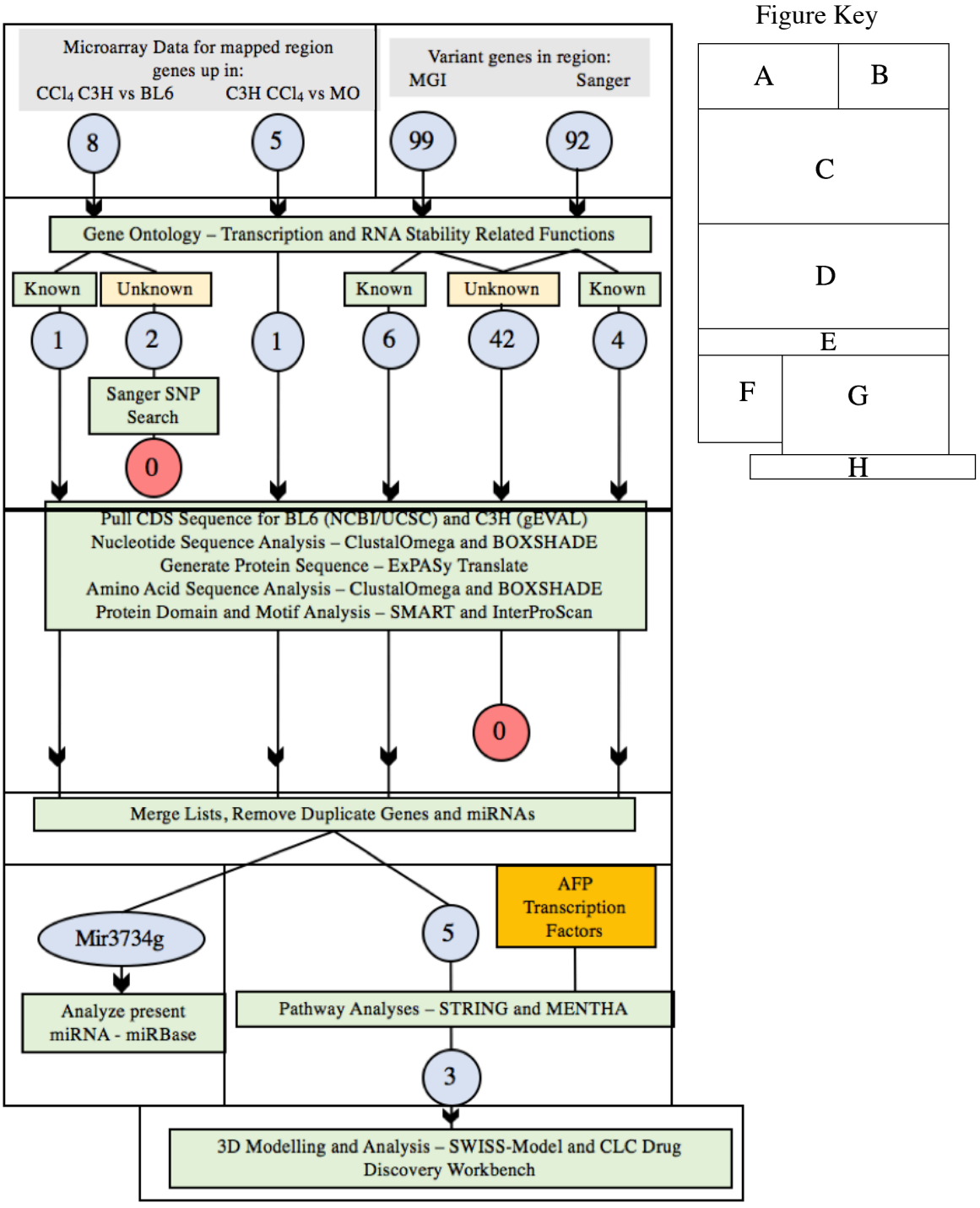

Figure 4. Bioinformatics Analytical Pipeline.

Left) Flowchart depicting the bioinformatics analytical pipeline that was developed.
Right) Figure key for the analytical flowchart.

Table 7. Candidate Genes Resulting from the Microarray Data Mining

Genes Upregulated in C3H/HeJ CCl_4 vs MO	Genes Upregulated in CCl_4 C3H/HeJ vs C57Bl/6
Chchd5	*Anapc1*
Ciao1	*Bcl2l11*
Mrps5	*Blvra*
Slc27a2	*ENSMUSG00000074822*
Stard7	*Il1a*
	OTTMUSGm14005
	Prom2
	Zfp661

Candidate genes located in the region of interest on chromosome 2 which are upregulated in C3H/HeJ following liver damage. Orange boxes indicate no known ontology. Green boxes indicate RNA stability-related ontology. Blue boxes indicate transcription related ontology.

Table 8. Combined Pool of Candidate Genes from the Variation Analyses

Combined Pool of Candidate Genes		
0610042E11Rik	*Gm14006*	*Gpat2*
1500011K16Rik	*Gm14007*	*Hdc*
1810024B03Rik	*Gm14008*	*Itpripl1*
4930402C16Rik	*Gm14009*	*Mal*
4933427J07Rik	*Gm14010*	*Mal1*
9830144P21Rik	*Gm14011*	*Mertk*
A730036I17Rik	*Gm14012*	*Mir3473g*
AcoxI	*Gm14022*	*Mrps5*
Adra2b	*Gm14024*	*Ncaph*
AI847159	*Gm14025*	*Nmf220*
Anapc1	*Gm14026*	*Nphp1*
Ap4e1	*Gm14027*	*Polr1b*
Astl	*Gm14028*	*Prom2*
Atp8b4	*Gm14029*	*RP23-160G19.10*
Bcl2l11	*Gm14212*	*RP23-206D14.7*
Blvra	*Gm14229*	*Secisbp2l*
Bub1	*Gm14244*	*Shc4*
Chchd5	*Gm14245*	*Slc27a2*
Ciao1	*Gm17555*	*Snrnp200*
Ckap2l	*Gm22411*	*Spdye4c*
Cops2	*Gm22613*	*Sppl2a*
Dtwd1	*Gm22859*	*Stard7*
Dusp2	*Gm22889*	*Tmem127*
Fahd2a	*Gm23101*	*Tmem87b*
Fam227b	*Gm23172*	*Trpm7*
Fbln7	*Gm23752*	*Ttl*
Fgf7	*Gm24739*	*Usp50*
Gabpb1	*Gm26496*	*Usp8*
Galk2	*Gm26697*	*Zc3h6*
Gm10762	*Gm27003*	*Zc3h8*
Gm10774	*Gm29010*	*Zfp661*
Gm14005	*Gm335*	

Candidates that are variant between C3H/HeJ and C57BL/6. Orange boxes indicate genes with no known ontology data. Green boxes indicate RNA stability related functions. Blue boxes indicate Transcription related functions.

Nucleotide sequences for genes with no known ontology data were translated and the amino acid sequences were used to search protein domain databases for clues into the function of these unknown genes (Figure 4D). None of the 42 genes that were analyzed were found to have domains that would indicate RNA stability or transcription related functions. The candidate pools were therefore narrowed to six genes that had known transcription related ontology (Figure 4E; Table 6). *Mir3473g* is a miRNA within the region of interest on chromosome 2 and was analyzed in miRBase for potential interactions with other candidates (Figure 4F). No interactions were found.

Since none of the potential *AFP* transcription factors were located within the appropriate region of chromosome 2, the candidate genes were searched using the STRING database for direct interactions with the potential *AFP* transcription factors (Figure 4G; Figure 5). These returned no evident, direct interactions with the potential *AFP* transcription factors identified in the earlier DNA binding predictions using PROMO (Appendices 1 and 2; Table 4). While STRING only searches for direct interactions, another database, MENTHA, searches for possible interactions across nodes and builds potential pathways between search terms. So, the search was repeated in the MENTHA interactome browser (Figure 4H; Figure 6). When evaluating the candidate pathways identified within the MENTHA database, candidate pathways containing both a candidate protein and at least one potential *AFP* transcription factor(s). This produced two viable candidate pathways, one involving Ciao1, the other Zfp661. Through the addition of Zhx2 to the analysis, the best candidate was determined to be the pathway containing Ciao1, Ywhae, and WT1. (Figure 6).

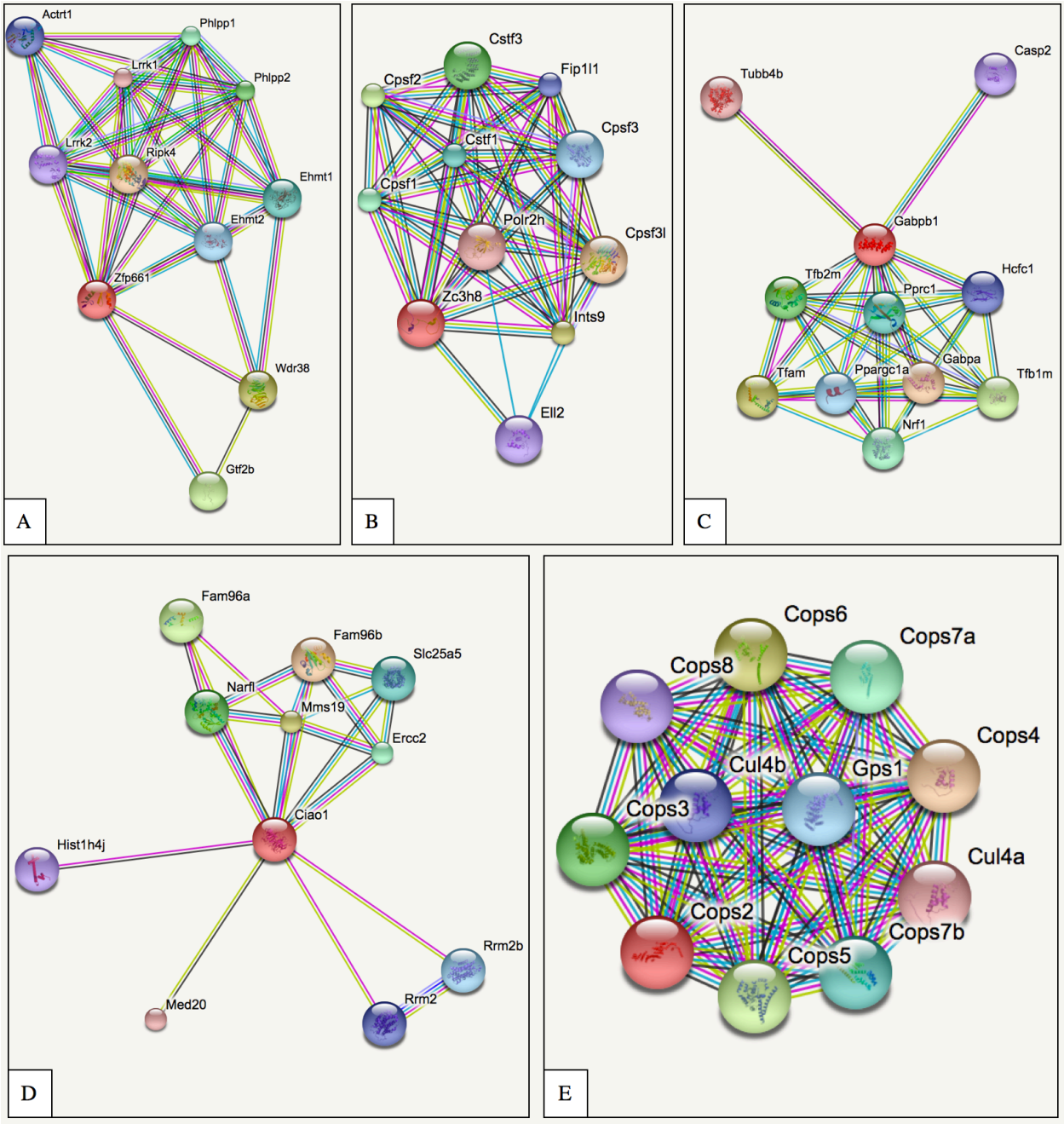

Figure 5. Direct Interactions for the Five Transcription Factor Candidate Genes from the STRING Database. A)*Zfp661*, **B)** *Zc3h8,* **C)** *Gabpb1,* **D)** *Ciao1,* and **E)** *Cops2.*

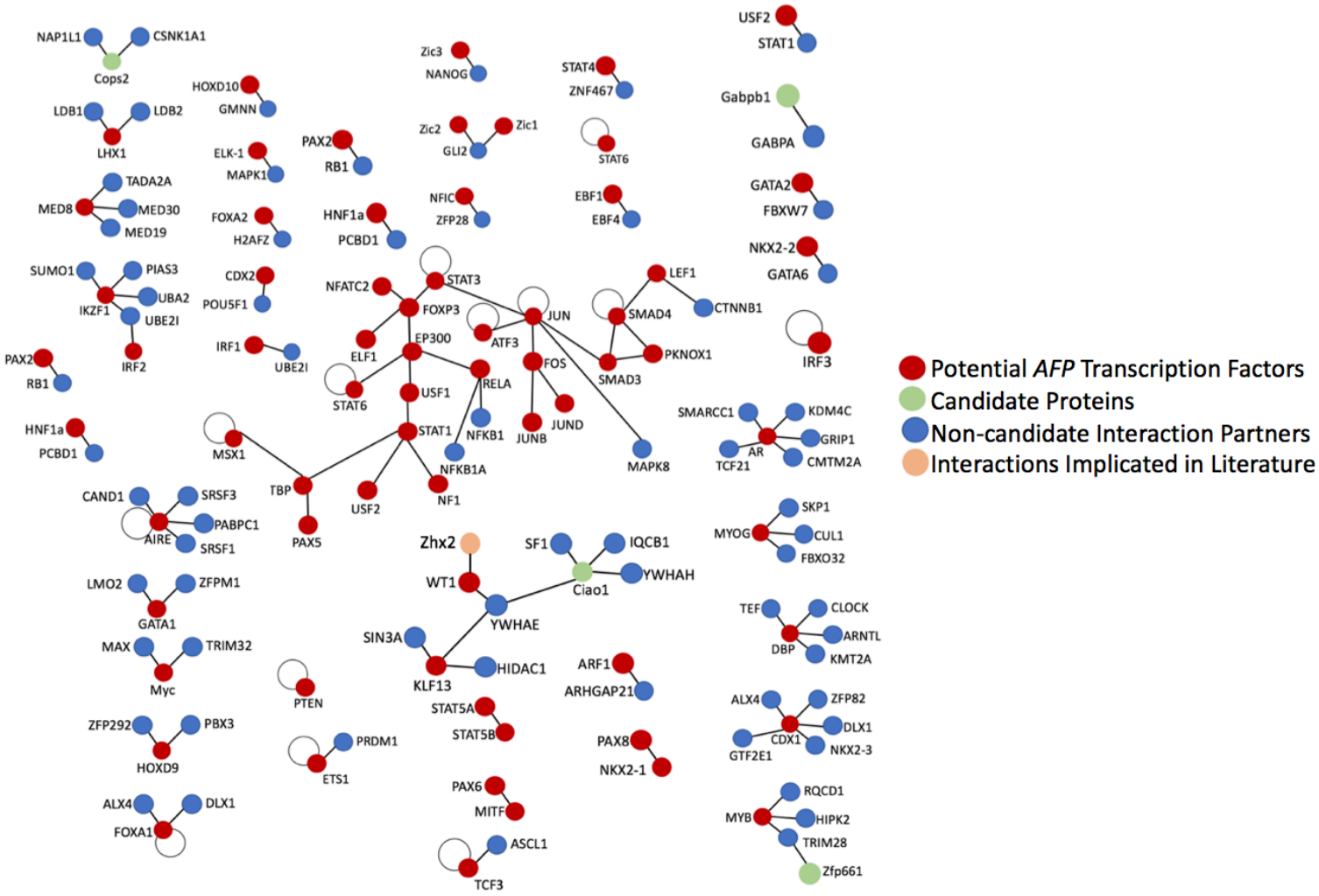

Figure 6. Predicted Interactome Pathways for Each of the Potential *AFP* Transcription Factors and Candidate Proteins with Transcription-Related Functions.

Ciao1 was subsequently analyzed for variance in the coding region between the two strains. Only synonymous variants were present, so the promoters were analyzed as far as to 2000 bp upstream (chr2:127247817-127249816) (Figure 7).

Once a candidate pathway was identified, a protein modeling approach was taken to gain additional information about specific protein-protein interactions between the members of the complex. Due to the lack of mouse protein models, homology models were generated from the SWISS-Model server for Ciao1, WT1, and Ywhae, (Figure 8; Figure 4H). When evaluating the quality of the model, sequence identity (similarity of the input sequence to the model sequence) as well as overall sequence coverage (amount of sequence that the generated model represents) were assessed and those with the highest quality in both were used in the model production. The models generated were also analyzed in the CLC Drug Discovery Workbench (Figure 4H). This allowed for manipulation and identification of individual portions of the models, specifically the: 1) identification of binding pockets in each one that could act as a docking site for the others, 2) identification of the predicted surface architecture, and 3) identification of the predicted electrostatic surface (Figures 9-11, respectively).

In order to model docking of one protein with another, ligands, which are surface contacts, must be extracted from the models. The Workbench failed to identify and extract ligands from the Ciao1 and Ywhae models. Zinc ligands were extracted from WT1, but the zinc ligands in WT1 are bound to DNA and therefore not able to participate in protein-protein interactions. This is not surprising since the WT1 model includes only the C-terminus, which is bound to DNA. With no ligands extracted from Ciao1 or Ywhae, the two could not be docked and analyzed. Although these models provide a

glimpse into the tertiary structure of each protein, they could not be used to further examine potential sites of protein-protein interaction.

Figure 7. Alignment of *Ciao1* Promoter in both Mouse Strains. C3H/HeJ is shown as C3H_Ciao_Pro, and C57BL/6 is shown as BL6_Ciao_Pro.

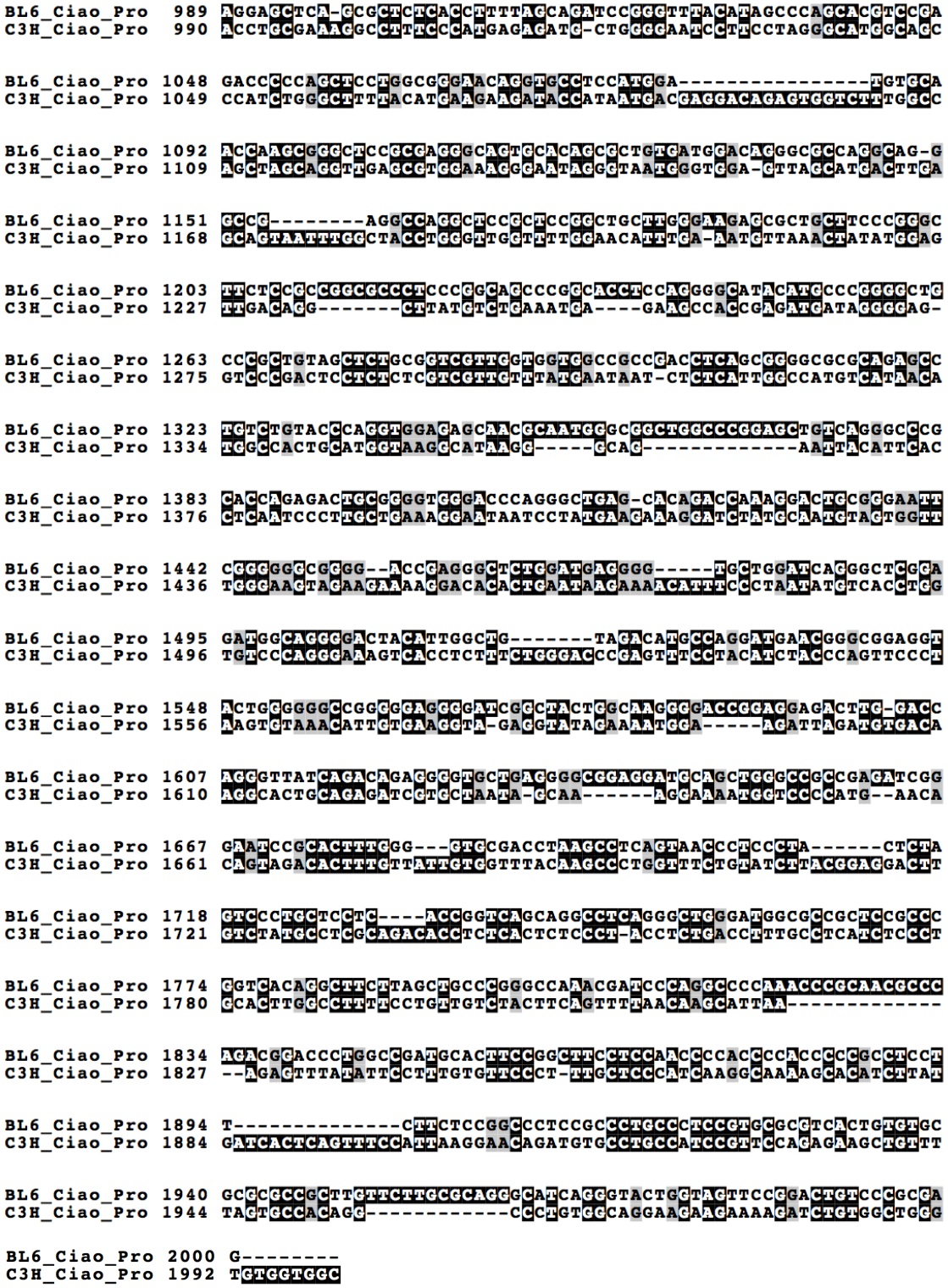

Figure 7. Alignment of *Ciao1* Promoter in both Mouse Strains (continued). C3H/HeJ is shown as C3H_Ciao_Pro, and C57BL/6 is shown as BL6_Ciao_Pro.

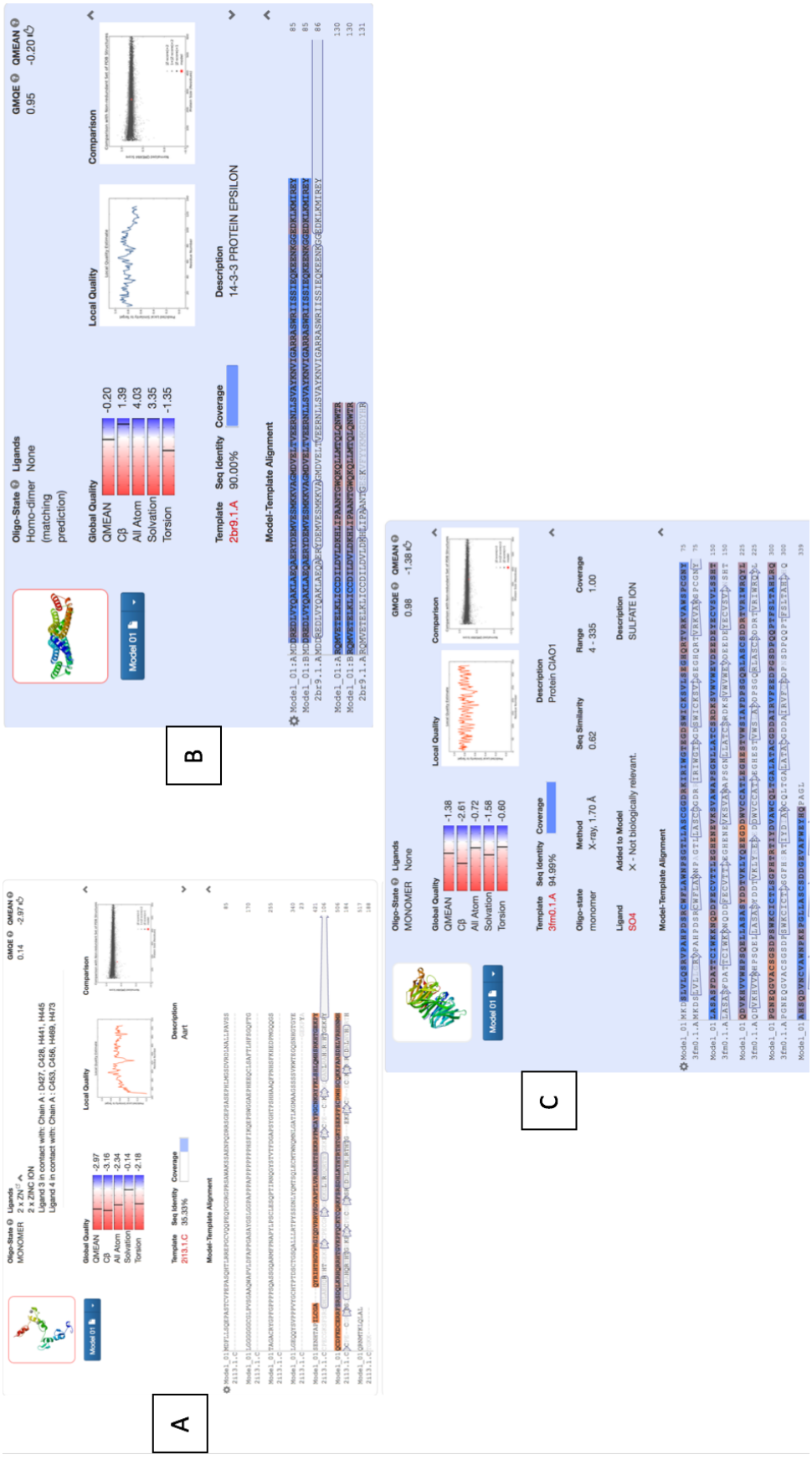

Figure 8. Templates Selected to be used in the Production of Homology Models. A) Selected template to be used for WT1 modeling. **B)** Selected template for Ywhae modeling. **C)** Selected template for Ciao1 modeling.

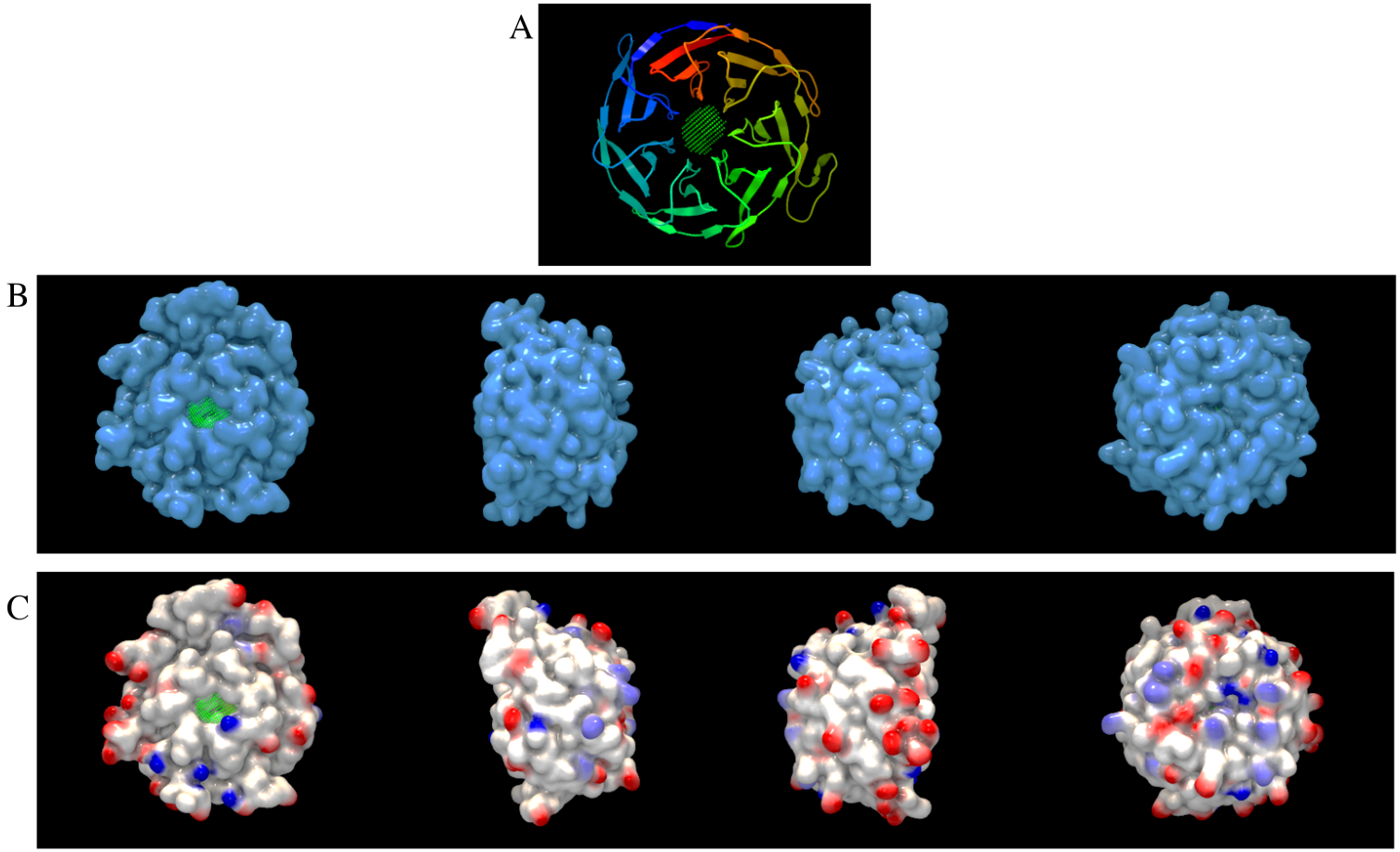

Figure 9. Modeling and Binding Pocket Analyses of Ciao1. Binding pocket is indicated by green dots. **A)** Ciao1 ribbon structure. **B)** Predicted surface of Ciao1. **C).** Predicted electrostatic surface of Ciao1.

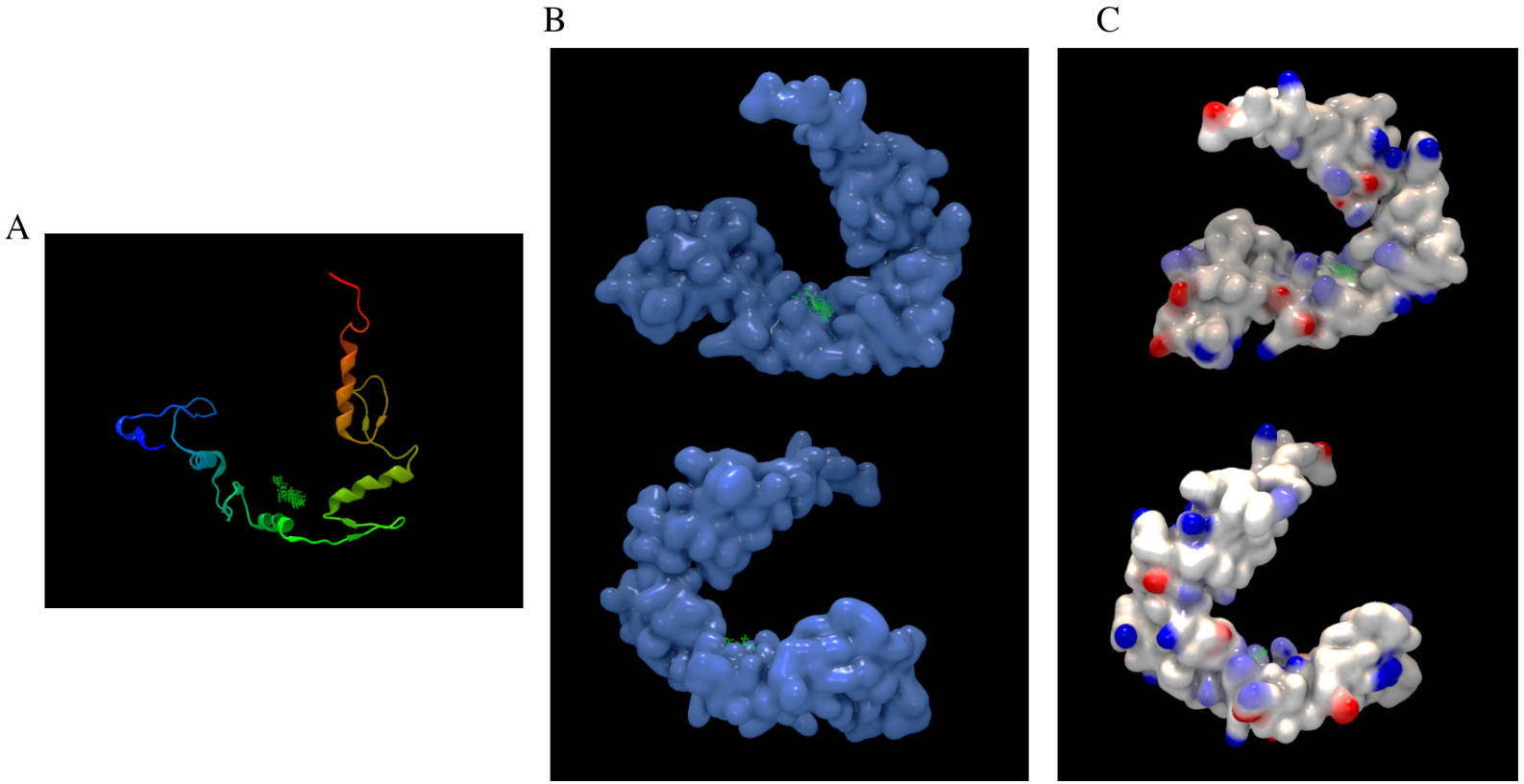

Figure 10. Modeling and Binding Pocket Analyses of WT1 C-Terminus. Binding pocket is indicated by green dots. **A)** WT1 ribbon structure. **B)** Predicted surface of WT1. **C)** Predicted electrostatic surface of WT1.

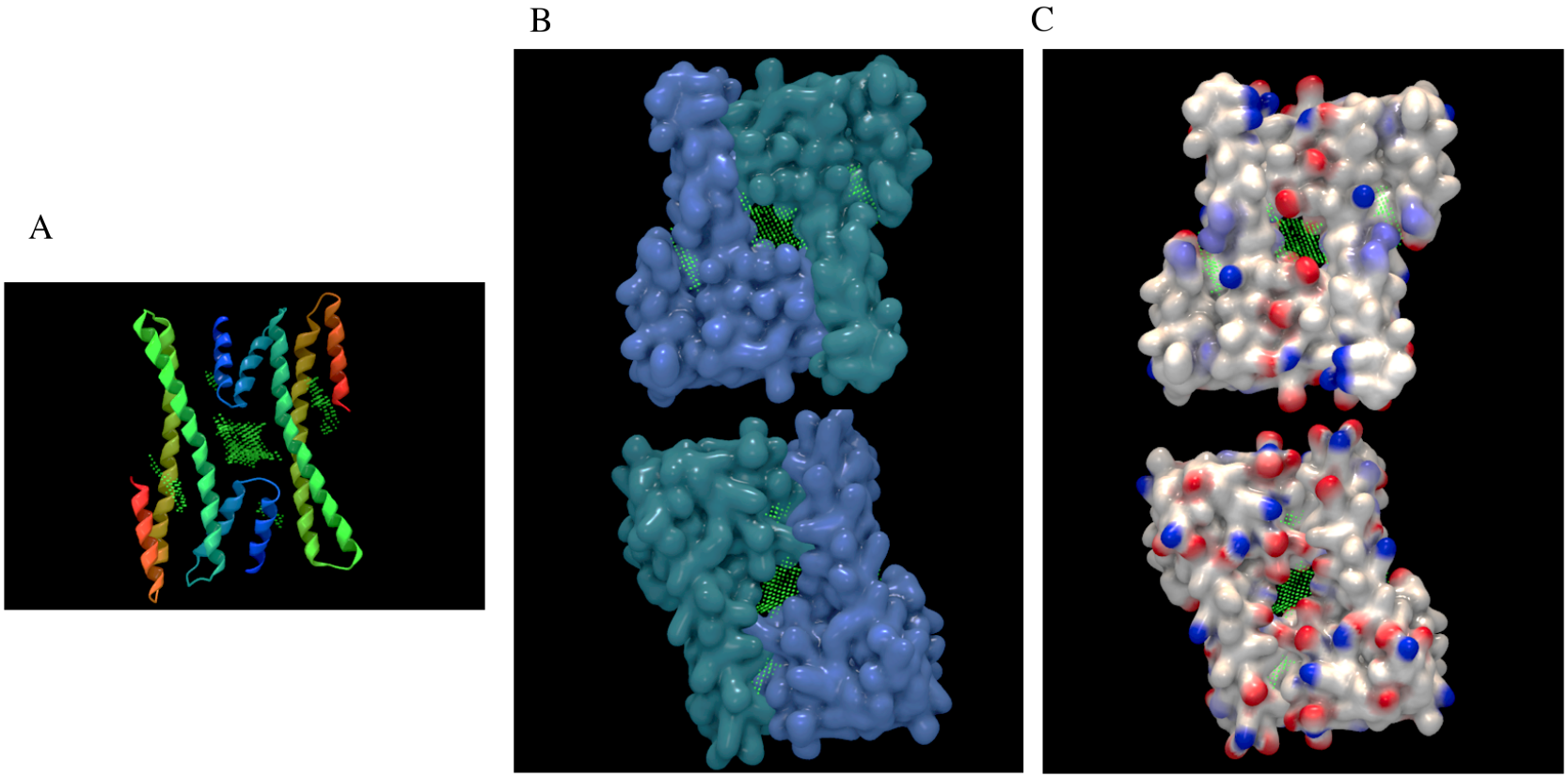

Figure 11. Modeling and Binding Pocket Analyses of Ywhae. Binding pocket is indicated by green dots. **A)** Ywhae ribbon structure. **B)** Predicted surface of Ywhae. **C)** Predicted electrostatic surface of Ywhae.

With the identification of the pathway, a molecular cloning approach was undertaken in order to clone each of the genes within the pathway (*Ciao1*:*Ywhae*:*WT1*). RNA was isolated from livers of mice that were treated with either mineral oil or CCl_4, and five livers tumor samples were obtained from Dr. Brett Spear. RNA samples were checked for quality on a formaldehyde agarose gel (Figure 12). Due to C3H/HeJ being the high expressing mouse, intact C3H/HeJ RNA samples from CCl_4-treated livers that exhibited both 28S and 18S bands on the formaldehyde gel were used due to generate cDNAs. Gradient PCR was used to optimize primer annealing temperatures for each of the primer sets for each candidate gene in the pathway. PCR products were then separated on a 1% agarose gel (Figure 13), to determine optimal annealing temperatures.

Once annealing temperatures were optimized, the PCR fragments were ligated into the vector pCR-Zero Blunt II TOPO. Ligations were transformed into *E.coli* and colonies were selected using kanamycin.

Colony growth indicated positive transformation results since this vector induces bacterial death if no insert is ligated. Eight colonies of each gene (*Ciao1*, *WT1*, and *Ywhae*) were selected for further screening. Plasmid DNAs isolated from each colony were first treated with EcoRI. This enzyme cuts the plasmid on either side of the insert so that the fragment is released (Figure 14). Colonies positive for insert were then cleaved with AvaII which cuts in each of the genes, as well as the plasmid, generating predicted sizes (Table 4).

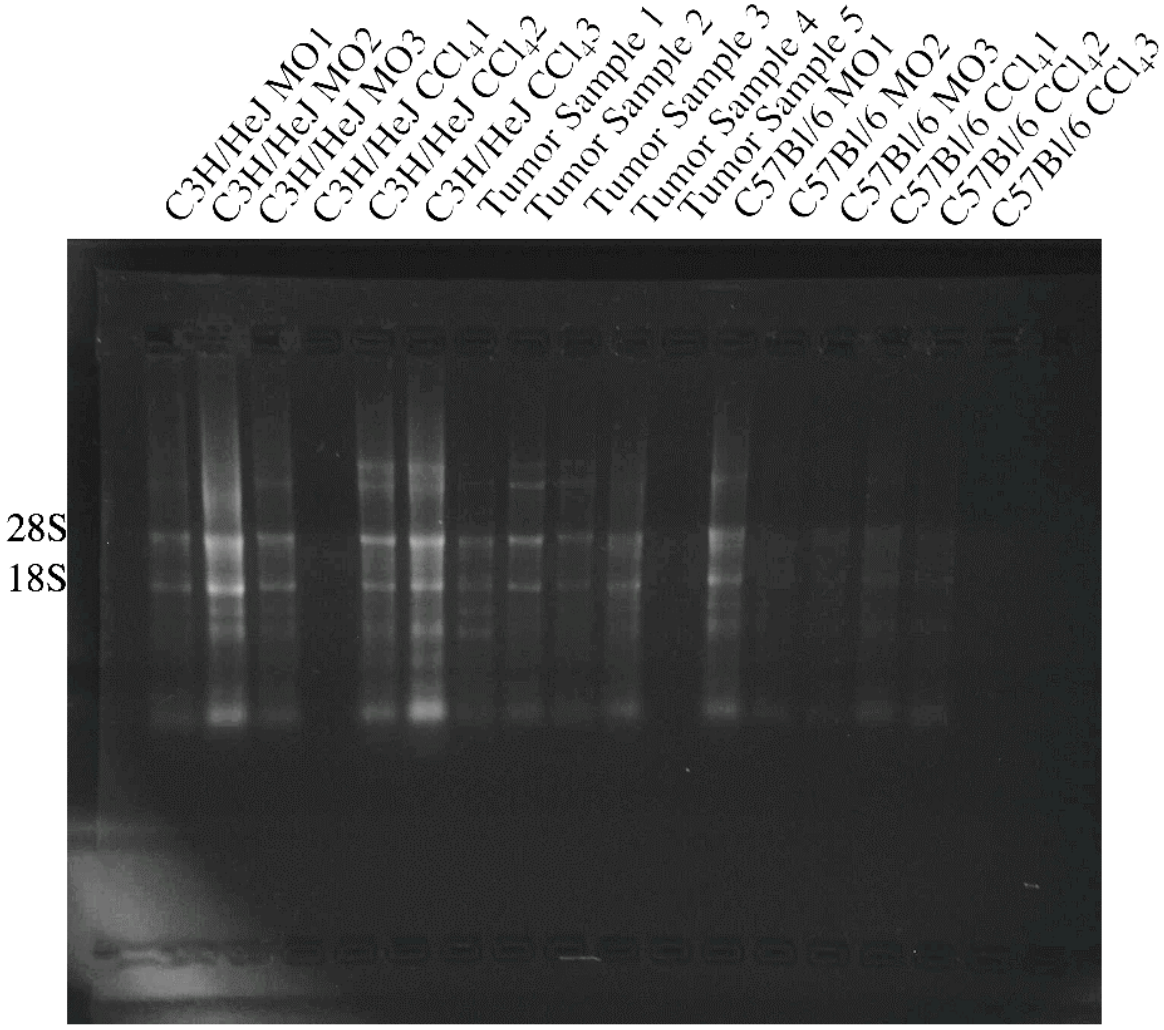

Figure 12. Formaldehyde Gel used to check Integrity of Liver RNA Samples

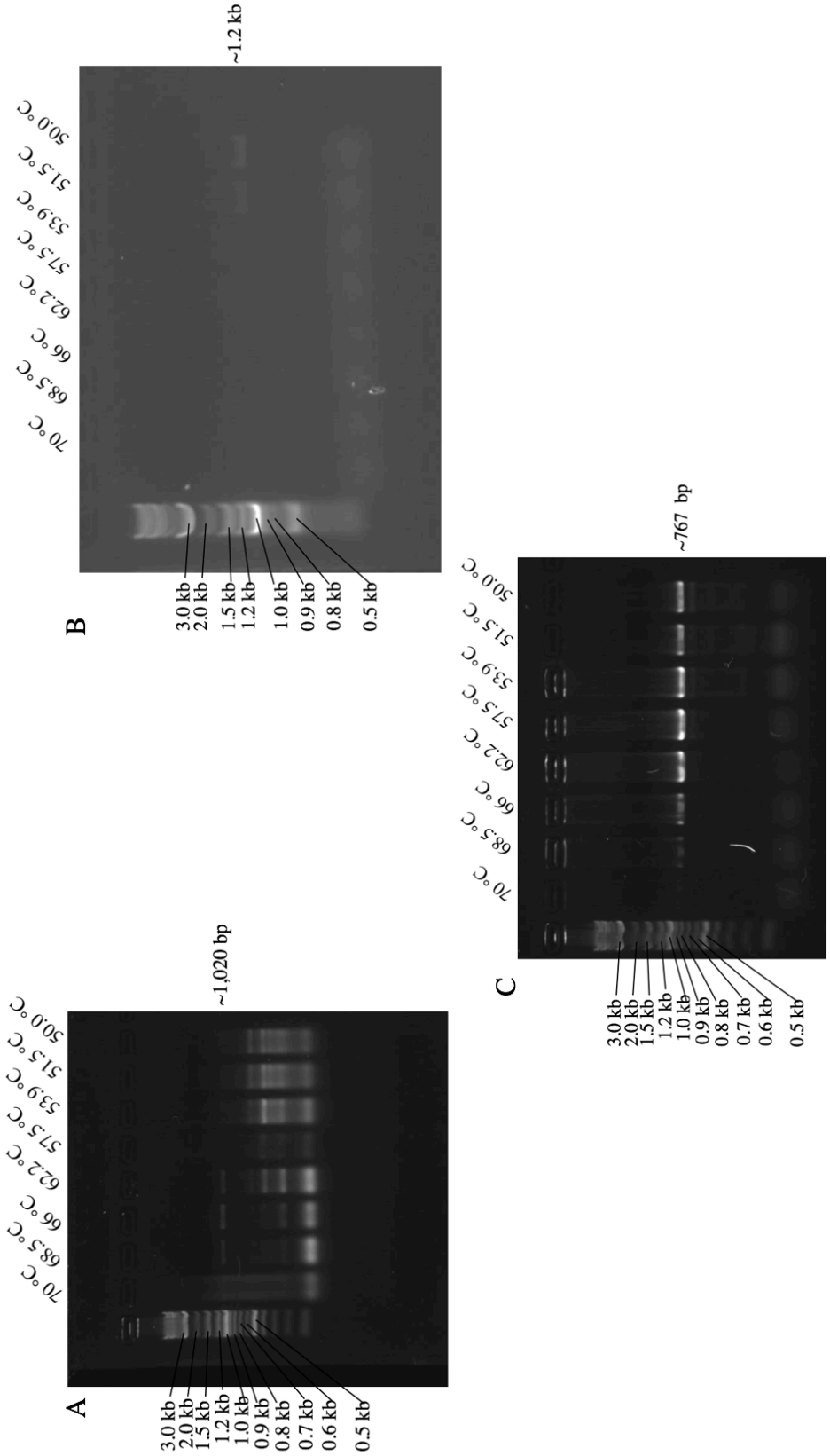

Figure 13. Optimal Annealing Temperatures for Primer Sets. Gradient PCR results for **A)** *Ciao1*, **B)** *WT1*, and **C)** *Ywhae*. Gradient temperatures corresponding to each lane are shown across the top. Ladder sizes are shown down the left side of the image. Fragment sizes are on the right side. Exact sizes are given where the expected sizes correspond to the present band (*Ciao1* and *Ywhae*), approximate sizes are noted where bands don't correspond to the expected size (*WT1*).

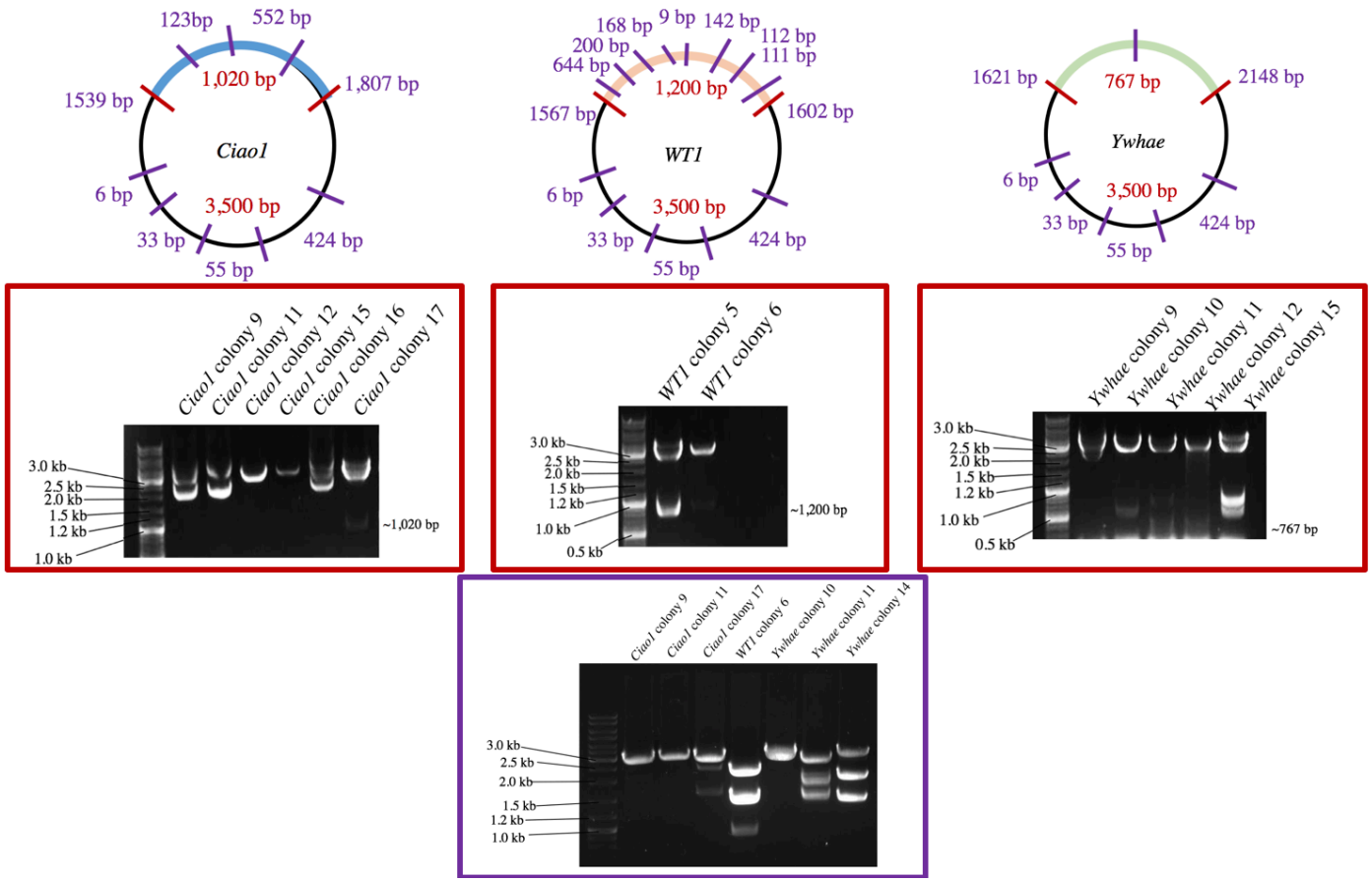

Figure 14. Cloning and Restriction Enzyme Analysis of Candidate Genes. Red lines and numbers within the vector maps indicate EcoRI sites and fragment sizes. Purple lines and numbers within the vector maps indicate AvaII cut sites and fragment sizes. Red bordered gel images are EcoRI digests while purple is the AvaII digest.

These results indicate that *WT1* colony W6 is likely correct and *Ywhae* colonies Y10, Y11, and Y14 are correct. *Ciao1* colony 17 has a fragment that is smaller than expected, but may be a positive colony. DNA sequencing will be used to confirm all clones. Having these genes cloned and the sequences will enable the next step in this research: functional studies using expressed proteins and an AFP reporter of transcription.

IV. DISCUSSION

Understanding key initiators, regulators, and markers of liver tumorigenesis are important for early diagnoses and treatment, and therefore, the reduction of mortality. Expression of RNA from the *AFP* gene in the liver, which is normally silent in the adult, has been used as a biomarker for liver tumorigenesis (Abelev 1971). The aim of this project was to identify regulators that lead to adult expression of AFP which could provide earlier biomarkers or genetic therapeutic targets for liver tumorigenesis.

In this study, a combination of bioinformatic analyses, computational modeling, mining of existing mapping data, genetic variation, and gene expression data was used to identify and clone three potential regulators in the mouse model. *Ciao1*, a known transcription factor whose genomic location is within the region of interest on chromosome 2 in mouse, was identified from genetic mapping, gene expression, and genetic variation analyses. *Ciao1* is a multi-WD40 subunit protein that has transcription factor ontology and sequence-specific DNA binding activity (Johnstone et al 1999). Variation analyses indicated only synonymous differences between the C3H/HeJ version of *Ciao1* and the C57BL/6. However, the promoter of *Ciao1* has differences which might lead to different levels of expression of the protein in C3H/HeJ and C57BL/6. These differences could account for the AFP phenotype differences in C3H/HeJ and C57BL/6 mouse strains. The effect of these mutations on *Ciao1* transcription, as well as the effect of the different Ciao1 protein levels on *AFP* transcription should be investigated further. Holding to the model of *AFP* transcription postulated here, increased or decreased levels of Ciao1 would be expected to increase or decrease level of AFP, respectively.

The second identified candidate, *WT1*, whose genomic location is outside of the region of interest on chromosome 2 in mouse, was implicated as a potential interaction partner with *Ciao1* through *AFP* promoter analysis and previous studies that indicate interaction and regulation by *Ciao1* (Johnstone et al 1999). The nucleotide sequence shows very little variation between strains, and the domain analysis shows no change in the predicted domains of the proteins across strains. Interestingly, *WT1* has also been shown to interact with *Zhx2* (*Afr1*) in kidney development and podocyte disease (Liu et al 2006). *WT1* expression data also shows that *WT1* expression levels mimic that of *AFP* – increased in the embryonic development stages and silenced in the adult liver (Mouse ENCODE Consortium et al 2012). *WT1* has been shown to be essential for the successful development and differentiation of multiple tissue types arising from the mesoderm through the binding targets' 3' untranslated regions (UTR) and subsequent mRNA downregulation through an decrease in mRNA stability (Bharathavikru et al 2017). It has also been shown to act as both a transcriptional activator and repressor (Essafi et al 2011, Toska and Roberts 2014), and has a binding affinity for both DNA and RNA (Bardeesy and Pelletier 1998). This observation could indicate that though initially identified and analyzed as a transcription factor of *AFP*, *WT1* could actually be influencing regulation through RNA stability as opposed to mRNA production.

The third potential regulatory gene identified is *Ywhae*. In the mouse, it is located on chromosome 11 and shows ontology related to protein domain specific binding. It was identified through the use of the MENTHA Interactome browser. *Ywhae* was predicted to be a potential intermediate between *Ciao1* and *WT1*. The known expression profiles match that of AFP in the liver in that expression decreases as embryonic development

occurs and the lowest level occurs in the adult liver (Mouse ENCODE Consortium et al 2012). There are no published data on *Ywhae* levels in liver tumorigenesis or regeneration.

Based upon these data, the proposed model of AFP reactivation is one in which a complex formed by Ciao1:Ywhae:WT1 is required for reactivation of *AFP* transcription (Figure 15A). This can be tested both *in vitro* and *in vivo* using purified proteins and transfected cells, respectively. The idea of a complex, rather than a single regulatory protein being responsible for this reactivation could potentially explain why the identity of the *Afr2* gene has remained elusive for so many years. That is, traditional methods such as 1-hybrid or 2-hybrid genetic screens would have been unable to identify the three-protein regulatory complex as it only utilizes a single, or two genes at a time.

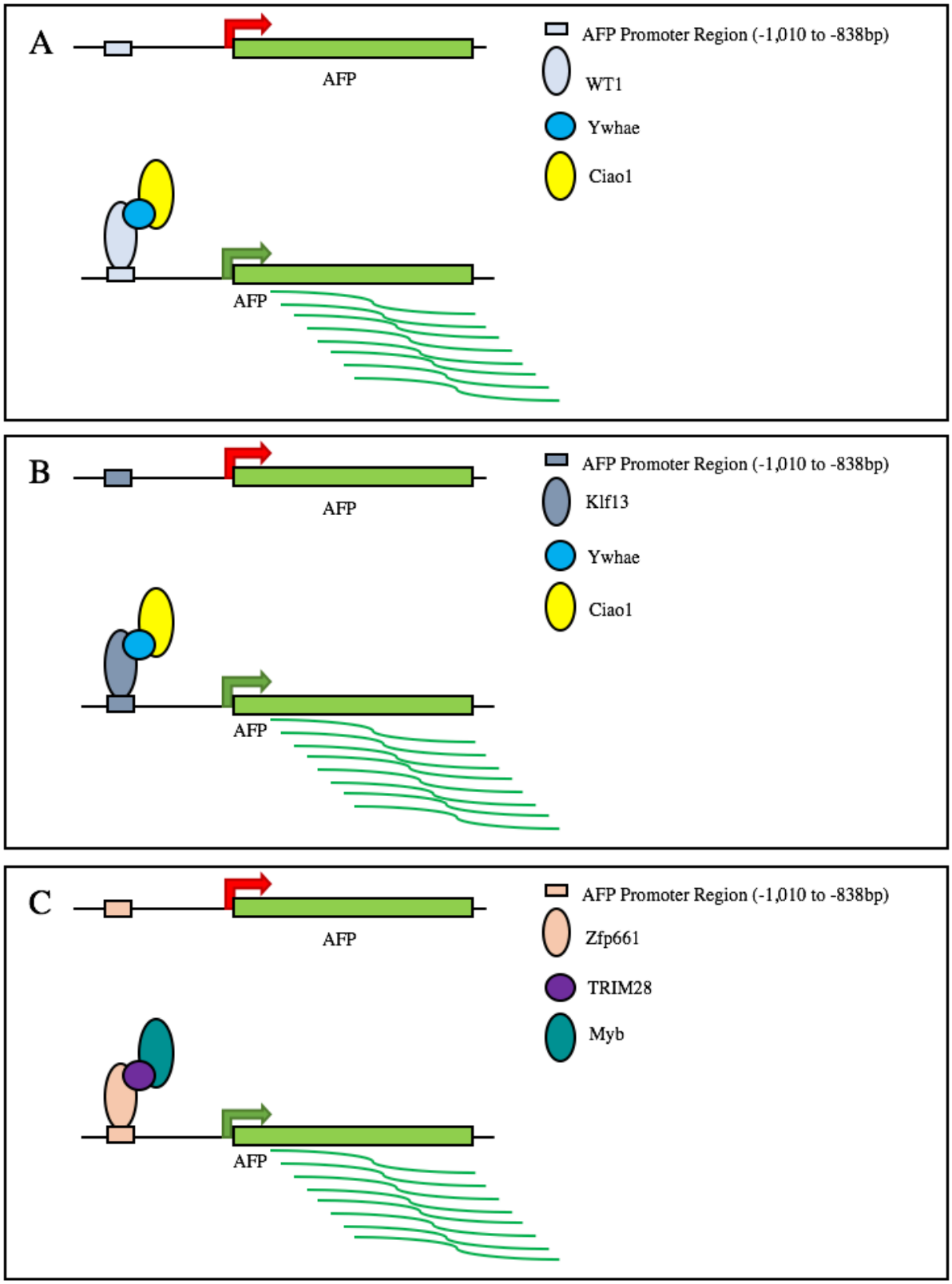

Figure 15. Proposed Models of *AFP* Transcriptional Reactivation

While the *Ciao1:Ywhae:WT1* pathway was identified as the candidate with the highest potential, there are two alternative pathways that show potential as well: *Ciao1:Ywhae:Klf13*, and *Zfp661:TRIM28:Myb* (Figure 15B-C).

Klf13 is a transcription factor, found on mouse chromosome 7, that regulates the production and development of many immune cells (Gordon et al 2008), though studies have indicated a reproductive phenotype associated with total deletion (Heard et al 2012). In a study using bioinformatics screening, *Klf13* was identified as a potential gene linked to gastric cancers (Li et al 2017), and Henson and Gollin showed that the overexpression of this gene was present in oral cancer cells (2010). While this candidate pathway has some potential, and *Klf13* shows transcription factor ontology, it also shows ontology related to repression of cellular proliferation (Gordon et al 2008). Mechanistically, this pathway could function through overexpression of this target, though no studies have been done on *Klf13* in liver cells.

The second alternative pathway consists of Zfp661, TRIM28, and Myb. *Zfp661* encodes a known zinc-finger containing transcription factor which shows high expression in embryonic livers, and a decreased expression in adult livers (Mouse ENCODE Consortium et al 2012) and is also within the candidate region on chromosome 2, indicating potential as a candidate for *Afr2*. There is a lack of liver studies using *Zfp661*, so no experimental conclusions about liver-related processes can be drawn. However, it has been shown to function in erythropoiesis (Papadopolous et al 2015) and to be active in development (Carter et al 2008).

TRIM28 is the intermediate in the second alternate pathway. TRIM28 is recruited to transcriptional start sites by zinc-finger proteins (Shibata et al 2011) and shows transcription factor ontology as well as transcription coactivator activity and contributes to sequence-specific DNA binding. This could indicate that it induces conformational changes within the binding partner leading to the recruitment of other transcriptional cofactors (Venkov et al 2007). Though ontology data suggests activator activity, TRIM28 has also been shown to interact with WTX, a tumor suppressor. This interaction is though to lead to epigenetic silencing and activity as a regulator in both cellular differentiation and tumorigenesis (Kim et al 2015). Other studies have shown that *TRIM28* regulates cancer stem cell populations in some cancer types (Czerwinska et al 2016). In this model, TRIM28 might bind to Zfp661 to regulate its binding or activation characteristics, or to act simply as a bridge to MYB. Other regulatory proteins are known to have opposing compelling activities in different molecular contexts.

Myb shows many transcription factor-related ontology terms, and mainly functions as a cell cycle checkpoint (Afroze et al 2003), a phenotype that if mutated could lead to tumor production. Studies suggest that this could also have some tumor suppressing function due to the deletion in various tissue types causing an increase in tumorigenesis (George and Ness 2014, Thorner et al 2010). It has also been implicated in the development of pancreatic cancers due to its presence or absence in the quiescent and cancerous organ, respectively (Srivastava et al 2015). This expression pattern mimicked in the liver, suggesting a potential role for *MYB* in the regulation of liver cancer, but not in liver regeneration, as would be expected of *Afr2*.

While the best candidate model, *Ciao1:Ywhae:WT1*, has support from several sources, it is possible that the connections between the members of the complex are coincidental since several pieces of data are based upon modeling. It will be important to test this model *in vivo* by expressing all three of the candidate proteins together in the same cell with an AFP promoter-reporter. To this aim, 2 of the 3 candidate genes have been successfully cloned from RNA isolated from regenerating liver from a C3H/HeJ mouse. If the model were to hold true, however, phenotypic studies in which one or more of the candidates in the complex are either deleted by CRISPR/Cas9 or knocked down by RNAi should be carried out in C3H/HeJ mice, or the introduction of the complex into C57BL/6 transgenic mice. Further, expression and/or variation should be investigated in humans as earlier biomarkers for liver cancer.

In summary, this study provides a candidate model of AFP reactivation involving a complex of three proteins, *Ciao1*, *WT1*, and *Ywhae*, as well as two alternative candidate models that are also complexes. These models, once confirmed, could aid in understanding and early diagnosis of liver cancer, which is the 5th and 9th leading cause of death in men and women, respectively.

LITERATURE CITED

[CDC] Centers for Disease Control and Prevention. Liver Cancer [Internet]. 2015 [cited 2015 October 15]. Available from http://www.cdc.gov/cancer/liver/index.htm

[NCI] National Cancer Institute A Snapshot of Liver Cancers [Internet]. 2014 [cited 15 October 2015]. Available from http://www.cancer.gov/research/progress/snapshots/liverandbileduct

[NIH] National Institutes of Health, National Cancer Institute. What you need to know about Liver Cancer [Internet]. Maryland: 2009 [cited 13 October 2015]. Available from http://www.cancer.gov/publications/patient-education/liver.pdf. Also available in paper copy from the publisher.

Abelev GI. Alpha-fetoprotein in ontogenesis and its association with malignant tumors. *Adv Cancer Res* 1971; 14:295-358.

Afroze T, Yang LL, Wang C, Gros R, Kalair W, Hoque AN, Mungrue IN, Zhu Z, Husain M. Calcineurin-independent regulation of plasma membrane Ca2+ ATPase-4 in the vascular smooth muscle cell cycle. *Am J Physiol Cell Physiol* 2003; 285(1):C88-95.

Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschi T. A uniform system for microRNA annotation. *RNA* 2003; 9(3):277-9.

Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gastreiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G,

Xenarios I, and Stockinger H. ExPASy: SIB bioinformatics resource portal. *Nucleic Acid Res* 2012; 40(W1):W597-603.

Bardeesy N, Pelletier J. Overlapping RNA and DNA binding domains of the wt1 tumor suppressot gene product. *Nuc Acid Res* 1998; 26(7): 1784-92.

Bharathavikru R, Dudnakova T, Aitken S, Slight J, Artibani M, Hohenstein P, Tollervey D, Hastie N. Transcription factor Wilms' tumor 1 regulates developmental RNAs through 3' UTR interaction. *Genes Dev* 2017; 31(4):347-52.

Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ, and the Mouse Genome Database Group. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nuc Acids Res* 2017; 45(D1): D723-9.

Calderone A, Castagnoli L, Cesareni G. Mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods* 2013; 10:690-1.

Carter MG, Stagg CA, Falco G, Yoshikawa T, Bassey UC, Aiba K, Sharova LV, Shaik N, Ko MSH. An *in situ* hybridization-based screen for heterogeneously expressed genes in mouse ES cells. *Gene Expr Patterns* 2008; 8(3):181-98.

Chow W, Brugger K, Caccamo M, Sealy I, Torrance J, Howe K. gEVAL – A web based browser for evaluating genome assemblies. *Bioinformatics* 2016.

Czerwinska P, Shah PK, Tomczak K, Klimczak M, Mazurek S, Sozanska B, Biecek P, Korski K, Filas V, Mackiewicz A, Andersen JN, Wiznerowicz M. TRIM28 multi-domain protein regulates cancer stem cell population in breast tumor development. *Oncotarget* 2016; 8:863-82.

Essafi A, Webb S, Berry RL, Slight J, Burn SF, Spraggon L, Velecela V, Martinez-Estrada OM, Wiltshire JH, Roberts SG, Brownstein D, Davies, JA, Hastie ND, Hohenstein P. A wt1-controlled chromatin switching mechanism underpins tissue-specific wnt4 activation and repression. *Dev Cell* 2011; 21(3):559-74.

George OL, Ness SA. Situational Awareness: Regulation of the Myb Transcription Factor in Differentiation, the Cell Cycle and Oncogenesis. *Cancers (Basel)* 2014; 6(4):2049-71.

Gordon AR, Outram SV, Karamatipour M, Goddard CA, Colledge WH, Metcalfe JC, Hager-Theodorides AL, Compton T, Kemp PR. Splenomegaly and modified erythropoiesis in KLF13 -/- mice. *J Biol Chem*. 2008; 283(18):11897-904.

Heard ME, Pabona JM, Clayberger C, Krensky AM, Simmen FA, Simmen RC. The Reproductive Phenotype of Mice Null for Transcription factor Kruppel-Like Factor 13 Suggests Compensatory Function of Family Member Kruppel-Like Factor 9 in the Peri-Implantation Uterus. *Biol Reprod* 2012; 87(5):115-26.

Henson BJ, Gollin SM. Overexpression of Klf13 and FGFR3 in oral cancer cells. *Cytogenet Genome Res* 2010; 128(4):192-8.

Jin DK, Feuerman MH. Genetic mapping of *Afr2 (Rif)*: regulator of gene expression in liver regeneration. *Mammalian Genome* 1997; 9:256-8.

Jin DK, Vacher J, Feuerman MH. α-fetoprotein gene sequences mediating *Afr2* regulation during liver regeneration. *Proc Natl Acad Sci USA* 1998; 95:8767-72.

Johnstone RW, Tommerup N, Hansen C, Vissing H, Shi Y. Structural organization, tissue expression, and chromosomal localization of *Ciao1*, a functional modulator of the Wilms' tumor suppressor, *WT1*. *Immunogenetics* 1999; 49(5):900-5.

Jones P, Binns D, Chang H, Fraser M, Li W, MnAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn A, Sangrador-Vegas A, Scheremetjew M, Yon S, Lopez R, Hunter S. InterProScan 5: genome-scale protein function and classification. *Bioinformatics* 2014.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004; 2:D493-6.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellåker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhomra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Bala S, Stalker J, Mott R, Durbin R, Jackson IJ, Czechanski A, Guerra-Assunção JA, Donahue LR, Reinholdt LG, Payseur BA, Ponting CP, Birney E, Flint J and Adams DJ. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 2011; 477(7364):289-94

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006.

Kibbe WA. OligoCalc: an online oligonucleotide properties calculator. *Nuc Acid Res* 2007; 35:W43-6.

Kim WJ, Wittner BS, Amzallag A, Brannigan BW, Ting DT, Ramaswamy S, Maheswaran S, Haber DA. The WTX Tumor Suppressor Interacts with the Transcriptional Corepressor TRIM28. *J Biol Chem* 2015; 290(23):14381-90.

Kosugi S, Ohashi Y. DNA binding and dimerization specificity and potential targets for the TCP protein family. *Plant J* 2002; 30(3):337–348.

Lelli, KM, Slattery M, Mann RS. Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. *Ann Rev Genet* 2012; 46:43-68.

Li Y, Zhang L, Yang C, Li R, Shang L, Zou X. Bioinformatic identification of candidate genes induced by trichostatin A in BGC-823 gastric cancer cells. *Oncol Lett* 2016; 13(2):777-783.

Lin J, Long L, Green MA, Spear BT. The alpha-fetoprotein enhancer region activates the albumin and alpha-fetoprotein promoters during development. *Dev Biol* 2009; 336(2):294-300.

Liu G, Clement LC, Kanwar YS, Avila-Casado C, and Chugh S. ZHX Proteins Regulate Podocyte Gene Expression during the Development of Nephrotic Syndrome. *J Biol Chem* 2007; 281(51):39681-92.

Messeguer X, Escudero R, Farre D, Nunez O, Martinez J, Alba MM. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics* 2002; 18(2):333-4.

Michalopolous GK. Liver Regeneration. *J Cell Physiol* 2007; 213(2):286-300.

Morford LA, Davis C, Jin, L, Dobierzewska A, Peterson ML, Spear BT. The Oncofetal Gene *Glypican 3* is Regulated in the Postnatal Liver by Zinc Fingers and Homeoboxes 2 and in the Regenerating Liver by Alpha-Fetoprotein Regulator 2. Hepatology 2007; 46(5):1541-7.

Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kau R *et al*. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* 2012; 13(8):418.

Papadopoulos P, Gutierrez L, Demmers J, Scheer E, Pourfarzad F, Papageorgiou DN, Karkoulia E, Strouboulis J, van de Werken HJG, van der Linden R, Vandenberghe P, Dekkers DHW, Philipsen S, Grosveld F, Tora L. TAF10 interacts with the GATA1 Transcription Factor and Controls Mouse Erythropoiesis. *Mol Cell Biol* 2015; 35(12):2103-18.

Peterson ML, Chunhong MA, Spear BT. Zhx2 and Zbtb20: Novel regulators of postnatal alpha-fetoprotein repression and their potential role in gene reactivation during liver cancer. *Sem Cancer Biology* 2011; 21(1): 21-7.

Schultz J, Milpetz F, Peer B, Ponting C. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc Natl Acad Sc USA* 1998; 95:5857-64.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Sys Bio* 2011; 7:539.

Shibata M, Blauvelt KE, Liem Jr KF, Garcia-Garcia MJ. TRIM28 is required by the mouse KRAB domain protein ZFP568 to control convergent extension and morphogenesis of extra embryonic tissues. *Development* 2011; 138(24): 5333-43.

Shin D, Monga SPS. Cellular and Molecular Basis of Liver Development. *Compr Physiol* 2013; 3(2):799-815.

Spear BT, Jiu L, Ramasamy S, Dobierzewska A. Transcriptional control in the mammalian liver: liver development, perinatal repression, and zonal gene regulation. *Cell Mol Life Sci* 2006; 63:2922-38.

Spear BT. Alpha-fetoprotein gene regulation: lessons from transgenic mice. *Sem Canc Biol* 1999; 9:109-16.

Srivestava SK, Bhardwaj A, Arora S, Singh S, Azim A, Tyagi N, Carter JE, Wang B, Singh AP. MYB is a novel regulator of pancreatic tumour growth and metastasis. *Br J Cancer* 2015; 113(12):1694-1703.

Thorner AR, Parker JS, Hoadley KA, Perou CM. Potential Tumor Suppressor Role for the c-Myb Oncogene in Luminal Breast Cancer. *PLoS One* 2010; 5(10):e13073.

Tomasi Jr TB. Structure and Function of Alpha-fetoprotein. *Ann Rev Med* 1977; 28:453-65.

Toska A, Roberts SG. Mechanisms of transcriptional regulation by WT1 (Wilms' tumour 1). *Biochem J* 2014; 461(1):15-32.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. Primer3Plus, an enhanced web interface to Primer3. *Nuc Acid Res* 2007; 35: W71-4.

Venkov CD, Link AJ, Jennings JL, Plieth D, Inoue T, Nagai K, Xu C, Dimitrova YN, Rauscher FJ, Neilson EG. A proximal activator in transcription in epithelial-mesenchymal transition. *J Clin Invest* 2007; 7(2):482-91.

Yalcin , Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellaker C, Goodstadt L, Nicod J, Bhomra A, Hernandez-Pliego P, Whitley H, leak J, Dutton R, Janowitz D, Mott R, Adams DJ, Flint J. Sequence-based characterization of structural variation in the mouse genome. *Nature* 2011; 477:326-9.

Zaret KS, Grompe M. Generation and Regeneration of Cells of the Liver and Pancreas. *Science* 2008; 322(5907):1490-4.

APPENDICES

APPENDIX A. PROMO Analysis of the *AFP* Promoter in C3H/HeJ

0 Ncx [T04368]
1 HNF-1A [T00368]
2 Cdx-1 [T01484]
3 HOX11 [T02054]
4 FOXO3a [T02938]
5 Pax-5 [T00070]
6 MYB2 [T02536]
7 ZF5 [T02349]
8 Msx-1 [T02072]
9 ABF1 [T00056]
10 HOXA3 [T00378]
11 Pax-2a [T00678]
12 USF-1 [T00875]
13 Zic1 [T04669]
14 Zic2 [T04670]
15 R [T00710]
16 AIRE [T05990]
17 HBP-1a(1) [T01394]
18 HBP-1a(c14) [T01395]
19 TGA1a [T00829]
20 C/EBPbeta(p20) [T01420]
21 C/EBPalpha [T00107]
22 HNF-3beta [T02344]
23 MEF1 [T00506]
24 muEBP-C2 [T00215]
25 USF-1 [T00877]
26 c-Myc [T00143]
27 LCR-F1 [T01599]
28 STE12 [T00772]
29 COE1 [T01112]
30 LyF-1 [T00479]
31 Elk-1 [T00250]
32 Pax-6 [T00682]
33 Pax-8 [T01828]
34 Zeste [T00918]
35 Zeste [T02100]
36 MF3 [T00507]
37 NF-1 [T01298]
38 AP-2alphaA [T00035]
39 Vpr [T02399]
40 LVb-binding factor [T00477]
41 TMF [T00835]
42 DBP [T00183]
43 AP-1 [T00029]
44 USF2 [T02115]
45 c-Fos [T00123]
46 Pax-2 [T01823]
47 PR B [T00696]
48 PR A [T01661]
49 GR-alpha [T00337]
50 PR B [T00697]
51 GR-beta [T01920]
52 USF [T00873]
53 YY1 [T00278]
54 C/EBPalpha [T00104]
55 CUTL1 [T00100]
56 MAC1 [T01265]
57 C/EBPalpha [T00105]
58 C/EBPalpha [T00108]
59 C/EBPbeta [T00459]
60 C/EBPbeta [T00581]
61 p300 [T01427]
62 FACB [T02841]
63 USF2b [T02377]
64 USF2 [T00878]
65 c-Jun [T00131]
66 JunB [T00436]
67 c-Fos [T00124]
68 c-Jun [T00132]
69 c-Jun [T00133]
70 JunD [T00437]
71 NHP-1 [T00621]
72 c-Fos [T00122]
73 ATF3 [T01095]
74 HMG I(Y) [T02368]
75 STAT4 [T01577]
76 NF-kappaB [T00588]
77 RelA [T00594]
78 COE2 [T05006]
79 c-Ets-1 [T00112]
80 MYBAS1 [T05553]
81 R2 [T00712]
82 DEAF-1 [T05885]
83 STAT6 [T01581]
84 UME6 [T01247]
85 STAT1 [T01575]
86 DEF:GLO [T03216]
87 HNF-3 [T02277]
88 Dl [T00196]
89 BR-C Z3 [T01479]
90 AP3:PI [T03251]
91 AP-3 (2) [T00039]
92 DSXF [T00955]
93 DSXM [T00956]
94 Nrf2:MafK [T05666]
95 DEF:GLO:SQUA [T03217]
96 IRF-2 [T00425]
97 Eve [T00272]
98 GR [T00333]
99 HNF-3beta [T02513]
100 HNF-3alpha [T00371]
101 RP58 [T05040]
102 BTEB3 [T05051]
103 TRM1 [T05311]
104 POU2F2 (Oct-2.1) [T00646]
105 TCF-1A [T00999]
106 LEF-1 [T02905]
107 TCF-4E [T02878]
108 HNF-4alpha2 [T02422]
109 HNF-4alpha1 [T02429]
110 LEF-1 [T00930]
111 GT-1 [T00339]
112 POU3F1 [T00969]
113 Cutl1 [T02042]
114 AGL3 [T03025]
115 POU1F1a [T00691]
116 C/EBPdelta [T00109]
117 unc-86 [T01882]
118 YY1 [T00915]
119 Pax-4a [T02983]
120 BR-C Z2 [T01478]
121 POU3F2 [T00630]
122 FOXP3 [T04280]
123 YY1 [T00865]
124 POU2F2C [T00665]
125 POU2F2 [T01032]
126 FOXD3 [T02290]
127 MATalpha2 [T00487]
128 Elf-1 [T01019]
129 HNF-3beta [T01049]
130 c-Myb [T00137]
131 GA-BF [T00297]
132 Olf-1 [T01040]
133 MZF-1 [T00529]
134 STAT5A [T04683]
135 STAT3 [T01493]
136 c-Ets-1 [T00111]
137 c-Ets-1 68 [T00115]
138 HELIOS [T06012]
139 C/EBPbeta [T00017]
140 CAT8 [T03227]
141 NF-1 [T00537]
142 ENKTF-1 [T00255]
143 NBF [T00951]
144 NBF [T01190]
145 INO2 [T01241]
146 AR [T00040]
147 C/EBP [T01386]
148 E47 [T00207]
149 CP2 [T00152]
150 WT1 I [T01840]
151 YY1 [T04970]
152 ARF1 [T04514]
153 NF-1 [T00538]
154 EIIaE-A [T00246]
155 TGGCA-binding protein [T00832]
156 LIM1 [T04817]
157 Nkx2-1 [T00856]
158 NFI/CTF [T00094]
159 DEC2 [T05845]
160 Mitf [T01554]
161 RC2 [T00724]
162 Zic3 [T04671]
163 Alfin1 [T04733]
164 CAC-binding protein [T00076]
165 ABI4 [T05743]
166 Pax-9a [T03593]
167 Pax-9b [T03594]

APPENDIX B. PROMO Analysis of the *AFP* Promoter in C57BL/6

0 MF3 [T00507]
8 HMG I(Y) [T02368]
16 XPF-1 [T00906]
24 TGGCA-binding protein [T00832]
32 BR-C Z2 [T01478]
40 NHP-1 [T00621]
48 STE12 [T00772]
56 CREMtau1 [T02108]
64 POU5F1 [T00651]
72 POU2F2 (Oct-2.1) [T01864]
80 Pax-2a [T00678]
88 GR [T00333]
96 DEF:GLO:SQUA [T03217]
104 Eve [T00272]
112 Nkx6-2 [T02050]
120 HOXD9 [T01756]
128 c-Myb [T00137]
136 YY1 [T04970]
144 MAC1 [T01265]
152 HNF-1A [T01211]
160 MED8 [T03491]
168 LF-A1 [T00467]
176 POU2F1b [T01862]
184 GATA-1 [T05705]
192 AP-2 [T00034]
200 YY1 [T00865]

1 MYB2 [T02536]
9 STAT4 [T01577]
17 EBF [T05427]
25 LIM1 [T04817]
33 NF-Y [T00613]
41 Nrf2:MafK [T05666]
49 HNF-3beta [T02344]
57 CREMtau2 [T02109]
65 POU2F2C [T00665]
73 POU2F2 (Oct-2.3) [T01865]
81 AP-2alphaA [T00035]
89 CDX2 [T03246]
97 TCF-3 [T02857]
105 HNF-3beta [T01049]
113 HNF-3alpha [T02512]
121 HOXD10 [T01757]
129 DSXF [T00955]
137 CP2 [T00152]
145 C/EBPalpha [T00105]
153 FOXD3 [T04166]
161 Nkx2-2 [T02384]
169 WT1 I [T01840]
177 POU2F1c [T01863]
185 GATA-1 [T00306]
193 TCF-1(P) [T01109]

2 Zic3 [T04671]
10 p300 [T01427]
18 LYS14 [T03472]
26 C/EBPalpha [T00107]
34 NFI/CTF [T00094]
42 STAT5A [T04683]
50 IRF-3 [T04673]
58 NERF-1a [T05021]
66 POU2F2 [T01032]
74 POU2F2 (Oct-2.4) [T01866]
82 PR B [T00696]
90 HOXD8 [T01426]
98 AGL3 [T03025]
106 GA-BF [T00297]
114 FOXI1 [T02474]
122 HOXD10 [T01758]
130 DSXM [T00956]
138 Zic1 [T04669]
146 C/EBPalpha [T00108]
154 GT-1b [T04286]
162 Pu box binding factor [T00704]
170 Oct-B1 [T00545]
178 NF-AT2 [T01945]
186 CRF [T00170]
194 Hlf [T01071]

3 YY1 [T00278]
11 FACB [T02841]
19 MZF-1 [T00529]
27 NF-Y [T00150]
35 GATA-2 [T01302]
43 IRF-1 [T00423]
51 HNF-3 [T02277]
59 myogenin [T00528]
67 POU2F1 [T00641]
75 POU2F2 (Oct-2.6) [T01867]
83 PR A [T01661]
91 HOXD8 [T01754]
99 STAT5B [T04761]
107 LCR-F1 [T01599]
115 FOXP3 [T04280]
123 HFH-1 [T02288]
131 C/EBPbeta [T00017]
139 Zic2 [T04670]
147 C/EBPbeta [T00459]
155 POU3F1 [T00969]
163 Elf-1 [T01019]
171 POU2F2B [T00662]
179 NF-AT1 [T01948]
187 CUTL1 [T00100]
195 Myf-3 [T00519]

4 Ncx [T04368]
12 c-Ets-2 [T01397]
20 COE1 [T01112]
28 Cutl1 [T02042]
36 GATA-2 [T00308]
44 PEA3 [T00684]
52 HNF-3alpha [T00371]
60 MyoD [T00526]
68 POU2F2B [T00964]
76 POU2F2 (Oct-2.1) [T01870]
84 MATalpha2 [T00487]
92 MNB1a [T01059]
100 PEND [T05977]
108 POU2F2 (Oct-2.1) [T00646]
116 FOXI1a [T02294]
124 MBF1 [T00492]
132 FOXO3a [T02938]
140 Smad3 [T04096]
148 C/EBPbeta [T00581]
156 NF-AT4 [T01949]
164 AIRE [T05990]
172 POU2F1 [T00959]
180 TRM1 [T05311]
188 mec-3 [T01076]
196 MyoD [T00525]

5 Cdx-1 [T01484]
13 c-Ets-1 [T00112]
21 Elk-1 [T00250]
29 POU1F1a [T00691]
37 Pax-6 [T00682]
45 LVb-binding factor [T00477]
53 HOXA3 [T00378]
61 Msx-1 [T02072]
69 POU2F2 [T00647]
77 Zta [T00923]
85 GR-alpha [T00337]
93 POU1F1b [T01516]
101 Pax-4a [T02983]
109 DEF:GLO [T03216]
117 HOXD9 [T01424]
125 Gt [T00328]
133 TGT3 [T00833]
141 Smad4 [T04292]
149 AP-3 (2) [T00039]
157 FOXJ2 (long isoform) [T04169]
165 AR [T00040]
173 POU2F2B [T01871]
181 NF-AT1 [T01944]
189 MYBAS1 [T05553]
197 MyoD [T01128]

6 Pax-8 [T01828]
14 NF-1 [T01298]
22 NF-1 [T00536]
30 C/EBPdelta [T00109]
38 RC2 [T00724]
46 HELIOS [T06012]
54 NF-AT1 [T00550]
62 POU2F1 [T01466]
70 POU2F2 [T00648]
78 DBP [T00183]
86 PR B [T00697]
94 POU1F1c [T01902]
102 POU3F2 [T00630]
110 Hb [T00395]
118 HOXD10 [T01425]
126 BR-C Z4 [T01480]
134 HNF-3beta [T03256]
142 C/EBPalpha [T00104]
150 DEC2 [T05845]
158 FOXO4 [T04176]
166 RP58 [T05040]
174 POU5F1 (Oct-5) [T00653]
182 STAT1beta [T01573]
190 EIIaE-A [T00246]
198 YY1 [T00915]

7 PKNOX1 [T04122]
15 R2 [T00712]
23 NF-1 [T00538]
31 unc-86 [T01882]
39 NF-X3 [T01514]
47 c-Ets-2 [T00113]
55 CREMtau [T01309]
63 C/EBP [T01386]
71 octamer-binding factor [T01225]
79 RF2a [T02811]
87 GR-beta [T01920]
95 HNF-3beta [T02513]
103 FOXD3 [T02290]
111 Croc [T02291]
119 HOXD9 [T01755]
127 BR-C Z3 [T01479]
135 E47 [T00207]
143 HNF-3gamma [T01050]
151 TBP [T00798]
159 GT-1 [T01089]
167 ALF1B [T01496]
175 POU3F1 [T00656]
183 Nkx2-1 [T00857]
191 AP-2beta [T02469]
199 INSAF [T00406]