2019

EVOLUTION OF THE ZHX TRANSCRIPTION FACTOR FAMILY AND ANALYSIS OF ZHX2 TARGET GENES CYP2A4 AND CYP2A5 IN MOUSE LIVER

Alexandra Nichole Nail
University of Kentucky, alexandranail12@gmail.com
Author ORCID Identifier: https://orcid.org/0000-0002-8008-3891
Digital Object Identifier: https://doi.org/10.13023/etd.2019.204

Click here to let us know how access to this document benefits you.

Recommended Citation
Nail, Alexandra Nichole, 'EVOLUTION OF THE ZHX TRANSCRIPTION FACTOR FAMILY AND ANALYSIS OF ZHX2 TARGET GENES CYP2A4 AND CYP2A5 IN MOUSE LIVER' (2019). Theses and Dissertations--Microbiology, Immunology, and Molecular Genetics. 20.
https://uknowledge.uky.edu/microbio_etds/20

This Doctoral Dissertation is brought to you for free and open access by the Microbiology, Immunology, and Molecular Genetics at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Microbiology, Immunology, and Molecular Genetics by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.
STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student’s advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student’s thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Alexandra Nichole Nail, Student
Dr. Brett T. Spear, Major Professor
Dr. Ken Fields, Director of Graduate Studies
EVOLUTION OF THE ZHX TRANSCRIPTION FACTOR FAMILY AND ANALYSIS OF ZHX2 TARGET GENES CYP2A4 AND CYP2A5 IN MOUSE LIVER

DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Medicine at the University of Kentucky

By
Alexandra Nichole Nail
Lexington, Kentucky

Co- Directors: Dr. Brett T. Spear Professor of Microbiology, Immunology, and Molecular Genetics
and Dr. Martha L. Peterson Professor of Microbiology, Immunology, and Molecular Genetics

Lexington, Kentucky
2019

Copyright © Alexandra Nichole Nail 2019
ABSTRACT OF DISSERTATION

EVOLUTION OF THE ZHX TRANSCRIPTION FACTOR FAMILY AND ANALYSIS OF ZHX2 TARGET GENES CYP2A4 AND CYP2A5 IN MOUSE LIVER

The liver is the largest internal organ and performs a wide variety of functions to maintain organismal homeostasis. While some liver functions are carried out by all hepatocytes, other functions are restricted to certain populations of hepatocytes within the liver. This phenomenon, called zonal gene regulation or liver zonation, controls many metabolic processes within the liver including ammonia detoxification; glucose homeostasis; bile acid and glutamine synthesis; and metabolism of xenobiotics, lipids, and amino acids. The liver also expresses many genes in a developmental or sex-biased manner. Some genes are expressed at higher levels early or late in development, or alternatively, in male or female liver.

Several years ago, our lab identified a transcription factor called Zinc finger and homeoboxes 2 (Zhx2) based on its ability to control the silencing of genes that are normally expressed in the fetal liver. Zhx2 belongs to a small gene family that also includes Zhx1 and Zhx3. These four exon genes have a rather unique structure in that their entire protein coding region is located on an unusually large third exon. Preliminary studies indicate that these proteins are found only in vertebrates. I have performed a comprehensive analysis of Zhx proteins across a number of chordate species to determine their relationship throughout chordate evolution. Using multiple sequence alignment and phylogenetic tree-building, my studies have found that the primordial Zhx gene is most related to Zhx3 and that this gene exists in lower chordates including lancelet, sea squirt, and sea lamprey.

Additional studies from our lab showed that Zhx2 regulates numerous hepatic genes in the adult liver, including cytochrome p450 (Cyp) genes as well as other genes that exhibit sex-biased expression. Previous studies have demonstrated that female-biased expression of Cyp2a4, is controlled, in part, by Zhx2. I have extended these studies to perform a comprehensive analysis of Cyp2a4 and the highly related Cyp2a5 gene. Despite the high similarity of these two Cyp genes, my data indicate that these genes exhibit different zonal expression patterns and are differentially regulated in the regenerating liver. In the course of these studies, I discovered and characterized antisense transcripts for both Cyp2a4 and Cyp2a5. Both Cyp2a4as and Cyp2a5as have positively correlated expression patterns compared to their sense counterparts. In contrast to Cyp2a4 and Cyp2a5, Cyp2a4as and Cyp2a5as show sex-biased expression patterns earlier in development, suggesting that they might contribute to later sex-biased patterns established for Cyp2a4 and Cyp2a5.

KEYWORDS: Zinc fingers and homeoboxes, Cyp2a4, Cyp2a5, antisense
Alexandra Nichole Nail
(Name of Student)

03/22/2019
Date
EVOLUTION OF THE ZHX TRANSCRIPTION FACTOR FAMILY AND ANALYSIS OF ZHX2 TARGET GENES CYP2A4 AND CYP2A5 IN MOUSE LIVER

By
Alexandra Nichole Nail

Brett T. Spear
Co-Director of Dissertation

Martha L. Peterson
Co-Director of Dissertation

Ken Fields
Director of Graduate Studies

03/22/2019
Date
DEDICATION

To my husband, mother, and grandparents
ACKNOWLEDGEMENTS

“If I have seen further it is by standing on the shoulders of Giants,” Sir Isaac Newton. I would like to thank my mentors, Drs. Brett Spear and Martha Peterson for their continued support and guidance. Thank you for always believing in me. I am grateful for my committee members Drs. Melinda Wilson and Rebecca Dutch for their guidance and feedback throughout my graduate career. I also want to thank Dr. Jeramiah Smith for serving as my outside examiner, and for any bioinformatic guidance that he has given to me over the past five years.

I am especially grateful for Drs. Jieyun Jiang and Shirley Qui who have provided me with their expertise over the years, thank you for your patience and your help. A special thanks to Courtney Turpin and Louise Song; we will always have our cherished memories in the lab. Thank you for making me laugh and reminding me to take things a little less seriously. To Katie Bernardo, thank you for all of your help over the years with experiments, you’ll always be my favorite undergraduate! I would like to thank the Microbiology, Immunology, and Molecular Genetics department; thank you to Dr. Beth Garvy for your hard work behind the scenes to make my graduate experience a success.

I want to thank my friend, Marion Coe, for all the late nights and pep talks that we will never forget. Thank you to Chelsea Barrett, who I admire and respect. I have cherished my time serving in a leadership role with you. A special thanks to my brother and sister in-laws, thank you for your encouragement and love.

Last but not least, I’d like to thank my family for their support through the trials and tribulations of graduate school. Thank you to my mom, who has always been there and supported me through every endeavor. Thank you to my grandparents for your love and support. To my mom and dad in-law, thank you for loving me like one of your own, and for your comfort. Finally, thank you to my husband Matt. Thank you for the late nights, for your love, for everything that you have provided me with, I love you.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ............................................................................................................................... iii

LIST OF TABLES ........................................................................................................................................ vii

LIST OF FIGURES ....................................................................................................................................... viii

CHAPTER I .................................................................................................................................................. 1

INTRODUCTION ........................................................................................................................................... 1
  Liver Architecture .................................................................................................................................. 1
  Liver Development and Gene Expression in Developing Liver ......................................................... 2
  Zinc Fingers and Homeoboxes 2 (Zhx2) ................................................................................................. 4
  Zhx Gene Family .................................................................................................................................. 7
  *Cytochrome p450* Genes in Mouse and Human .................................................................................. 9
  Liver Zonation and Regulation of *Cyps* ................................................................................................. 10
  Sex-Biased Gene Regulation .................................................................................................................. 14
  Non-coding RNAs ................................................................................................................................. 18
  Antisense LncRNAs ............................................................................................................................... 19
  Hypotheses ............................................................................................................................................ 21

CHAPTER II ............................................................................................................................................... 32

MATERIALS AND METHODS ...................................................................................................................... 32
  Identification and Characterization of *Zinc Finger and Homeobox* Genes from Chordate Species ................................................................................................................................. 32
  Multiple Sequence Alignment and Phylogenetic Tree Building ................................................................. 33
  Sea Lamprey Tissue Collection ................................................................................................................ 34
  RNA Isolation for Mouse and Sea Lamprey Tissues ................................................................................ 34
  cDNA Synthesis and Reverse Transcription Quantitative Real-Time PCR (RT-qPCR) .................................................. 35
  Gel Electrophoresis and Gel Visualization ............................................................................................... 36
  DNA Isolation for Mouse and Sea Lamprey Tissues ............................................................................... 36
  Mice ...................................................................................................................................................... 37
  Breeding of Zhx2 Mice and Genotyping ................................................................................................. 37
  Euthanasia and Mouse Tissue Collection ................................................................................................. 38
  Developmental Timepoint Mice ................................................................................................................ 39
  Alignment of Gene Sequences and JASPAR Analysis ............................................................................ 39
  Slp Plasmid Construction ....................................................................................................................... 40
  Stat5b and Stat5b* Plasmid Construction ............................................................................................... 41
  Midiprep of Plasmids for Luciferase Assay ............................................................................................. 42
CHAPTER III .................................................................................................................. 70
A COMPREHENSIVE ANALYSIS OF ZHX GENES AND PROTEINS
THROUGHOUT CHORDATE EVOLUTION................................................................. 70
Identification of gnathostome and early chordate Zhx genes............................ 75
Sea lamprey contains a single functional Zhx-like gene with multiple coding exons .......................................................................................................................... 76
Sea lamprey Zhx has two spliced isoforms present in multiple tissues ............. 76
Characterization of the sea lamprey Zhx pseudogene......................................... 78
Zhx gene structures are dynamic throughout chordate evolution .................... 78
Annotation of ZF and HD regions in Zhx proteins ............................................. 80
Alignment of gnathostome Zhx proteins reveals the highest conservation within the amino terminus ................................................................. 81
Alignment of early chordate and gnathostome Zhx predicted proteins .......... 82
Evolutionary relationships of sea lamprey and gnathostome Zhx proteins ........ 83
Lancelet and sea lamprey Zhx proteins are most highly related to Zhx3 based on ZF region alignments ......................................................................................... 86
HD1 is the most conserved gnathostome Zhx HD .............................................. 88
Lancelet and sea lamprey Zhx HDs are most highly related to Zhx3 ................. 89
DISCUSSION ........................................................................................................... 90

CHAPTER IV ................................................................................................................ 122
ZHX2 REGULATES SEX-LIMITED PROTEIN (Slp) GENE EXPRESSION IN MALE LIVER ......................................................................................................................... 122
Slp and C4 genes are highly similar ..................................................................... 125
Slp mRNA levels are significantly reduced in Zhx2 liver-specific knockout male liver ................................................................. 126
Sex-biased Slp expression is established in the postnatal liver ......................... 126
Sex-biased transcription factor gene expression is unaltered in Zhx2 liver-specific knockout mouse liver ................................................................. 127
Zhx2 positively regulates the Slp promoter ...................................................... 128
Zhx2 and Stat5b positively regulate the Slp 2000 promoter ............................... 130
Igf-1 mRNA levels are not altered in Zhx2 liver-specific knockout mouse liver .. 132
Zhx2, Stat5b, and HNF4α overlapping regulated genes .................................... 132
DISCUSSION ...................................................................................................... 136

CHAPTER V ........................................................................................................ 156
DIFFERENTIAL EXPRESSION OF HIGHLY HOMOLOGOUS MOUSE CYP2A4
AND CYP2A5 GENES .......................................................................................... 156
Cyp2a4 and Cyp2a5 loci are highly similar .......................................................... 159
Cyp2a4 and Cyp2a5 mRNA levels are higher in female liver .............................. 160
Cyp2a4 and Cyp2a5 expression increases after birth ............................................. 160
Cyp2a4 and Cyp2a5 display different zonal expression patterns ......................... 161
Cyp2a4 and Cyp2a5 expression differences may be regulated by transcriptional or
post-transcriptional mechanisms ........................................................................... 163
Cyp2a4as and Cyp2a5as transcripts extend into Cyp2a4 and Cyp2a5 gene
regions ...................................................................................................................... 166
Cyp2a4as transcripts are present as multiple spliced isoforms .............................. 167
Cyp2a4as and Cyp2a5as expression patterns ...................................................... 168
Potential transcription factor binding differences between Cyp2a4as and Cyp2a5as
promoters .................................................................................................................. 170
Cyp2a4as and Cyp2a5as are nuclear-enriched transcripts ..................................... 170
Cyp2a4 and Cyp2a4as are upregulated in a mouse model of HCC ...................... 171
Additional Cyp2 family antisense RNAs identified in male and female C57BL/6
liver .............................................................................................................................. 172
DISCUSSION ......................................................................................................... 173

CHAPTER VI ...................................................................................................... 215
Summary and Future Directions ........................................................................... 215

APPENDIX TABLE 1 .............................................................................................. 220
REFERENCES ...................................................................................................... 223
VITA ...................................................................................................................... 237
LIST OF TABLES

Table 1.1 Sex-biased male and female cytochrome p450 gene expression in mouse liver .................................................................................................................................................................................. 24

Table 2.1 Oligos for RT-qPCR ........................................................................................................................................................................................................................................... 56

Table 2.2 Oligos used for RT-PCR and PCR .............................................................................................................................................................................................................. 61

Table 2.3. Genotyping PCR conditions .................................................................................................................................................................................................................................. 62

Table 2.4. Liver perfusion media recipe .................................................................................................................................................................................................................... 63

Table 2.5. Liver digestion media recipe ....................................................................................................................................................................................................................... 63

Table 2.6. Cyp2a4as and Cyp2a5as oligos used for 5’/3’ RACE and RT-PCR ........................................................................................................................................................................................................ 64

Table 2.7. Cyp2a4as and Cyp2a5as 5’ RACE PCR conditions ........................................................................................................................................................................................................... 65

Table 2.8. Cyp2a4as 5’ RACE PCR conditions .................................................................................................................................................................................................................................. 66

Table 2.9. Cyp2a5as 5’ RACE PCR conditions .................................................................................................................................................................................................................................. 66

Table 2.10. Cyp2a4as 5’ RACE nested PCR conditions ............................................................................................................................................................................................................. 67

Table 2.11. Cyp2a5as 5’ RACE nested PCR conditions ............................................................................................................................................................................................................ 67

Table 2.12. 3’ RACE PCR conditions .......................................................................................................................................................................................................................................... 68

Table 2.13. 3’ RACE nested PCR conditions .................................................................................................................................................................................................................................. 68

Table 2.14. Cyp2a4as and Cyp2a5as nuclear RT-PCR conditions ........................................................................................................................................................................................................ 69

Table 3.1. Percent pairwise identity of Zhx predicted protein zinc fingers and zinc finger regions .................................................................................................................................................................................................................................. 110

Table 3.2. Amino acid matches between Zhx predicted protein ZF regions .................................................................................................................................................................................................................................. 112

Table 3.3. Percent pairwise identity of homeodomains across Zhx predicted proteins ............................................................................................................................................................................. 115

Table 3.4. Amino acid matches between Zhx predicted protein HD regions .................................................................................................................................................................................................................................. 118

Table 4.1. Sex-biased genes significantly affected (FC>2, p<0.05) in BALB/cJ, Stat5b knockout, and HNF4α liver-specific knockout compared to wildtype male mice ........................................................................................................................................................................................................................................ 155

Table 5.1. ApoB1 versus Cyp2a4as and Cyp2a5s average RNA abundance in C3H 8-week male and female liver ........................................................................................................................................................................................................................................ 196

Table 5.2. Coding Potential Assessment Tool (CPAT) evaluation of the Cyp2a4as and Cyp2a5as transcripts ........................................................................................................................................................................................................................................................................................................ 204

Table 5.3 Ratio among Cyp2a4, Cyp2a5, Cyp2a4as, and Cyp2a5as average RNA levels in C57BL/6 male and female liver ........................................................................................................................................................................................................................................................................................................ 207

Table 5.4. Ratio of Cyp2a4, Cyp2a5, Cyp2a4as, and Cyp2a5as average RNA levels between C57BL/6 female and male liver ........................................................................................................................................................................................................................................................................................................ 207
LIST OF FIGURES

Figure 1-1. Diagram of mammalian liver architecture .......................................................... 23
Figure 1-2. Growth hormone activation of Stat5b in hepatocytes ...................................... 25
Figure 1-3. Simplified model for the regulation of sex-biased genes by Stat5b, Bcl6, and Cux2 in adult male and female liver ................................................................................. 27
Figure 1-4. Characterization of long non-coding RNAs throughout the genome ............ 28
Figure 1-5. Mechanisms by which long non-coding RNAs (lncRNAs) can affect expression of protein coding genes ............................................................ 30
Figure 3-1. A phylogenetic tree of chordate evolution within the context of geologic time...................................................................................................................................................... 96
Figure 3-2. Characterization of the sea lamprey Zhx gene ................................................. 97
Figure 3-3. The 2018 sea lamprey genome assembly reveals a predicted Zhx pseudogene .............................................................................................................................. 99
Figure 3-4. Zhx genes in lower chordates show dramatically different structures ........ 101
Figure 3-5. Amino terminus, zinc finger, and homeodomain regions are most conserved across gnathostome Zhx proteins .................................................................................. 103
Figure 3-6. MUSCLE MSA reveals alignment of lancelet, sea squirt, and sea lamprey ZFs and HDs to gnathostome ZFs and HDs .................................................................................. 105
Figure 3-7. Determining the relationship of sea lamprey Zhx to gnathostome Zhx proteins .............................................................................................................................. 108
Figure 3-8. Zinc finger MSA shows high divergence of Zhx3 ZFs compared to those of Zhx1 and Zhx2 ................................................................................................................. 111
Figure 3-9. Neighbor-distance joining phylogeny using zinc finger regions (ZFRs) places sea lamprey within the gnathostome Zhx3 lineage .................................................. 113
Figure 3-10. Gnathostome HD1 helix 3 is the most conserved region in Zhx homeodomains ......................................................................................................................... 116
Figure 3-11. Bayesian phylogeny of all Zhx homeodomains confirms Zhx3 as the primordial chordate Zhx gene ........................................................................................................ 119
Figure 3-12. Proposed model of Zhx and HDs throughout chordate evolution ............... 120
Figure 4-1. Slp and C4 mRNA are highly similar ............................................................. 140
Figure 4-2. Slp is not expressed in Zhx2 liver-specific knockout male mice ................. 142
Figure 4-3. Male-biased Slp expression is established after birth .................................. 143
Figure 4-4. Known sex-biased regulators are unaffected by the absence of Zhx2 in adult liver .............................................................................................................................. 144
Figure 4-5. Stat5b binding sites are present within the Slp promoter region ................ 145
Figure 4-6. Zhx2 positively regulates the Slp promoter .................................................... 146
Figure 4-7. Constitutively active Stat5b positively regulates NTCP GAS elements in vitro ........................................................................................................................................ 147
Figure 4-8. Zhx2 and Stat5b coordinately regulate the Slp promoter .............................. 148
Figure 4-9. Igf-1 mRNA levels are the same in the livers of Zhx2 liver-specific knockout and Zhx2 floxed adult mice ...................................................................................... 150
Figure 4-10. Methods for identifying common genes between Zhx2, Stat5b, and HNF4α microarray datasets ......................................................................................................... 152
Figure 4-11. Common differentially regulated sex-biased and sex-independent genes between Zhx2, Stat5b, and HNF4α microarrays ......................................................................................................... 154
Figure 5-1. The sequences of the mouse Cyp2a4 and Cyp2a5 genes are highly similar ......................................................................................................................................... 182
Figure 5-2. Cyp2a4 and Cyp2a5 are more highly expressed in female liver compared to male liver ..................................................................................................................................... 184
Figure 5-3. Cyp2a4 and Cyp2a5 mRNA levels increase in postnatal mouse liver .... 185
Figure 5-4. Cyp2a4 and Cyp2a5 mRNA display different zonal expression patterns .... 187
Figure 5-5. Cyp2a4 periportal and Cyp2a5 pericentral expression patterns in the CCl4 experimental mouse model ............................................................................................. 188
Figure 5-6. JASPAR predictions of transcription factor binding sites within the mouse Cyp2a4 and Cyp2a5 promoters ........................................................................................................ 189
Figure 5-7. An EST annotated antisense to the Cyp2a4 gene ......................................................................................................................................... 190
Figure 5-8. Initial detection of the Cyp2a4 antisense transcript in male and female mouse hepatocytes ......................................................................................................................................... 191
Figure 5-9. RT-PCR analysis of Cyp2a4as and Cyp2a5as transcripts and regions of similarity between Cyp2a4as and Apob genes ......................................................................................................................................... 193
Figure 5-10. 5’ RACE for Cyp2a4as and Cyp2a5as transcripts ......................................................................................................................................... 197
Figure 5-11. 3’ RACE identifies the 3’ ends of Cyp2a4as and Cyp2a5as transcripts .... 200
Figure 5-12. RT-PCR of full length Cyp2a4as and Cyp2a5as transcripts ......................................................................................................................................... 202
Figure 5-13. Diagram of the Cyp2a4as spliced isoforms and the single Cyp2a5as spliced isoform identified by cloning and sequencing RT-PCR products ......................................................................................................................................... 203
Figure 5-14. Cyp2a4as and Cyp2a5as expression patterns mimic those of their overlapping mRNA counterparts ......................................................................................................................................... 205
Figure 5-15. JASPAR predictions of transcription factor binding sites within the potential promoters of mouse Cyp2a4as and Cyp2a5as ......................................................................................................................................... 208
Figure 5-16. Cyp2a4as and Cyp2a5as are nuclear-enriched transcripts ......................................................................................................................................... 209
Figure 5-17. Cyp2a4as and Cyp2a5as expression patterns are similar to Cyp2a4 and Cyp2a5 in DEN-induced tumors ......................................................................................................................................... 210
Figure 5-18. Additional Cyp2 family antisense and intergenic RNA transcripts are not sex-biased and are expressed at very low levels in mouse liver ......................................................................................................................................... 211
Figure 5-19. Cyp2a4 and Cyp2a5 DNAseI hypersensitivity regions in females overlap with the Cyp2a4as promoter ......................................................................................................................................... 212
CHAPTER I
INTRODUCTION

Liver Architecture

The liver is the largest internal organ in mammals and performs numerous functions including: the production of serum proteins, clotting factors, and transport proteins such as albumin and transferrin; the removal and breakdown of serum proteins, red blood cells, and microbes; the production or breakdown of glucose; the metabolism of fatty acids and triglycerides; maintenance of cholesterol homeostasis; synthesis and interconversion of non-essential amino acids; the breakdown of toxic endogenous compounds such as ammonia; the production and excretion of bile components; and the detoxification of xenobiotic agents (Spear et al., 2006). Two blood sources supply the liver, the hepatic artery and portal vein. The hepatic artery delivers oxygen-rich blood to the liver, whereas the portal vein delivers nutrients absorbed by the intestinal system. In order to carry out its variety of functions and maintain metabolic homeostasis, the liver is divided into hexagonal functional units called liver lobules. Lobules contain three to six portal triads (comprised of the hepatic artery, portal vein, and bile duct) at the edge, and lobules converge to a central vein in the center (Figure 1-1). Bile ducts contain bile that flows out of the liver through right and left hepatic ducts, which drain into a common hepatic bile duct. The common hepatic bile duct joins with the cystic duct from the gallbladder and empties bile into the small intestine where it can be used for the absorption of lipids or be eliminated in feces. The liver is composed of multiple cell types including hepatocytes (also called parenchymal cells), biliary epithelial cells
(cholangiocytes), stellate cells, Kupffer cells (resident liver macrophages), and liver sinusoidal endothelial cells (Trefts et al., 2017). Hepatocytes comprise the majority of the liver, representing 60% of cells and 80% of the liver mass (Spear et al., 2006).

Liver Development and Gene Expression in Developing Liver

The process of liver organogenesis is well-conserved across vertebrate species (Gordillo et al., 2015). Both hepatocytes and cholangiocytes are derived from endoderm that emerges from the anterior primitive streak of the gastrulating embryo by embryonic day 7.5 (E7.5) in mouse (Gordillo et al., 2015). The embryonic endoderm gives rise to the liver, pancreas, lungs, thyroid, and gastrointestinal tract (Tremblay and Zaret, 2005). Mouse liver organogenesis begins at roughly embryonic day 9 (E9), when the ventral domain of the foregut adjacent to the cardiac mesoderm and septum transversum thickens to form the liver diverticulum (Gordillo et al., 2015; Houssaint, 1980). Early studies by Douarin first showed the requirement of mesoderm for endoderm differentiation into hepatic parenchymal cells (Douarin, 1975). Chimeric embryo and tissue co-culture experiments in chicken, frog, mouse, and zebrafish have shown that liver development is controlled by signals from the mesoderm (Gordillo et al., 2015). Fibroblast growth factor 1 (FGF1), FGF2, bone morphogenetic protein (BMP), and GATA binding protein 4 (GATA4) produced from mesoderm coordinate to induce liver bud formation in both chicken and mouse (Jung et al., 1999; Zhang et al., 2004). In mice, by E10.5, the hepatocyte marker alpha-fetoprotein (AFP) mRNA can be detected in ventral endodermal cells that are contacted by septum transversum mesenchyme (Shiojiri et al., 1991). The septum transversum mesenchyme forms the liver diverticulum, and the diverticulum
thickens and transitions to form hepatoblasts which spread, proliferate, and invade the septum transversum to form the liver bud. The fetal liver at E12.5 contains hematopoietic stem cells, that are responsible for hematopoiesis, and hepatoblasts, that are marked by AFP and albumin (ALB) mRNA and protein (Gordillo et al., 2015). By E13.5, hepatoblasts differentiate into cholangiocytes (bile duct epithelial cells) and hepatocytes. As hepatocytes mature, the primary function of the liver switches from hematopoiesis to metabolism. By E17, roughly 30% of the liver is responsible for hematopoiesis (Sasaki and Sonoda, 2000). At this point, hematopoiesis becomes established within bone marrow (Mikkola and Orkin, 2006).

Liver gene expression profiles shift during postnatal development and dictate changes in hepatocyte phenotype and function (Li et al., 2009). Microarray studies in C57BL/6 mice that span E11.5 through 18 weeks showed that genes activated or repressed during different stages of development were enriched for specific transcription factor binding motifs. This analysis identified that early in development, Sp1 transcription factor (Sp1), E2F transcription factor 1 (E2f1), and myocyte enhancing factor 2 (Mef2) motifs were enriched for genes that were expressed; whereas hepatocyte nuclear factor 4 alpha (Hnf4α), nuclear factor 1 (NF1) and Smad family member 3 (Smad3) were motifs that were underrepresented during this same time frame. Inversely, Hnf4α and Smad3 motifs were enriched for genes activated late in development [postnatal day 7 (P7) to 18 weeks]. More recently, single cell RNA-sequencing and single-cell qPCR of liver collected from developmental timepoints E11.5, E12.5, E13.5, E14.5, E16.5, E18.5, P2.5, and P3.25 has provided a more comprehensive understanding
of the developmental gene expression changes that occur throughout liver development (Su et al., 2017).

**Zinc Fingers and Homeoboxes 2 (Zhx2)**

Understanding gene regulation throughout perinatal development, including switches that occur to inactivate and activate genes, is important because genes that are activated during fetal periods are often reactivated in cancer. The *AFP* gene normally remains silent in the adult liver, but can be reactivated during liver regeneration and in hepatocellular carcinoma (Abelev and Eraiser, 1999; Spear et al., 2006).

Initial insight into AFP regulation came from the investigations of serum levels from 27 different strains of mice; the objective was to find a model system for AFP regulation where AFP was not efficiently “switched off” (Olsson et al., 1977). Among these strains, BALB/cJ mice continued to have higher adult serum AFP levels compared to all other mouse strains. Backcrossing BALB/cJ to DBA/2, a strain with low adult serum AFP levels, revealed that high AFP levels in BALB/cJ mice were due to a recessive monogenic trait.

Tilghman and colleagues were the first to determine that steady state liver AFP mRNA levels continuously decreased in BALB/cJ similarly to other strains until one week after birth; by week two there was a considerable slowing of decrease rate; by four weeks AFP mRNA levels persisted at levels roughly ten-fold higher in BALB/cJ mice compared to C3H/He and C57BL/6 strains (Belayew and Tilghman, 1982). They named the gene responsible for this incomplete AFP repression phenotype, regulator of AFP, or
Raf. Shortly after this, the Tilghman lab cloned H19 mRNA, which was later shown to be a long noncoding RNA, and another gene regulated by Raf (Brannan et al., 1990; Pachnis et al., 1984; Pachnis et al., 1988). Raf, later renamed alpha-fetoprotein regulator 1 (Afr1), was mapped to chromosome 15 (Blankenhorn et al., 1988). Peyton et al., using transgenic mice, were the first to demonstrate that Afr1 repressed AFP through its 250 bp promoter (Peyton et al., 2000). In addition to Afr1 transcriptional regulation of target genes, it was also proposed that Afr1 may regulate AFP at the post-transcriptional level. This proposal was based on nuclear run on assays that showed no differences in rate of AFP transcription between Afr1 heterozygous and homozygous mice (Vacher et al., 1992).

Positional cloning identified Zhx2 as the gene responsible for the Afr1 phenotype (Perincheri et al., 2005). RT-PCR confirmed that Zhx2 mRNA levels were very low in BALB/cJ mice compared to other strains of mice. Dramatic reduction of BALB/cJ Zhx2 mRNA levels is caused by insertion of a MERV-K family retrotransposon in Zhx2 intron 1, which interrupts normal splicing and dramatically reduces wild-type Zhx2 transcripts. Taken together, lower levels of Zhx2 in BALB/cJ livers results in higher mRNA levels of target genes AFP and H19 in the adult liver.

In addition to AFP and H19, Zhx2 has been found to influence the expression of many genes in mice including Glypican 3 (Gpc3), Lipoprotein lipase (Lpl), and several sex-biased genes [major urinary proteins (Mups), and cytochrome p450 (Cyps)] (Creasy et al., 2016; Gargalovic et al., 2010; Ma et al., 2015; Morford et al., 2007; Yue et al., 2012). Gpc3, like AFP and H19, is expressed abundantly in the fetal liver, and HCC, is silent in normal adult liver, and has persistent mRNA expression in BALB/cJ compared to C3H/HEJ adult mouse liver (Morford et al., 2007).
A major phenotypic trait of BALB/cJ mice is reduced triglycerides, total cholesterol, and LDL/VLDL cholesterol levels, in comparison to other mouse strains, when put on an atherogenic, high fat diet (Wang et al., 2004). This trait was attributed to the hyperlipidemia 2 (Hyplip2) locus mapped to chromosome 15. Quantitative trait locus and transgenic mouse studies identified Zhx2 as the causal gene for the Hyplip2 phenotype (Gargalovic et al., 2010). Microarray analysis from the Lusis lab, between Sub13 male mice, which express normal Zhx2 mRNA levels, and BALB/cJ male mice, revealed 1084 differentially expressed liver genes (Gargalovic et al., 2010). This microarray analysis has been instrumental in discovering new gene targets that are regulated by Zhx2.

The Hyplip2 microarray data expanded our views of potential Zhx2 sex-biased gene targets that were upregulated and downregulated in BALB/cJ versus Sub13 mice. Expression of one family of genes, Mups, were dramatically decreased in BALB/cJ mice. A separate study showed decreased Mup proteins isolated from male BALB/cJ urine (Cheetham et al., 2009) and that this trait mapped to chromosome 15 (Duncan et al., 1988). Mups are predominately expressed in liver and are male-biased. Mup gene expression has been shown to be under multihormonal regulation; growth hormone, thyroxine, and testosterone all contribute to Mup gene expression (Knopf et al., 1983; Kuhn et al., 1984; Shaw et al., 1983). Our lab has confirmed several highly expressed Mups including, Mup20, Mup3, Mup7, Mup10, and Mup19, are expressed at much lower levels in Zhx2 hepatocyte-specific knockout (Zhx2ΔHep) and whole body Zhx2 knockout (Zhx2KO) male livers (Jiang et al., 2017). In addition, in vitro luciferase assays suggest that Zhx2 regulates Mup20 expression through an element within the Mup20 promoter.
Also, chromatin immunoprecipitation experiments (ChIP) showed binding of Zhx2 to the 
*Mup20* promoter *in vivo*.

Similarly to Mups, expression of several *Cyps* genes is significantly changed in 
BALB/cJ versus Sub13 mice. A dramatic example of female-biased *Cyp* gene expression 
is seen with *Cyp2a4*; in normal female and male liver, Cyp2a4 mRNA levels are high and 
low, respectively. Creasy et al. showed elevated mRNA expression of Cyp2a4 in male 
Zhx2^D^Hep livers compared to Zhx2^fl/fl^ (Creasy et al., 2016). In addition, adult male 
Zhx2^D^Hep liver showed elevated mRNA levels for female-biased Cyps, including Cyp2a5, 
Cyp2b9, and Cyp2b13 (Table 1.1). However, male-biased and sex-independent Cyps 
were unaffected in male and female Zhx2^D^Hep livers.

**Zhx Gene Family**

Homeobox genes contain homeobox sequences that encode for a protein domain 
named the homeodomain (Holland et al., 2007). Homeodomain (HD) transcription 
factors are involved in a variety of biological functions (Burglin and Affolter, 2016). The 
HD is 60 amino acids long and contains key residues important for the hydrophobic core 
of the HD; these include leucine (L16), phenylalanine (F20), tryptophan (W48), and 
phenylalanine (F49) (Burglin and Affolter, 2016). However, in some HDs these 
conserved residues are substituted with amino acids that have similar properties.

The Zinc finger (ZF) class of homeobox genes contain C2H2 and C2H2-like ZF’s 
in addition to HD’s. C2H2-type ZFs are typically involved in DNA binding and the 
number of zinc fingers and number of HDs can vary (Burglin and Affolter, 2016). The
ZF structure is maintained by the zinc ion; in the case of C2H2-type ZFs, two cysteines in one chain and two histidines in the other chain are coordinated by one zinc ion (Cassandri et al., 2017). In vertebrates, five families comprise the ZF class: Adnp, Tshz, Zeb, Zfhx, and Zhx. The Adnp and Zhx families seem to be vertebrate-specific, whereas Tshz, Zeb, and Zfhx are conserved across the bilaterian divide (i.e. Drosophila) (Burglin and Affolter, 2016). Zhx2, along with Zhx1 and Zhx3, are members of a small family of genes, the Zinc fingers and homeoboxes (Zhx) family. Zhx genes have unusual structures where the entire protein coding regions are found on an exceptionally large internal exon; Zhx3 has a short coding region in the fourth exon (Spear et al., 2006). Zhx proteins are ubiquitously expressed and contain two C2-H2-type ZF motifs and four or five HDs (Kawata et al., 2003a). Interestingly, HOMEZ is also included within the Zhx class but does not contain any zinc fingers (Holland et al., 2007). Evidence for this relationship is shown by multiple sequence alignment of Homez and Zhx predicted protein homeodomain regions (Bayarsaihan et al., 2003).

Zhx1 was the first Zhx protein identified by mouse bone marrow stromal cell line cDNA library screening to find proteins important for myeloid progenitor cell differentiation (Barthelemy et al., 1996; Lee et al., 1996). Mouse Zhx1 was later found to interact with NF-YA in a yeast two hybrid screen (Yamada et al., 1999a). Human ZHX1 was found to also interact with NF-YA by yeast two-hybrid screening and was mapped to chromosome 8 (Yamada et al., 1999b). Subsequent studies found that rat Zhx1 homodimerizes and is localized to the nucleus (Hirano et al., 2002; Yamada et al., 2002). Human ZHX1 was later shown to homodimerize and heterodimerize with ZHX3, ZHX2, as well as NF-YA (Kawata et al., 2003b; Yamada et al., 2002; Yamada et al., 2003).
Additional studies revealed that mouse Zhx2 and Zhx3 could also form heterodimers based on yeast-two hybrid screening (Kawata et al., 2003a). Taken together, these data suggest that Zhx1, Zhx2, and Zhx3 can heterodimerize with one another and with NF-YA.

**Cytochrome p450 Genes in Mouse and Human**

Cytochrome p450s (Cyts; mouse, CYPs; human) are heme-containing membrane-bound enzymes critical for removal of xenobiotics such as drugs, foreign chemicals, and other pollutants; arachidonic acid metabolism; eicosanoid biosynthesis; cholesterol, sterol and bile acid biosynthesis; steroid catabolism; vitamin D₃ synthesis and catabolism; and retinoic acid hydroxylation, as well as other functions (Nelson et al., 2004). There are 57 putatively functional CYP genes identified in humans, whereas there are 102 putatively functional Cyp genes in the mouse (Nelson et al., 2004). The liver is the main site for Cyp-mediated metabolism, however other tissues such as intestine, lung, and kidneys also express Cyts. Compared to the human, mice have undergone a significant expansion of Cyp genes in seven clusters; the Cyp2abfgst cluster, the Cyp2c cluster, the Cyp2d cluster, the Cyp2j cluster, the Cyp3a cluster, the Cyp4abx cluster, and the Cyp4f cluster (Nelson et al., 2004). The expansion of these Cyp families are proposed to offer evolutionary advantages for mice that encounter various plant toxins. Many liver-enriched transcription factors and nuclear receptors appear to be involved in regulating liver-specific expression of Cyts (Akiyama and Gonzalez, 2003).
Liver Zonation and Regulation of Cyts

Liver metabolic homeostasis is maintained by hepatocytes that carry out different functions depending on their spatial location within the liver lobule. Hepatocytes are deemed periportal (PP) if they are closer to the portal triad, whereas those localized closer to the central vein are pericentral (PC). This zonation of metabolic function across the liver lobule includes ammonia detoxification, glucose and energy metabolism, xenobiotic metabolism, lipid metabolism, bile acid synthesis, amino acid metabolism, and glutamine synthesis. Some metabolic processes within hepatocytes are not zonal, including synthesis of serum proteins, such as albumin. Zonation is a dynamic feature that can change in response to nutrition, drugs, hormones, oxygen content, and is governed by differential transcriptional and post-transcriptional regulation between periportal and pericentral regions.

Gene expression patterns can vary within the liver lobule. For instance, zonal regulation can exhibit a gradient pattern across the liver lobule. However, expression of some zonal genes is highly restricted to PC or PP regions. For example, glutamine synthetase (GS) expression is highly restricted to one-two layers of hepatocytes surrounding the central veins. Previously, our lab has demonstrated that three enhancers upstream of AFP (E1, E2 and E3), confer differential zonal expression patterns of linked transgenes in mouse liver. The most striking pattern is seen with E3, which is active only in 1-2 layers of hepatocytes around the central vein similarly to GS; E1 and E2 exhibit a more gradual reduction in expression across the PC-PP axis (Ramesh et al., 1995).

Although many enzymes exhibit zonal expression patterns, mechanisms that regulate zonal expression patterns are not fully understood. Currently, zonal gene
expression is thought to be regulated primarily at the level of transcription (Spear et al., 2006). A major known regulator of zonal expression is the wingless (Drosophila gene) integrated (vertebrate homolog of the Drosophila wingless gene) (Wnt) / beta-catenin (β-cat) pathway. In the absence of Wnt signaling, adenomatous polyposis coli (APC), glycogen synthase (GSK-3β), casein kinase 1(CK1), and Axin scaffolding protein (Axin) form the degradation complex (Stamos and Weis, 2013). In the absence of Wnt proteins, CK1 and GSK-3β phosphorylate the β-cat amino terminal region, and β-Tcrp, an E3 ubiquitin ligase subunit, subsequently ubiquitinates β-cat and targets it for proteosomal degradation (MacDonald et al., 2009). As a result, β-cat is prevented from reaching the nucleus, and Wnt target genes are repressed by DNA-bound T-cell factor/lymphoid enhancer factor (TCF/LEF) family proteins that recruit co-repressors, such as Groucho and histone deacetylases.

In the presence of Wnt signaling, Wnt protein binds to the seven-pass transmembrane Frizzled receptor (Fz), and its co-receptor, low-density lipoprotein receptor related protein 6 (LRP6), or LRP5. Disheveled protein is then recruited to Fz and LRP6. Relocation of the degradation complex to the plasma membrane, mediated by Disheveled protein, prevents phosphorylation and degradation of β-cat (Cliffe et al., 2003). As a result, β-cat accumulates and translocates to the nucleus, where it interacts with TCF/LEF bound to target genes, dissociates repressors from this complex, and activates target genes. Gene activation occurs through CBP/p300 coactivator recruitment to the TCF/LEF/β-cat complex (Cadigan and Waterman, 2012).
Previous studies by Benhamouche et al. revealed that activated β-cat is highly enriched in PC hepatocytes whereas APC is enriched in PP hepatocytes (Benhamouche et al., 2006). Conditional knockout of APC in hepatocytes resulted in β-cat stabilization and subsequent activation of PC genes GS and Ornithine aminotransferase (OAT) throughout the liver lobule. Complementary experiments blocking the β-cat pathway using an adenoviral vector expressing Dickkopf-1 (Dkk1), an extracellular antagonist of Wnt signaling, leads to the adoption of a PP phenotype across the liver lobule. Microarray analysis between conditional APC \textsuperscript{KO} and wildtype mice demonstrated that genes such as GS, Axin2, and OAT were significantly upregulated in APC \textsuperscript{KO} mice. These studies establish β-cat as a master regulator of zonal gene regulation.

Mechanistic studies utilizing mouse primary hepatocytes have also demonstrated that in vitro differentiated resident liver stem cell (RLSC) gene expression patterns can be changed from a PP to PC phenotype by inhibition GSK-3β (Colletti et al., 2009). Parviz and colleagues further investigated the transcriptional machinery downstream from β-cat activation and demonstrated direct protein interactions between LEF1 and HNF4α (Colletti et al., 2009). Chromatin immunoprecipitation (ChIP) experiments showed that HNF4α bound to HNF4α consensus sequences upstream of GS (PC gene) and glutaminase 2 (Gsl2), a PP gene, in cultured RLSC’s in the absence of GSK-3β inhibition. RLSC’s treated with GSK-3β inhibitor displaced HNF4α, and LEF1 bound to both upstream HNF4α and LEF consensus sequences of GS and Gsl2. These data suggested an activating function for HNF4α in PP hepatocytes, whereas HNF4α might have a repressive role in PC hepatocytes. These results were consistent with a previous
report by Stanulovic et al, where HNF4a was required for PP-specific suppression of GS and OAT protein levels (Stanulovic et al., 2007). Additionally, interactions between β-cat, HNF4a, and TCF4 led to a proposed model where TCF4 and HNF4a activate PP-specific genes, and TCF4 and β-cat activate PC-specific genes (Gougelet et al., 2014).

More recently, Wnt/β-cat signaling has been found to determine the hepatic zonation of Cyp gene expression. Early evidence for Wnt/β-cat signaling regulation of mouse Cyp genes combined N-nitrosodiethylamine- (DEN) treatment and phenobarbital-induced mutations within the β-cat gene, CTNNB1 (Loeppen et al., 2005). Tumors containing CTNNB1 mutations showed strong upregulation of GS, as well as Cyp isoforms Cyp1a, Cyp2b, Cyp2c, and Cyp2e1 based on immunohistochemical staining.

These results were extended by Sekine and colleagues, where Cyp1a1, Cyp3a11, Cyp1a2, Cyp2e1, and Cyp2c29 mRNA levels were compared between wildtype and hepatocyte-specific β-catKO male mouse livers (Sekine et al., 2006). Cyp1a1 and Cyp3a11 were not affected by the presence or absence of β-cat, whereas Cyp1a2, Cyp2e1, and Cyp2c29 gene expression was significantly decreased in the absence of β-cat. Additional analyses using digitonin/collagenase perfusion of male livers, which separates PP and PC hepatocyte populations, identified additional Cyp isoforms that were zonally regulated and restricted to PC hepatocytes (Braeuning et al., 2007; Braeuning et al., 2006). Microarray analysis revealed that a majority of Cyps were preferentially expressed in PC hepatocytes. Two isoforms, Cyp2f2 and Cyp39a1, were preferentially expressed in PP hepatocytes. Taken together, these results suggest that the majority of Cyp isoforms analyzed are pericentrally expressed whereas few are expressed periportally, and several are not affected by changes in β-cat.
Sex-Biased Gene Regulation

In 1932, Nicholas and Barron were the first to recognize that female rats required half the dose of barbiturate to be anesthetized compared to males (Barron, 1932; Mode and Gustafsson, 2006). Sex differences in barbiturate hepatic metabolism were found to result in dosage disparity. Since then, research has established that the gonadal-hypothalamus-pituitary liver axis determines sex-biased liver gene expression and the metabolism of steroids and drugs (Mode and Gustafsson, 2006; Waxman and Celenza, 2003).

Growth hormone (GH) is a 191-amino acid protein hormone secreted by the anterior pituitary gland that is responsible for mediating the effects of testosterone and estrogen on liver drug and steroid metabolism (Waxman and Holloway, 2009). Neonatal exposure to testosterone imprints male neuroendocrine control of pulsatile pituitary GH secretion first seen at puberty and androgen absence during the neonatal period results in female GH secretory patterns (Fernandez-Perez et al., 2013). Anterior pituitary GH release is mediated by hypothalamic GH-releasing hormone (GHRH) and ghrelin, and inhibited by somatostatin (Chilton and Hewetson, 2005).

In many species, including rats, mice, and humans, pituitary pulsatile GH secretion is more frequent in females than in males. In females, a high pulse frequency results in a near-continuous GH presence in plasma. In contrast, the male pituitary release intermittent pulses of GH; peaks of GH in plasma are often higher than female peaks, and troughs between peak secretions are characterized by almost undetectable GH levels. Disruption of the GH-free trough in males can be achieved by treating mice or rats with exogenous, continuous GH. The continuous infusion of GH induces mRNA expression of
female-biased genes and reduces mRNA expression of male-biased genes in the liver. Additionally, GH pulsing can be abolished by hypophysectomy (removal of the pituitary gland) and male-biased genes can be induced by administration of exogenous, pulsing GH (Waxman et al., 1991).

Growth hormone regulates liver gene expression by binding to growth hormone receptor (GHR) which exists as a dimer at the hepatocyte cell surface (Figure 1-2) (Chilton and Hewetson, 2005; Waters, 2016). GH binding induces a conformational change in the GHR, leading to phosphorylation of GHR cytoplasmic domains by Janus kinase 2 (JAK2)(Argetsinger et al., 1993; Carter-Su et al., 2015; Holloway et al., 2006; Waters, 2015). Binding of Jak2 induces tyrosine phosphorylation of both Jak2 and GHR (Argetsinger et al., 1993; Herrington and Carter-Su, 2001). Phosphorylation of the GHR produces docking sites for downstream signaling proteins, including signal transducers and activators of transcription 5a and 5b (STAT5a/b)(Waxman et al., 1995). STAT5b is the predominant form of Stat5 in hepatocytes, based on mRNA and western blot analysis (Choi and Waxman, 1999). Stat5a/b are normally found within the cytoplasm of cells and become activated by the male pulsatile GH pattern in liver (Waxman et al., 1995). Although Stat5b is also present in females, activated Stat5b protein levels are much lower based on phosphotyrosine antibody detection (Waxman et al., 1995). Stat5 binding to Jak2 induces STAT5a/b phosphorylation. Phosphorylated Stat5a/b proteins can then homodimerize and translocate into the nucleus to regulate transcription of target genes (Holloway et al., 2006; Quelle et al., 1996). Although Stat5a and Stat5b contain >90% sequence similarity, Stat5a is not able to compensate for Stat5b loss in mouse knockout models (Udy et al., 1997; Waxman and Holloway, 2009). Whereas Stat5a is required for
normal mammary development, Stat5b is required for sex-biased gene expression in adult liver (Waxman and Holloway, 2009). Stat5b signaling can be terminated by dephosphorylation of activated GHR-Jak2 phosphorylated complexes, recycling or degradation of GHR-Jak2 complex, or dephosphorylation of Stat5b (Gebert et al., 1999).

The GH signaling cascade leads to multiple changes in gene transcription (Waxman and O’Connor, 2006). One target of transcriptional regulation by GH and Stat5b is Igf-1, a protein hormone secreted by the liver and other target tissues that mediates the growth promoting effects of GH (Bichell et al., 1992; Eleswarapu et al., 2008; Udy et al., 1997; Wang and Jiang, 2005). To understand the sex-biased regulation imparted by Stat5b specifically, whole-body Stat5b knockout mice (Stat5b−/−) were generated by Udy et al. (Udy et al., 1997). Stat5b global loss resulted in decreased growth rates in males, decreased Mup protein levels, increased Cyp2a4 mRNA levels, and decreased Igf-1 plasma protein levels. Microarray analysis from these livers revealed marked losses in sex-biased liver gene expression patterns for male livers, whereas female livers were modestly affected. Therefore, Stat5b is thought to play a more important role in male-biased sex patterns in liver. Stat5b−/− males had decreased expression of many male-biased genes in liver including Mup1, Mup3, Mup4, Cyp4a12, Cyp2d9 and Slp (Clodfelter et al., 2006; Holloway et al., 2006). In addition, several female-biased genes were upregulated in male liver including Cyp2b9, Cyp2b10, Cyp2b13 and Cyp4a14.

Similarly to Stat5b, HNF4α contributes to male-biased gene expression and negatively regulates female-specific Cyps in male mouse liver (Holloway et al., 2006). Male-biased expression of Mup1, Mup3, Cyp4a12, Cyp2d9, and Slp are significantly
decreased in male liver lacking HNF4α or Stat5b (Holloway et al., 2006). Concomitantly, female specific liver Cyp2b9 and Cyp2a4 mRNA levels are significantly increased in male liver lacking HNF4α or Stat5b (Holloway et al., 2006).

Global gene expression analysis has shown that some genes respond rapidly to GH, many of which are known to be targets of Stat5b, including Igf-1 (Wauthier et al., 2010). Genes that exhibit delayed (+30 min, up to 90 minutes) responses to exogenous GH administration are proposed to be regulated by other transcription factors. One of these early response genes (response within 30 minutes of GH pulse), B cell CLL/Lymphoma 6 (Bcl6), is activated by male GH pulsing and is repressed by continuous female GH (Meyer et al., 2009). Bcl6 and Stat5b ChIP revealed overlap binding of Bcl6 and Stat5b within female-biased genes that are repressed in male liver. This data suggests that Bcl6 is a male-biased transcriptional repressor of female genes in male liver (Zhang et al., 2012) (Figure 1-3).

Microarray analysis has shown that Cut-Like Homeobox 2 (Cux2 or Cutl2) is highly expressed in female liver (Clodfelter et al., 2006; Laz et al., 2007). Further investigation showed that Cux2 expression is activated by female GH patterns (Laz et al., 2007) (Figure 1-3). Stat5b and Bcl6 have been shown to bind to the upstream region of the Cux2 gene in male liver (Zhang et al., 2012). Therefore, Cux2 expression is negatively regulated by binding of Stat5b and Bcl6 in male liver. Furthermore, male liver-specific HNF4α knockout mice (HNF4αΔHep) have increased Cux2 mRNA levels, suggesting negative regulation of Cux2 mRNA expression by HNF4α in male liver (Laz et al., 2007). Taken together, these results suggest that Cux2 expression is repressed in male liver, whereas Cux2 expression is activated in female liver.
Non-coding RNAs

High throughput sequencing of mammalian transcriptomes has revealed nearly ubiquitous low-level transcription of the genome and the presence of a large number of noncoding transcripts (ncRNA) (Melia et al., 2016). Many of these pervasive ncRNA transcripts have low species conservation, low transcript levels, and uncharacterized function. Some investigators have challenged the functional relevance of these transcripts, arguing that technical limitations, alternative splicing, or extensions of known protein coding genes may account for these spurious, low abundant transcripts (Kung et al., 2013). However, given that is has been long known that numerous noncoding transcripts such as transfer RNAs (tRNAs), ribosomal RNA (rRNAs), and spliceosomal RNAs are critical components of many cellular machines, it seems likely that additional ncRNAs play key regulatory and functional roles (Wilusz et al., 2009).

Within the broad category of ncRNAs, transcripts are grouped based on length, localization, and function (Kashi et al., 2016). Non-coding RNAs include transfer RNA (tRNA), ribosomal RNA (rRNA), small noncoding RNA (sRNA), and long non-coding RNA (lncRNA). Classic ncRNAs that regulate mRNA splicing and translation include snRNA, rRNA, and tRNA (Li and Chang, 2014). Small noncoding RNA include small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), piwi-interacting RNA (piRNA), microRNA (miRNA), and small interfering RNA (siRNA). Short 21-28 nt regulatory RNAs involved in transcriptional and posttranscriptional gene silencing include piRNA, miRNA, and siRNA. LncRNA are defined as non-coding transcripts longer than 200 nucleotides that do not code for proteins. Based on their location within the genome with respect to protein coding genes, lncRNAs can be classified into the following categories:
sense, antisense, intronic, divergent (or bidirectional), intergenic, and enhancer lncRNAs (Figure 1-4) (Kazemzadeh et al., 2015; Rinn and Chang, 2012; Zhang et al., 2014).

LncRNAs can participate in the control of gene expression by mechanisms in both the cytoplasm and the nucleus (Figure 1-5) (Zhang et al., 2014). Within the cytoplasm, lncRNAs can facilitate mRNA decay, stabilize mRNAs, and promote or inhibit the translation of target mRNAs through extended base-pairing. Additionally, lncRNAs can function as a precursor for miRNAs or compete for miRNA-mediated binding that leads to the increased expression of mRNA (Zhang et al., 2014). In the nucleus, lncRNA can control the epigenetic state of genes, participate in transcriptional regulation, regulate alternative splicing, and organize or regulate subnuclear compartments (i.e. silencing lncRNA Neat1 prevents paraspeckles, or nuclear bodies)(Bond and Fox, 2009).

Antisense LncRNAs

Antisense transcripts are a class of lncRNAs that are transcribed from the strand opposite to sense transcripts of either protein-coding or non-protein coding genes (Pelechano and Steinmetz, 2013). Antisense transcription is carried out by RNA polymerase II and can arise from independent promoters, shared bidirectional promoters, or cryptic promoters that are located within transcribed regions of their sense genes (Core et al., 2008; Pelechano and Steinmetz, 2013; Spicer and Sonenshein, 1992). Once transcribed, antisense RNA contain sequences that may be partially or completely complementary to sense transcripts and can interact with sense RNAs through complementary base pairing (Lin et al., 2015). In contrast to protein-coding RNAs which
accumulate in the cytoplasm, antisense lncRNAs preferentially accumulate in the nucleus (Pelechano and Steinmetz, 2013). Furthermore, antisense lncRNAs are typically less abundant than their sense counterparts (Faghihi and Wahlestedt, 2009).

Antisense lncRNAs are functionally diverse and can act as positive and negative regulators of protein-coding genes. The function of an antisense transcript can be facilitated by the transcript itself or the act of its transcription, and antisense RNA can act both in cis and in trans (Pelechano and Steinmetz, 2013). Cis-acting mechanisms include local interactions, such as promoter or gene interactions, whereas trans-acting mechanisms act distally, such as by enhancer-gene interactions or by affecting expression of other genes. Antisense lncRNAs affect many steps within the biogenesis and localization of mRNA including transcription, mRNA splicing, RNA stability, and translation efficiency (Pelechano and Steinmetz, 2013; Villegas and Zaphiropoulos, 2015). In addition, the functions of antisense lncRNA can fall into three categories: lncRNA-DNA, lncRNA-RNA, and lncRNA-protein interactions.

Antisense lncRNAs can affect gene transcription; acting as scaffolds to guide proteins to specific parts of the genome and affecting histone modifications. APOA1-AS, an antisense transcript that is expressed opposite of the APOA1 gene, is an antisense lncRNA that can function in protein recruitment, histone modification, lncRNA-DNA, and lncRNA-protein interactions. APOA1-AS recruits chromatin modifying complexes to the APO gene cluster and induces gene silencing (Halley et al., 2014). Another study has demonstrated that Air, another antisense lncRNA, interferes with transcription of Slc22a3 by histone methyltransferase G9a recruitment which silences Sl22a3 expression (Nagano et al., 2008). In addition to these examples, antisense RNAs have been shown to
influence DNA methylation as well as affect transcription initiation by transcriptional interference (Pelechano and Steinmetz, 2013).

Sense and antisense RNA can hybridize and form RNA duplexes by complementary base pairing (Villegas and Zaphiropoulos, 2015). ZEB2-AS, an antisense lncRNA that prevents Zeb2 5’ splicing, ultimately increases Zeb2 protein levels. Zeb2 then downregulates E-cadherin inducing an epithelial to mesenchymal transition. This example illustrates a posttranscriptional and lncRNA-RNA mechanism in which antisense RNAs can affect mRNA splicing. Antisense lncRNAs have also been shown to affect translational efficiency and RNA stability in bacteria, mice, and humans (Pelechano and Steinmetz, 2013).

Hypotheses

My dissertation explores two aspects surrounding Zhx2. In the first part of my thesis, I explore Zhx proteins throughout chordate evolution using phylogenetic analysis and identify that a Zhx gene exists in the basal chordate, Branchiostoma floridae (lancelet). Furthermore, I identify that early chordate Zhx genes are most highly related to Zhx3. Based on my data, I hypothesize that a Zhx3-like (Zhx3L) gene is the most ancestral Zhx gene that gave rise to Zhx3L and Zhx1L genes.

The second part of my thesis explores regulation of Zhx2 sex-biased target genes. Previous analysis from our lab has identified that some female-biased genes are upregulated in male Zhx2\textsuperscript{Hep} liver, whereas some male-biased genes are downregulated. In chapter four, I find that male Zhx2\textsuperscript{Hep} liver has ablated sex-limited protein (Slp)
expression. I hypothesize that male S1p expression is dependent upon both Stat5b and Zhx2.

Previous results from our lab showed that Cyp2a4 mRNA levels are significantly upregulated in Zhx2^ΔHep^ male liver, whereas Cyp2a5 mRNA levels are not significantly affected by Zhx2 loss. Additionally, Cyp2a4 expression is highly female-biased whereas Cyp2a5 expression is slightly female-biased, and previous research has demonstrated high sequence homology between these two genes. I hypothesize that Cyp2a4 and Cyp2a5 gene expression is differentially regulated by transcriptional or post-transcriptional mechanisms. Data from chapter five expand the current knowledge about differential regulation of Cyp2a4 and Cyp2a5 gene expression, including developmental and zonal patterns, and identify long non-coding antisense RNAs (Cyp2a4as and Cyp2a5as) which mirror expression patterns of Cyp2a4 and Cyp2a5. I hypothesize that Cyp2a4as and Cyp2a5as positively regulate Cyp2a4 and Cyp2a5 gene expression.
Figure 1-1. Diagram of mammalian liver architecture. **Left:** The lobule structure in the adult liver shows repeating hexagonal lobular units. **Middle:** Each lobule converges to a central vein in the center. Each corner of the hexagonal lobule contains a portal vein, hepatic artery, and bile ducts, termed a portal triad. Plates of hepatocyte (usually 15-20 cells) extend outward from the center of the liver lobule. **Right:** The portocentral axis of the liver lobule. Blood enters the liver through the portal vein and hepatic artery (periportal region) and flows along sinusoids towards the central vein (pericentral region). Transfer of nutrients between blood and hepatocytes occurs in the space of Disse. Canaliculi transport bile produced from hepatocytes to the bile duct. Used with permission from (Spear et al., 2006).
Table 1.1. Sex-biased male and female cytochrome p450 gene expression in mouse liver.

<table>
<thead>
<tr>
<th>Female Biased</th>
<th>Zhx2^fl/fl Female/Male Ratio</th>
<th>Male Zhx2^ΔHep/Zhx2^fl/fl Ratio</th>
<th>Female Zhx2^ΔHep/Zhx2^fl/fl Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyp2b13</td>
<td>10,936</td>
<td>33.7</td>
<td>1</td>
</tr>
<tr>
<td>Cyp2a4</td>
<td>981</td>
<td>7.9</td>
<td>1</td>
</tr>
<tr>
<td>Cyp2b9</td>
<td>770</td>
<td>6.8</td>
<td>1</td>
</tr>
<tr>
<td>Cyp4a10</td>
<td>7.1</td>
<td>5.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Cyp39a1</td>
<td>1.5</td>
<td>4.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Cyp2a5</td>
<td>3.2</td>
<td>3.1</td>
<td>1.1</td>
</tr>
<tr>
<td>Cyp2c40</td>
<td>5.8</td>
<td>1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Cyp7a1</td>
<td>7</td>
<td>1.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Male Biased</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyp4a12</td>
<td>0.002</td>
<td>0.9</td>
<td>0.6</td>
</tr>
<tr>
<td>Cyp2d9</td>
<td>0.4</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Cyp8b1</td>
<td>0.6</td>
<td>1.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Unbiased</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyp3a25</td>
<td>0.6</td>
<td>1</td>
<td>1.1</td>
</tr>
<tr>
<td>Cyp2d10</td>
<td>0.6</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>Cyp3a16</td>
<td>0.8</td>
<td>2.1</td>
<td>1</td>
</tr>
</tbody>
</table>

Liver mRNA expression ratios of 14 Cyp genes examined in 8-week old Zhx2^fl/fl and Zhx2^ΔHep mice. The mRNA levels were first quantitated by RT-qPCR and normalized to ribosomal protein L30 (L30) mRNA levels. Ratios between mice were then calculated as indicated. Adapted from (Creasy et al., 2016).
**Figure 1-2. Growth hormone activation of Stat5b in hepatocytes.** Growth hormone (GH) binds to growth hormone receptor (GHR) on the hepatocyte cell membrane, which dimerizes and activates GHR-associated tyrosine kinase, Janus kinase 2 (Jak2). Jak2 phosphorylates tyrosines on the cytoplasmic domain of GHR at multiple sites, creating docking sites for Stat5 and other SH2 domain containing proteins. Previous studies have shown that Stat5 is the major Stat isoform that is activated by GH signaling in rat liver (Waxman, 1995). Stat5 binds to phosphorylated tyrosines on the GHR, and is itself phosphorylated on tyrosine 699, which is required for Stat5 DNA binding and transcriptional activity. Stat5 then undergoes dimerization and nuclear translocation to regulate transcription of target genes. Phosphotyrosine phosphatases activated by Stat5 deactivate Stat5 and Jak2/GHR. Stat5 can be reactivated by multiple rounds of tyrosine phosphorylation.
phosphorylation. In males, Stat5b is activated by pulsing of GH which occurs in short bursts. In females, continuous GH consistently activates phosphotyrosine phosphatases that dephosphorylate Stat5 and Jak2/GHR leading to differential regulation of sex-biased genes (Gebert, Park, & Waxman, 1999a). Figure adapted from (Waxman & O'Connor, 2006).
Figure 1-3. Simplified model for the regulation of sex-biased genes by Stat5b, Bcl6, and Cux2 in adult male and female liver. In male liver (left panel), Stat5b is intermittently activated by plasma GH pulses that facilitate Stat5b activation of male-biased genes, including Bcl6. Bcl6 represses many female biased genes, including the female-specific repressor of male-biased genes, Cux2. Cux2 is de-repressed in Stat5b-deficient male liver, indicating that pulsatile GH-activated Stat5b helps repress Cux2 expression in male liver. Continuous GH in females (right panel) results in persistent activation of Stat5b and continued dephosphorylation of Stat5 resulting in female-biased gene expression patterns. Cux2 is expressed in female liver and acts to repress male-biased gene activation. Figure adapted from (Lau-Corona, Suvorov, & Waxman, 2017).
Figure 1-4. Characterization of long non-coding RNAs throughout the genome.

LncRNAs can be classified by genomic location, relative to protein-coding genes, into six non-mutually exclusive categories: sense, antisense, intronic, divergent, intergenic (or stand-alone), and enhancer-associated. The above diagram shows these types of LncRNAs in the context of Gene A exons (pink squares, numbered). LncRNAs are depicted as multi-exon spliced transcripts (black boxes). Transcription direction is indicated by arrows from antisense or Gene A exon 1. A. Sense lncRNAs are transcribed from the same strand as the sense transcripts. B. Antisense RNA are transcribed opposite of protein-coding genes and can overlap regions on the 5’ end of genes, including the promoter (Promoter antisense); within the gene (Nested antisense); or on the 3’ end of the gene (Terminal antisense). Natural antisense tend to be enriched around the 5’ and 3’ ends of sense transcript. C. Intronic of protein-coding genes can harbor lncRNAs (Intronic
lncRNAs). D. LncRNAs can be produced from the vicinity of transcription start sites for protein-coding genes in both sense and antisense directions (Divergent lncRNAs). Shown is an antisense directed divergent transcript. E. LncRNAs can be transcribed in a region of the genome that does not overlap with any protein-coding gene (Intergenic lncRNAs). An intergenic lncRNA is shown between two genes, Gene A (pink box, 1) and Gene B (gray box, 1) which are transcribed in separate directions. F. Enhancer regions, located either upstream or downstream of their respective protein-coding gene, are genomic regulatory elements which can produce lncRNAs (Enhancer lncRNAs). Shown are two enhancer RNAs that are transcribed in the opposite direction within an enhancer region (yellow) that is upstream of gene A exon 1.
Figure 1-5. Mechanisms by which long non-coding RNAs (lncRNAs) can affect expression of protein coding genes. LncRNAs (purple lines) can influence transcription initiation by recruiting co-activators, co-repressors, or chromatin modifying proteins (black circles) by complementary lncRNA-DNA binding; these interactions can activate or repress transcription of protein-coding genes. LncRNAs can affect post-transcriptional processing of mRNAs (orange lines) by base-pairing with target mRNA and regulating alternative splicing through masking of splice sites. LncRNA-RNA interactions can initiate nuclear RNA-editing that can cause nuclear retention and RNA degradation of RNA transcripts by nucleases. LncRNA-RNA interactions in the cytoplasm can increase or decrease translational efficiency of mRNA, or increase mRNA stability by masking miRNA sites. LncRNA can be processed in the cytoplasm into miRNA that can target mRNA transcripts for degradation. Additionally, lncRNA act as competitors for miRNAs that might normally target mRNAs. All RNAs are shown as polyadenylated (AAAAAA)
and 5’ capped (red circle), although not all lncRNAs are have these modifications The Argonaute miRNA complex is shown as a yellow circle.
CHAPTER II

MATERIALS AND METHODS

Identification and Characterization of Zinc Finger and Homeobox Genes from Chordate Species

Zhx protein sequences were downloaded from the NCBI database for multiple and pairwise sequence alignments (Appendix Table 1). The rabbit Zhx2 sequence available from the NCBI database contains a gap in its sequence that spanned homeodomain 4 which resulted in the incorrect prediction of the carboxy-terminus of the protein; aligning the rabbit Zhx2 genomic sequence to pig predicted the correct end of the rabbit Zhx2 protein. The Petromyzon marinus (Sea Lamprey) Zhx-like (ZhxL) gene was found by BLAST analysis of the newly assembled SIMRbase/Lamprey Genome Browser for homologous regions to the mouse Zhx1, Zhx2, and Zhx3 genes (Smith et al., 2018). Zhx genes were manually annotated for zinc finger and homeodomain regions using Geneious software (Version 6.1.8). C2H2 Zinc finger domains were anchored starting with the first cysteine at amino acid (aa) position 1, the second cysteine at aa position 4, the first histidine at aa position 17, and the second histidine at aa position 22 for each zinc finger. Homeodomains were anchored by aa positions 48 and 49 (W48/F48/Y48 and F49/Y49/W49 within each homeodomain) that are the most highly conserved aa residues among previously studied homeodomain proteins (Burglin and Affolter, 2016). Each homeodomain was annotated as 60 a.a.’s in length. Furthermore, conserved residues
L16/F16/M16/A16 and R53/L53 within each homeodomain were minimal criteria in order to annotate a region as a predicted homeodomain.

Multiple Sequence Alignment and Phylogenetic Tree Building

Multiple sequence alignments were completed in Geneious 11.1.5 software using MUSCLE. Multiple sequence alignments of full length Zhx proteins, Zhx homeodomain regions (60 aa), and zinc finger regions (ZFR) (22 aa for each zinc finger and the 10 aa region between zinc fingers), were used to build phylogenetic trees across species.

Distance tree building was completed using Geneious Tree Builder with the Jukes-Cantor distance model and Neighbor Joining Tree building method. Lancelet Zhx full-length protein or lancelet ZFR was used for full length Zhx tree building and ZFR tree building, respectively. No outgroup was used for gnathostome Zhx tree building. Consensus trees were built using bootstrap resampling with 100 replicates implementing a greedy clustering algorithm and a support threshold of 50%.

For Bayesian tree building, MUSCLE MSA were exported from Geneious software as phylip alignment files (.phy) and tested for best-fit models of evolution using Prottest3.4.2 software. Prottest3.4.2 execution was completed using the University of Kentucky Dell Intel Xenon64 Linux Cluster II. MUSCLE MSA were then exported from Geneious software as nexus alignment files (.nex) and analyzed using MrBayes3.2.6 software. MrBayes execution was completed using the University of Kentucky Dell Intel Xenon64 Linux Cluster II. For Zhx full length proteins, lancelet was used as an outgroup, and among site variation was set to invgamma for likelihood parameters. For the
parameters of the phylogenetic model; we used a Jones Taylor Thornton rate matrix, with 1,000,000 generations, sample frequency of 200, and burnin fraction to 0.25. For HDs analysis, no outgroups were used and among site variation was set to gamma for likelihood parameters. For the parameters of the phylogenetic model; we used a Jones Taylor Thornton rate matrix, with 30,000,000 generations, sample frequency of 200, and burnin fraction to 0.25. Sumt and Sump commands were completed after runs for both Zhx full length and HD runs, and consensus trees were uploaded into FigTree1.4.4 for visualization. Compilation of MrBayes3.2.6 and Prottest3.4.2 to run on the cluster environment, including the incorporation of new Java script for Prottest3.4.2 compatibility, were completed by Vikram Gazula.

Sea Lamprey Tissue Collection

Lamprey were euthanized in MS222, tissues were dissected over ice, and snap frozen in liquid nitrogen immediately following dissection. Tissue collection was completed by Dr. Jeramiah Smith. Tissues were stored in -80°C until RNA or DNA isolation.

RNA Isolation for Mouse and Sea Lamprey Tissues

Tissues were weighed and homogenized using tube pestle and RNAzol®RT (1mL/100mg tissue) (Molecular Research Center, Inc. Catalog # RN 190). To precipitate DNA, protein, and polysaccharides, 0.4 mL of water per 1 mL of RNAzol®RT was added to the homogenate. The mixture was then shaken vigorously for 15 sec and stored
at room temperature for 15 min. Samples were then centrifuged at 12,000 g for 15 min at 4°C. 700 μL of supernatant was transferred to a new tube and RNA was precipitated by mixing with 700 μL of isopropanol. Samples were centrifuged at 12,000 g for 10 min at 4°C. Supernatant was removed after centrifugation and the RNA pellet was washed twice with 600 μL 75% ethanol (v/v). After each wash, pellets were centrifuged at 8,000 g for 3 min and supernatant was removed from pellet. RNA pellets were dried under vacuum for 15 min and solubilized using 30-100 μL of water depending on pellet size. RNA concentrations were measured using NanoDrop One (Thermo Scientific).

cDNA Synthesis and Reverse Transcription Quantitative Real-Time PCR (RT-qPCR)

1.0 μg of RNA prepared from tissues was reverse transcribed to cDNA using the High Capacity cDNA Reverse Transcription Kit (ThermoFisher Scientific, Catalog # 4368814). qPCR reactions were prepared with SsoAdvanced Universal SYBR Green Supermix (Bio-Rad, Catalog #172-5271) and amplified in a Bio-Rad CFX96 real-time PCR system. Oligonucleotides were obtained from Integrated DNA Technologies (Coralville, IA) and their sequences are shown in Table 2.1. All RT-qPCR runs were done in duplicate and included non-template controls, as well as RT negative controls where needed. Primer specificity was determined by melting curve analysis; a single peak at a temperature above 80°C indicated absence of primer dimer formation or non-specific product amplification. Additional analyses to resolve amplicons using agarose or polyacrylamide gel electrophoresis was completed for specific amplicons as indicated in the text. qPCR Ct values were normalized to either ribosomal protein L30 (L30) or Serine
and Arginine Rich Splicing Factor 4 (SFRS4) and reported as normalized expression of the indicated gene using the ΔCt method unless otherwise noted. L30 and SFRS4 mRNA levels are stable through mouse liver development and in adult liver, whereas SFRS4 mRNA levels are stable in our mouse model of liver cancer and during liver regeneration. In all cases, primers were designed to span introns to circumvent amplification of products from potential contaminating genomic DNA.

**Gel Electrophoresis and Gel Visualization**

Amplicons from RT-qPCR, RT-PCR, or PCR were resolved using 1.0% or 0.7% agarose gel electrophoresis. Agarose gels, except those presented in Figure 5-9, contained ethidium bromide in the gels and were run using 1X TBE Buffer conditions. Agarose gels presented in Figure 5-9 were stained with 0.1% ethidium bromide for 30 min following gel electrophoresis. Following staining, gels were washed twice with deionized water for 15 min while rocking at room temperature. Additionally, for increased resolution, some amplicons were resolved using 6.0% acrylamide vertical gel electrophoresis at 4°C. Acrylamide gels were stained with ethidium bromide for 15 min following gel electrophoresis and washed twice with deionized water for 15 min while rocking at room temperature. All gels were viewed using an Amersham Imager 600 transilluminator using automatic high resolution settings (GE Healthcare Life Sciences). The GeneRuler 100 bp plus ladder (Thermo Scientific, catalog # SM0321) was used as the size marker on the gels.
DNA Isolation for Mouse and Sea Lamprey Tissues

DNA was extracted by tissue incubation in lysis buffer (100 mM Tris-HCl pH 8.5, 5mM EDTA, .2% SDS, 200 mM NaCl) and 2.5µL Proteinase K (20 mg /mL stock). Samples were lysed by incubation overnight in a 55°C water bath. Following lysis, samples were vortexed to ensure that lysis was complete and centrifuged at 13,000 g for 12 minutes at room temperature. Supernatants were transferred to separate tubes containing 500 µL of isopropanol. Tubes were inverted and DNA was precipitated by centrifugation at 13,000 g for 2 minutes. Supernatant was then decanted and excess liquid was drained by inversion. DNA pellets were washed with 500 µL of 75 % ethanol and supernatant was decanted and the pellet was allowed to dry for 30 minutes on the lab benchtop. DNA pellets were resuspended in 50-250 µL of water depending on pellet size.

Mice

All mice were housed in the University of Kentucky Division of Laboratory Animal Research (DLAR) facility according to Institutional Animal Care and Use Committee (IACUC) approved protocols. All mice had *ad libitum* access to food and water and were maintained on a 14/10 light/dark cycle. Six hours prior to sacrifice, all mice were restricted to water only and all mice were sacrificed between 12-4 p.m. to limit circadian rhythm variances in gene expression.
Breeding of Zhx2 Mice and Genotyping

Breeding pairs of C57BL/6 mice with a Zhx2 floxed allele \((\text{Zhx2}^{\text{fl/fl}})\) were purchased from the Knockout Mouse Project Repository at the University of California-Davis. In these mice, exon 3 has been floxed by LoxP sites (Creasy et al., 2016). Since the entire coding region for Zhx2 is found in exon 3, deletion of this exon results in loss of the entire protein. These mice were crossed with C57BL/6 mice expressing cre recombinase driven by the liver-specific \textit{Albumin} promoter (Alb-Cre) (Jackson Labs cat #016832) Mice were bred to achieve homozygous floxed \textit{Zhx2} alleles with Alb-Cre expression \((\text{Zhx2}^\Delta\text{Hep})\) or without Alb-Cre \((\text{Zhx2}^{\text{fl/fl}})\) serving as littermate controls. At approximately 10-14 days of age, each mouse pup was given an ear punch for identification and ear tissue was collected into a correspondingly labeled tube. DNA was extracted (as previously described) and genotyped by PCR amplification using Thermo Scientific DreamTaq PCR Master Mix \((\#K1071)\) and two primer pairs; 1651/1654, which distinguish floxed or wildtype \(\text{Zhx2}\) alleles, and 1354/1355 that amplifies the Cre recombinase, if present (Table 2.2). PCR amplification was completed in the SimpliAmp Therma Cycler (Applied Biosystems) using primers \((0.4 \mu\text{M})\), \(2 \mu\text{L}\) of diluted genomic DNA, and PCR conditions listed in Table 2.3.

Euthanasia and Mouse Tissue Collection

C57BL/6 \textit{Zhx2}^{\text{fl/fl}} and \textit{Zhx2}^\Delta\text{Hep} littermates were euthanized by \text{CO}_2 asphyxiation at 8-weeks of age. Livers were dissected and immediately snap frozen in a 100% ethanol and dry ice bath, and stored at -80\(^\circ\text{C}\) until RNA isolation.
Developmental Timepoint Mice

Female C3H/HeJ (C3H) mice were bred with C57BL/6J (Bl/6) mice and female mice were monitored for vaginal plugs to estimate the time of fertilization. Pregnant females were euthanized by CO2 asphyxiation at 17.5 days post-conception and the amniotic sac was removed containing the mouse embryo. Neonatal and perinatal pups were euthanized by decapitation, and pups aged by more than 14 days were euthanized by CO2 asphyxiation. All pups were dissected and livers were isolated, frozen on dry ice, and stored at -80°C until RNA isolation. Tissue samples were prepped for RNA as previously described and RNA was stored in -80°C. Drs. Kate Townsend Creasy and Jieyun Jiang completed mouse dissections and RNA isolation. For my analysis, RNA stored at -80°C was used for cDNA synthesis as previously described.

Alignment of Gene Sequences and JASPAR Analysis

C4 (C4b), Slp (C4a), Cyp2a4, and Cyp2a5 gene sequences were downloaded from the mm10 assembly from the UCSC Genome Browser. C4, Slp, Cyp2a4, Cyp2a5, Cyp2a4as, and Cyp2a5as sequences, were aligned using Geneious 11.1.5 pairwise alignment with free end gaps and a 65% cost matrix. Sequences 2000 base pairs or 1500 base pairs upstream from the transcriptional start site were used for JASPAR CORE (Khan et al., 2018) Mus musculus transcription factor (TF) binding prediction analysis (176 transcription factors total), or were selected specifically for Stat5a:Stat5b predicted binding only (Slp promoter analysis). Files were exported, and only TF’s with a score of 7.5 and Pearson correlation coefficient greater than 0.8 were considered in order to
ensure a higher probability of TF binding. TF’s meeting the score criteria were annotated on aligned sequences in Geneious to find TF’s that were predicted to differentially bind because of nucleotide differences between aligned sequences.

**Slp Plasmid Construction**

*Slp* 1500 promoter primers (2185 and 2186; table 2.1) and *Slp* 2000 promoter primers (2227 and 2186) were used to amplify *Slp* promoter regions from C57BL/6 liver genomic DNA using Thermo Scientific DreamTaq PCR Master Mix (catalog #K1071). Both 2185 and 2227 primers contained XhoI restriction enzyme sites whereas 2186 contained a HindIII restriction enzyme site for later cloning into pGL4.14. The amplicons were separated on 0.7% agarose gels and DNA was extracted using the Wizard®SV Gel and PCR Clean-Up System (Promega, catalog # A9281). DNA was eluted in 35 µL of sterile water and DNA concentration was measured using Nanodrop One (Thermo Scientific). Amplicons were ligated overnight into the pGEM®-T Easy Vector. Ligation reactions were transformed into DH5α *E. coli* and grown on plates containing both ampicillin and X-gal. White colonies were selected for PCR colony screening using pGEM-T Easy Vector primers (386 and 387) which bind to the vector sequence and amplify the region of the plasmid within the multiple cloning site and product insert. Colonies that were positive by PCR were selected and grown overnight in LB broth containing 100 µg/ml ampicillin. 1.0 mL of bacterial culture from each isolated colony was used for Miniprep plasmid isolation (Thermo Scientific GeneJET Plasmid Miniprep Kit, catalog #K0502). Plasmid DNA was eluted into 50 µL of sterile water and sequenced.
using both forward and reverse DNA strands and 386 or 387 primers (ACGT Inc., Wheeling, IL). pGEM-T Easy plasmids containing the desired insert and the pGL4.14 [luc2/Hygro] vector (Promega, catalog # E699A) were digested with XhoI and HindIII and the products were separated on 0.7% agarose gels. pGL4.14-digested backbone and Slp 2000 and 1500 promoter fragments were isolated using Wizard® SV Gel and PCR Clean-Up System (Promega, catalog # A9281), ligated overnight with T4 ligase at 4°C (New England Biolabs, catalog # M0202S), and transformed into DH5α E. coli. Colonies were selected by ampicillin and checked by PCR colony screening using 2185 and 2186, or 2227 and 2186, primers. All primers used for cloning within this section are listed in Table 2.2.

Stat5b and Stat5b* Plasmid Construction

Stat5b pME18S (Stat5b) and Stat5b pCI (Stat5b*) were kindly provided by Dr. David Waxman (Boston University, Boston, MA). Plasmids were transformed into DH5α E. coli and selected with ampicillin. 1.0 mL of bacterial culture from each isolated colony was used for Miniprep plasmid isolation (Thermo Scientific GeneJET Plasmid Miniprep Kit, catalog #K0502). Stat5b* and Stat5b plasmids were then digested with EcoRI and NotI, and digested products were loaded into a 0.7% agarose gel and resolved by gel electrophoresis. The Stat5b* and pME18S bands were isolated from the gel and DNA was extracted from gel slices using the Wizard® SV Gel and PCR Clean-Up System (Promega, catalog # A9281). DNA was eluted in 35 µL of sterile water and DNA concentration was measured using Nanodrop One (Thermo Scientific). Stat5b* was
ligated into pME18S overnight with T4 ligase at 4°C (New England Biolabs, catalog # M0202S), and transformed into DH5α E. coli. Colonies were selected by ampicillin. For colony PCR, primers were designed against pME18S (2308 and 2313) and used to check for Stat5b* inserts. Colonies that were positive by PCR colony screening were selected and grown up overnight in LB broth containing ampicillin. 1.0 mL of bacterial culture from each isolated colony was used for Miniprep plasmid isolation (Thermo Scientific GeneJET Plasmid Miniprep Kit, catalog #K0502). Both Stat5b and Stat5b* pME18S plasmids were eluted into 50 µL of sterile water and first sequenced using the primer 2310. Primer 2310 binds within Stat5b and was used for forward sequencing to check for H299R mutation (CAC> CGC) mutation; the 3’ end of the 2310 primer lies 71 bp upstream of this mutation site. A second round of sequencing using Primer 2312, confirmed the second site specific mutation needed to confer constitutive Stat5b activity for pME18S Stat5b*. Primer 2312 was used for forward sequencing to check for the S711F mutation (TCC>TTC). All primers used for cloning within this section are listed in Table 2.2.

Midiprep of Plasmids for Luciferase Assay

The 4XNTCP pT109 Luciferase plasmid was kindly provided by Dr. Young-Kwon Hong (University of Southern California, Los Angeles, CA)(Ganguly et al., 1997). Zhx2-HA pcDNA 3.1 (-) (Jiang et al., 2017), rat HNF4α pMT7 from Dr. Frances Sladek (UC Riverside), HSV-thymidine kinase promoter Renilla luciferase control plasmid (Renilla, Promega, #E2241), pGL4.14 (empty luciferase vector), Slp 1500 pGL4.14, Slp
2000 pgl4.14, Stat5b, and Stat5b* plasmids were transformed into DH5α E. coli and selected using ampicillin. Following selection, single colonies from each plate were inoculated into a 1.0 mL LB broth starter culture containing 100 µg/mL ampicillin and grown for 8 hours shaking at 37°C. 50 µL of starter culture was then transferred into a 50 mL LB broth culture containing 100 µg/mL ampicillin and grown overnight (16 hours) shaking at 37°C. Cells were harvested by centrifugation at 5,000 g for 10 min. Supernatant was discarded and plasmid DNA was purified using the Thermo Scientific GeneJET Plasmid Midiprep Kit (Catalog # K0481). Plasmids were eluted into 350 µL of sterile water and DNA concentrations were measured using the NanoDrop One (Thermo Scientific).

**Transient DNA Transfections**

Hepa 1.6 hepatoma cells (derived from a C57BL/6 hepatoma) were grown in T25 flasks (Corning) in Dulbecco’s minimal eagle’s media (DMEM, Sigma, catalog # D6421) supplemented with 10% fetal bovine serum (FBS, Gibco). The cells were maintained at 37°C and 5 % CO2; cells were split using 0.5 mL of trypsin every few days and plated at a 1:5 ratio to prevent confluency. For luciferase assays, cells were counted using a hemocytometer and 1.0 × 10⁵ cells were plated 24 hours in advance of transfection experiments.

Hepa 1.6 cells were seeded at 1.0 × 10⁵ cells/ well in 24-well plates containing 500 µL of 10 % FBS/ DMEM media for 24 hours prior to transfection. For Slp single driver transfections, serum free DMEM media (50 µL), 10 ng (Renilla), 175 ng
(pGL4.14, Slp 1500 pGL4.14, Slp 2000 pGL4.14), and 350 ng (pcDNA 3.1, Zhx2-HA pcDNA3.1) was added together and gently mixed. 2 µL Turbofect reagent (Thermo Scientific, catalog #R0532) was added to the serum free DMEM media and DNA mixture, gently mixed, and incubated at room temperature for 20 min. Each mixture was then added in a dropwise fashion to designated wells and gently swirled to mix. Forty-eight hours after transfection, cells were harvested for Dual Luciferase Reporter assay.

For NTCP and Slp single driver transfections, 10 ng (Renilla), 100 ng (4XNTCP pT109 Luciferase plasmid), 200 ng (pcDNA 3.1, Zhx2-HA pcDNA3.1, Stat5b, Stat5b*, or HNF4α pMT7), and another 200 ng (pcDNA3.1) was added to serum free DMEM media and the protocol was followed as described above. For NTCP and Slp double driver transfections, 10 ng (Renilla), 100 ng (4XNTCP pT109 Luciferase plasmid), 200 ng (pcDNA 3.1, Zhx2-HA pcDNA3.1, Stat5b, Stat5b*, or HNF4α pMT7) was added to serum free DMEM media and the same transfection protocol was followed.

Dual Luciferase Reporter Assay

Cells were harvested 48 hours post transfection and washed with sterile 1X phosphate buffered saline (PBS), followed by lysis with 1X Passive Lysis Buffer (100 µL, Promega, catalog # E1910) for 15 min at room temperature while rocking. Cell lysates (20 µL) were placed in duplicate wells of a 96-well luciferase plate (Falcon, catalog #353296). Luciferase Assay Reagent II (LAR II) was prepared by resuspending lyophilized Luciferase Assay Substrate in 10 mL of supplied Luciferase Assay Buffer II. One volume of Stop and Glo substrate was added to fifty volumes of Stop and Glo Buffer
and mixed by inversion. Dual luciferase analysis was performed using the GloMax Explorer (Promega). Both LarII and Stop and Glo were injected sequentially (100 µL) into each well. Files were exported into Microsoft Excel, and Luciferase values were first normalized to Renilla values. The number of experiments is indicated in legends below each figure.

**Bioinformatics for Zhx2, Stat5b, and HNF4α Overlap**

Supplementary Excel spreadsheets containing microarray data were downloaded and visualized with Microsoft Excel for Zhx2 (Gargalovic et al., 2010), Stat5b (Clodfelter et al., 2006), and HNF4α (Holloway et al., 2008). Unigene IDs for each microarray were found by uploading all accession numbers from each microarray into Database to Database conversions software (db2db-BioNet). Sex-biased genes were first identified within Stat5b or HNF4α microarrays by male to female wildtype expression ratios > 2, and p-value < 0.005 using VLOOKUP, and IF(AND) functions. Common sex-biased genes between both microarrays were identified using VLOOKUP function. Using the above criteria, 116 common sex-biased genes were identified in total, with 105 genes that matched for sex-bias between both datasets. There were 11 genes that did not match either with sex-bias (in one dataset the gene was reported as female-biased, and in the other it was reported as male-biased), or contained other probes that did not show sex-biased expression, and as a result, these genes were not counted within the common sex-biased genes. Common sex-biased genes were then applied to Stat5b, HNF4a, and Zhx2 datasets to filter for sex-biased and sex-independent genes from all three microarrays.
Sex-biased and sex-independent genes were then filtered for differential expression within each microarray using stringent (FC>2, p-value<.005) or relaxed (FC>2, p-value<.05) criteria using IF(AND) functions. Overlapping differentially expressed genes were found between datasets using VLOOKUP function.

Liver Perfusion and Preparation of Hepatocytes for Flow Activated Cell Sorting

E3-β-gl-Dd mice have been described elsewhere (Ramesh et al., 1995). Briefly, these mice contain the mouse H-2Dd structural gene fused to the human β-globin promoter. Directly upstream of the β-globin promoter lies alpha-fetoprotein enhancer 3. This enhancer confers highly restricted expression of the H-2Dd structural gene to hepatocytes directly surrounding the central vein.

Mouse liver perfusions were completed by Dr. Jieyun Jiang. Mice were anesthetized with intraperitoneal injection of xylazine (50 mg/kg) and ketamine (100 mg/kg) and surgery was not initiated until breathing slowed and loss of limb sensitivity was evident. Livers were perfused by the portal vein using 40 mL liver perfusion media (Table 2.4) with a peristaltic pump (P-1, Pharmacia) set to level 5. Briefly after portal vein entry and flow of perfusion media, the inferior vena cava was clipped to allow for drainage of blood from liver. After 4 minutes, liver perfusion media was changed to liver digest media (Table 2.5) containing collagenase (1mg/mL). Collagenase was added to liver digest media immediately before the start of perfusion surgery. After 5 minutes, the peristaltic pump was turned off and the inserted tube was removed from the portal vein. The liver was removed from the abdominal cavity and placed onto cold serum free
DMEM media on ice. The remainder of steps for hepatocyte isolation were performed within a laminar flow hood. The membrane was removed from the liver and cells were released into the media by gently swirling back and forth to release all cells. The connective tissue and membrane was then discarded. Cells were then gently pipetted into a 50 mL conical tube (BD Falcon) and centrifuged at 50 g for 2 min at 4°C. Cells were placed on ice following centrifugation and all but 2 mL of supernatant was discarded. Cells were resuspended in 10 mL cold serum free DMEM media. Percoll gradient was prepared using 9 mL Percoll stock and 1 mL sterile 10X PBS. 8 mL of the Percoll mix was added to 12 mL of hepatocytes in serum free DMEM media and gently pipetted up and down to mix thoroughly. Cells were centrifuged at 200 g for 15 min at 4°C. Supernatant was removed from cells, replaced with 20 mL serum free DMEM media, and cells were gently resuspended by pipetting. Cells were then centrifuged at 50 g for 2 min at 4°C. Supernatant was removed after following centrifugation and cells were gently resuspended by circular motion of the into 10 mL of 1% FBS in sterile 1X PBS. Cells were then gently resuspended by pipetting.

To check viability, 90 μL of cells and 10 μL of Trypan blue was mixed in a 96-well culture plate well, and 10 μL of this mix was added to a hemocytometer for counting. The ratio of live cells (cells that exclude blue) minus dead cells (cells that stain blue) over total cells counted was used to calculate the viability. All perfusions that were performed had 95-98% viability.

The concentration of cells per mL was then calculated, and for every 1.0 × 10^6 cells or 1.0 mL of cells, 20 μL of H2-Dd-FITC antibody (BD Biosciences, cat #
553579C) was added to allow for maximal antibody binding. Cells were gently mixed by agitation and placed on ice for 1 hour in dark, and mixed every 30 minutes to ensure antibody binding. As a control, cells were incubated with IgG antibody. Following 1 hour incubation, FITC antibody-treated cells and control cells were centrifuged at 50 g for 2 min at 4°C. Supernatant containing unbound antibody was removed and pellets were placed back on ice. Cells were gently rinsed with 1.0 mL 10% FBS 1% sodium azide (10% sodium azide (w:v) in 1X PBS solution) 1X PBS followed by centrifugation at 50 g for 2 min at 4°C and supernatant removal, three times. Following last centrifugation, supernatant was removed and cells were filtered through a 70 µM cell strainer (Fisher Scientific, cat # 22363548) and filtered cells were collected into 5.0 mL tubes and centrifuged at 50 g for 2 min at 4°C. Cells were placed on ice and supernatant was removed. 2.0 mL sorting buffer (1X PBS, 5 mM EDTA, 25 mM HEPES (pH 7.0), 1% FBS, 1% sodium azide) was added to cells treated with antibody and control cells. Cells were sorted using the Becton-Dickinson FACSCalibur flow cytometer and analyzed by BD CellQuest™ Pro software (San Jose, CA). Total RNA was isolated from unsorted, Dd negative, and Dd positive cells. For my analysis, I helped with the preparation of cells following perfusion, and performed cDNA synthesis and RT-qPCR on isolated RNA.

**Liver Regeneration**

Mouse liver regeneration was induced by a single intraperitoneal injection of carbon tetrachloride (CCl₄). 8-month old C3H males were administered either mineral oil or 0.05 mL 10 % CCl₄ diluted in MO. After 3 days, animals were euthanized by CO₂
asphyxiation, mice were dissected and livers were removed, and RNA was isolated for downstream RT-qPCR analysis. Regeneration experiments were previously performed by Hui Ren. For my analysis, I used RNA stored at -80°C for cDNA synthesis (as described earlier).

**DEN Cancer**

14-day old male Znx2fl/fl mice were injected with diethylnitrosamine (DEN) or PBS at a dose of 10 µL/g body weight. Mice were weaned at 21 days of age and maintained under normal conditions for 34 weeks. Mice were euthanized by CO2 asphyxiation and examined for HCC tumor development. DEN mouse experiments were performed by Dr. Jieyun Jiang. For these studies, I helped collect liver tissue and isolated RNA from all liver tissues used. RNA was stored in -80°C until cDNA synthesis.

**5' Rapid Amplification of cDNA Ends (5' RACE)**

5’ RACE was performed using the 5'/3'RACE Kit 2nd Generation (Roche, catalog # 03353621001). Oligonucleotides for 5’ RACE reactions were obtained from Integrated DNA Technologies (Coralville, IA) and their sequences are shown in Table 2.6. For the first set of 5’ RACE reactions presented in Figure 5-10A and B, cDNA was synthesized from 8-week old female whole liver RNA using Transcriptor Reverse Transcriptase, dNTP mix, cDNA synthesis buffer, sterile ddH2O and P7 primer (12.5 µM). Contents were mixed and spun down briefly and incubated at 55°C for 60 min, followed by 85°C
for 5 min. Following reaction, contents were mixed and spun down briefly. cDNA was isolated using the Wizard® SV Gel and PCR Clean-Up System (Promega, catalog # A9281). Briefly, an equal volume of Membrane Binding Solution was added to the RT reaction volume and filtered through a minicolumn by centrifugation at 16,000 g for 1 min at 4°C. Flow through was discarded and 700 µL of Membrane Wash solution was added to the minicolumn and centrifuged at 16,000 g for 1 min at 4°C. Flow through was discarded and 500 µL of Membrane Wash solution was added to the minicolumn and centrifuged at 16,000 g for 5 min at 4°C. Flow through was discarded and the minicolumn was transferred to a clean microcentrifuge tube. 35 µL of ddH2O was added to the minicolumn and incubated at room temperature for 1 min and cDNA was collected by centrifugation at 16,000 g for 1 min at 4°C. Poly A-tailing of cDNA was completed using 19 µL of cDNA, 2.5 µL of 10X concentrated Reaction Buffer, and 2.5 µL of ATP (2.0 mM). Contents were mixed, spun briefly, and incubated at 94°C for 3 min. Following this incubation, the tube was chilled on ice for 5 min, and contents were then mixed and spun down briefly. 1.0 µL of terminal deoxynucleotidyl transferase was added and the contents incubated at 37°C for 30 min, followed by 70°C for 10 min. Contents were mixed, spun down briefly, and placed on ice. For PCR amplification, 1/20 diluted dA-tailed cDNA (2.0 µL), Oligo-dT-Anchor Primer (0.4 µL), Thermo Scientific DreamTaq PCR Master Mix (#K1071 (25 µL), ddH2O (21.6 µL), and specific primers (P8 or P6) were used. The PCR conditions for this reaction are presented in Table 2.7. Following this PCR, contents were mixed, briefly spun down, and a portion of the first PCR reaction was diluted 1/20 and added into a nested PCR. For nested PCR, 1.0 µL 1/20 diluted cDNA, PCR Anchor Primer (0.4 µL), Thermo Scientific DreamTaq PCR Master Mix (#K1071 (12.5 µL),
ddH₂O (10.6 µL), and specific primers (P9 or P10) were used. The PCR conditions for this reaction are presented in Table 2.7. Contents of nested PCR reaction were then resolved on a 1.0 % agarose gel.

For the second set of 5′RACE reactions presented in Figure 5-10C and D, cDNA was synthesized from the same 8-week old female whole liver RNA using Transcriptor Reverse Transcriptase, dNTP mix, cDNA synthesis buffer, sterile ddH₂O and P6 or P9 primer (12.5 µM). cDNA synthesis, cDNA isolation, and poly-A-tailing was completed as described before. For Cyp2a4as PCR amplification, dA-tailed cDNA (2.0 µL), Oligo-dT-Anchor Primer (0.4 µL), Thermo Scientific DreamTaq PCR Master Mix (#K1071 (12.5 µL), ddH₂O (21.6 µL), and specific primer P11 (1.0 µL) were used. The PCR conditions for this reaction are presented in Table 2.8. Previous to the annealing temperature presented in Table 2.8, other annealing temperatures (57, 58, 60, 65, 67, 70°C) had been tried and failed to produce specific product. For Cyp2a5as PCR amplification, dA-tailed cDNA (5.0 µL), Oligo-dT-Anchor Primer (1.0 µL), Q5 High Fidelity Polymerase (0.5 µL), 5X Q5 Reaction Buffer (10 µL), 10 mM dNTP mix (1.0 µL), 5X Q5 GC Enhancer (10 µL), ddH₂O (20.5 µL), and specific primer P10 (2.0 µL) were used. The PCR conditions for this reaction are presented in Table 2.9. For Cyp2a4as nested PCR, 1.0 µL 1/20 diluted cDNA, PCR Anchor Primer (0.4 µL), Thermo Scientific DreamTaq PCR Master Mix (#K1071 (12.5 µL), ddH₂O (10.6 µL), and specific primer P12 (0.5 µL) were used. The PCR conditions for this reaction are presented in Table 2.10. For Cyp2a5as nested PCR, 1.0 µL 1/20 diluted cDNA or undiluted cDNA, PCR Anchor Primer (1.0 µL), Q5 High Fidelity Polymerase (0.25 µL), 5X Q5 Reaction Buffer (5.0
µL), 10 mM dNTP mix (0.5 µL), 5X Q5 GC Enhancer (5.0 µL), ddH₂O (11.25 µL), and specific primers P13 or P14 were used. The PCR conditions for this reaction are presented in Table 2.11. Positive colonies were selected and screened, and plasmids were isolated as previously described. Cyp2a4as plasmids were sequenced using primer 386, whereas Cyp2a5as plasmids were sequenced using T7 or SP6 primers provided by ACGT sequencing services (ACGT Inc., Wheeling, IL).

3’ RACE

3’ RACE was performed using the 5’/3’RACE Kit 2nd Generation (Roche, catalog # 03353621001). Oligonucleotides for 3’ RACE reactions were obtained from Integrated DNA Technologies (Coralville, IA) and their sequences are shown in Table 2.6. cDNA was synthesized from the same 8-week old female whole liver RNA using Transcriptor Reverse Transcriptase (1.0 µL), dNTP mix (2.0 µL), cDNA synthesis buffer (4.0 µL), sterile ddH₂O (10.99 µL) and oligo-Dt-Anchor Primer (1.0 µL). Following cDNA synthesis, PCRs were completed using cDNA (1.0 µL), PCR Anchor Primer (0.4 µL), Thermo Scientific DreamTaq PCR Master Mix (#K1071 (12.5 µL), ddH₂O (10.6 µL), and specific primer P16, P17, P18, P5, P19 or P20 (0.5 µL). The PCR conditions for this reaction are presented in Table 2.12. For nested PCR, 1/10 diluted P18 or P5 PCR cDNA reactions (1.0 µL), PCR Anchor Primer (0.4 µL), Thermo Scientific DreamTaq PCR Master Mix (#K1071 (12.5 µL), ddH₂O (10.6 µL), and specific primer P19 or P20 (0.5 µL) were used. The PCR conditions for this reaction are presented in Table 2.13.
Fragments from second nested PCR were cloned and sequenced using primer 387 (ACGT Inc., Wheeling, IL).

RT-PCR

RNA was reverse transcribed as described previously. Primers P8 and P15, or P6 and P15, were used to amplify Cyp2a4as and Cyp2a5as from male and female nuclear and cytoplasmic RNA (nuclear and cytoplasmic fractionation protocol below). For PCR reactions 1.0 µL of cDNA, Q5 High Fidelity Polymerase (1.0 µL), 5X Q5 Reaction Buffer (10 µL), 10 mM dNTP mix (1.0 µL), 5X Q5 GC Enhancer (10 µL), ddH2O (21 µL), specific primers (2.5 µL) were used. The PCR conditions for these reactions are presented in Table 2.14.

Nuclear and Cytoplasmic Fractionation

6-7 week old C57BL/6 male and female mice were sacrificed and their livers were removed. Livers were immediately immersed into Buffer I (0.25 M sucrose, 60 mM KCl, 15 mM NaCl, 15 mM Hepes pH 7.4, 2 mM EDTA, 0.5 mM EGTA, 0.15 mM spermine, 0.5 mM spermidine, 14 mM β-mercaptoethanol). Using a pre-chilled Dounce homogenizer, liver was homogenized in Buffer I over ice. 10 mL of Buffer II (0.75 M sucrose, 60 mM KCl, 15 mM NaCl, 15 mM Hepes pH 7.4, 2 mM EDTA, 0.5 mM EGTA, 0.15 mM spermine, 0.5 mM spermidine, 14 mM β-mercaptoethanol) was added into a 30 mL tube (Corex) and homogenate was layered over Buffer II. Tube was then spun at
1000 g for 10 minutes at 4°C. The cytoplasmic fraction in the top supernatant was removed to a new tube and placed on ice. Residual liquid and the nuclei pellet was resuspended in 3.2 mL of Buffer III (2.0 M sucrose, 60 mM KCl, 15 mM NaCl, 15 mM Hepes pH 7.4, 0.1 mM EDTA, 0.1 mM EGTA, 0.15 mM spermine, 0.5 mM spermidine, 14 mM β-mercaptoethanol). The resuspended nuclei pellet was then layered over 1.0 mL of Buffer III and spun at 36,000 rpm for 1 hour at 4°C. Supernatant was decanted off of the nuclei pellet. 300 µL of Trizol reagent was used to resuspend and lyse the nuclei pellet. Another 500 µL of Trizol reagent was added and nuclei were further mixed. Nuclei and cytoplasmic tubes were then stored in -80°C until RNA extraction. To extract RNA, TRIzol® reagent (Life Technologies, catalog #15596-026) was used. These experiments were performed by Dr. Guofong Qui. For my experiments, I generated cDNA from stored RNA.

Cloning and Sequencing

Amplicons were cut from the gel and isolated using the Wizard®SV Gel and PCR Clean-Up System (Promega, catalog # A9281) as described earlier. Fragments were ligated into pGEM®-T Easy Vector and transformed into DH5α E.coli. Positive colonies were selected and screened and plasmids were isolated as previously described. Plasmids were sequenced using primer 387 (ACGT Inc., Wheeling, IL).
Statistics

All values within a group were averaged and plotted as mean with standard deviation. P-values were calculated between two groups using student’s t-test and between three or more groups by One-Way or Two-Way ANOVA followed by Tukey’s test. p-values ≤ 0.05 were considered significant. Data was graphed and statistical tests were performed in GraphPad PRISM 7d Mac OS X software.
Table 2.1. Oligos for RT-qPCR.

<table>
<thead>
<tr>
<th>Primer Number</th>
<th>Gene</th>
<th>Primers</th>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. Smith Lab</td>
<td>EF1A</td>
<td>F</td>
<td>Sea Lamprey</td>
<td>GTCTACAAAATTTGGCGGTATT</td>
</tr>
<tr>
<td>J. Smith Lab</td>
<td>EF1A</td>
<td>R</td>
<td>Sea Lamprey</td>
<td>ACATCCTTGACAGACACGTT</td>
</tr>
<tr>
<td>1859</td>
<td>Zhx</td>
<td>F</td>
<td>Sea Lamprey</td>
<td>CGGAAAACACACGAGCAGTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AAC</td>
</tr>
<tr>
<td>1861</td>
<td>Zhx</td>
<td>R</td>
<td>Sea Lamprey</td>
<td>TTGGAACGACTGCACACAGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GA</td>
</tr>
<tr>
<td>1522</td>
<td>L30</td>
<td>F</td>
<td>Mouse</td>
<td>ATGGTGGCGCAAGGAAGAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GAA</td>
</tr>
<tr>
<td>1523</td>
<td>L30</td>
<td>R</td>
<td>Mouse</td>
<td>CCTCAAAGCTGGACAGTTG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GGCA</td>
</tr>
<tr>
<td>1588</td>
<td>Zhx2</td>
<td>F</td>
<td>Mouse</td>
<td>AGGCGGGACAAGCCTAGACA</td>
</tr>
<tr>
<td>1589</td>
<td>Zhx2</td>
<td>R</td>
<td>Mouse</td>
<td>TGAGGTTGGCCCAGCCACT</td>
</tr>
<tr>
<td>2157</td>
<td>Slp</td>
<td>F</td>
<td>Mouse</td>
<td>GGCTCTTTCATTCTACTTCG</td>
</tr>
<tr>
<td>2158</td>
<td>Slp</td>
<td>R</td>
<td>Mouse</td>
<td>CCCATTATGATACTCCTTGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2217</td>
<td>C4</td>
<td>F</td>
<td>Mouse</td>
<td>GCTGTGTCTAAGATTCTCCAA ATTG</td>
</tr>
<tr>
<td>2218</td>
<td>C4</td>
<td>R</td>
<td>Mouse</td>
<td>CAGAACCATAGTCTCCAAGG</td>
</tr>
<tr>
<td>2411</td>
<td>Stat5b</td>
<td>F</td>
<td>Mouse</td>
<td>CAGACCAAGTTTGAGCCAC</td>
</tr>
<tr>
<td>2412</td>
<td>Stat5b</td>
<td>R</td>
<td>Mouse</td>
<td>GGATCTGCGGCTGATTCATCA</td>
</tr>
<tr>
<td>2446</td>
<td>Stat5a</td>
<td>F</td>
<td>Mouse</td>
<td>CAATGCATCCACAGATGCCG</td>
</tr>
<tr>
<td>2447</td>
<td>Stat5a</td>
<td>R</td>
<td>Mouse</td>
<td>TCATCCAGGTCAAAACCTGCCCC</td>
</tr>
<tr>
<td>2101</td>
<td>HNF4α</td>
<td>F</td>
<td>Mouse</td>
<td>GGAAGCTGTCCAAAATGAGCG</td>
</tr>
<tr>
<td>2102</td>
<td>HNF4α</td>
<td>R</td>
<td>Mouse</td>
<td>ATGTCGCCATTGATCCCAGAG</td>
</tr>
<tr>
<td>P903</td>
<td>Bcl6</td>
<td>F</td>
<td>Mouse</td>
<td>CCCACAGCATACAGAGATGTG</td>
</tr>
<tr>
<td>P904</td>
<td>Bcl6</td>
<td>R</td>
<td>Mouse</td>
<td>GAAGAAGGTCCCATTCTCACA</td>
</tr>
<tr>
<td>P905</td>
<td>Cux2</td>
<td>F</td>
<td>Mouse</td>
<td>GGACATGAAGAAGCCTGAGAGA</td>
</tr>
<tr>
<td>P906</td>
<td>Cux2</td>
<td>R</td>
<td>Mouse</td>
<td>GCCTGCATGAGACTCAACT</td>
</tr>
<tr>
<td>2244</td>
<td>Igf1</td>
<td>F</td>
<td>Mouse</td>
<td>GAGACTGGAGATGTACTGTGC</td>
</tr>
<tr>
<td>2245</td>
<td>Igf1</td>
<td>R</td>
<td>Mouse</td>
<td>CTCCTTTGCAGCTTCTTTTCTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2195</td>
<td>Cyp4a14</td>
<td>F</td>
<td>Mouse</td>
<td>CAGTCCAATTCTACTTACGAAAGGC</td>
</tr>
<tr>
<td>2196</td>
<td>Cyp4a14</td>
<td>R</td>
<td>Mouse</td>
<td>CCCCGAGAGACACTGTAAGC</td>
</tr>
<tr>
<td>1688</td>
<td>SFRS4</td>
<td>F</td>
<td>Mouse</td>
<td>CTCGCACAGAGTACAGACTTAT</td>
</tr>
<tr>
<td>1689</td>
<td>SFRS4</td>
<td>R</td>
<td>Mouse</td>
<td>TTGCCTCCTTGTGAGCATCT</td>
</tr>
<tr>
<td>1963</td>
<td>Cyp2a4</td>
<td>F</td>
<td>Mouse</td>
<td>GGAAGACGAACGGTGCTTTC</td>
</tr>
<tr>
<td>1964</td>
<td>Cyp2a4</td>
<td>R</td>
<td>Mouse</td>
<td>TTCCAGCATCATTCTAAGA</td>
</tr>
<tr>
<td>1965</td>
<td>Cyp2a5</td>
<td>F</td>
<td>Mouse</td>
<td>GGAAGACGAACGGTGCTTTT</td>
</tr>
<tr>
<td>1966</td>
<td>Cyp2a5</td>
<td>R</td>
<td>Mouse</td>
<td>TTCCAGCATCATTCGAAGC</td>
</tr>
<tr>
<td>1972</td>
<td>Sds</td>
<td>F</td>
<td>Mouse</td>
<td>GTCACCCCTGGCCCAAGATC</td>
</tr>
<tr>
<td>1973</td>
<td>Sds</td>
<td>R</td>
<td>Mouse</td>
<td>TTCTCAAGAGCAGACACAGC</td>
</tr>
<tr>
<td>1649</td>
<td>OAT</td>
<td>F</td>
<td>Mouse</td>
<td>AAGCAGGCGTTATCGTTC</td>
</tr>
<tr>
<td>1650</td>
<td>OAT</td>
<td>R</td>
<td>Mouse</td>
<td>ATCCACAGCCAGCCATCTAC</td>
</tr>
<tr>
<td>1647</td>
<td>β-cat</td>
<td>F</td>
<td>Mouse</td>
<td>GGTGCCACCACCCACAGCTCC</td>
</tr>
<tr>
<td>1648</td>
<td>β-cat</td>
<td>R</td>
<td>Mouse</td>
<td>GCTCGGACCCTCTGAGCCCT</td>
</tr>
<tr>
<td></td>
<td>Primer Pair</td>
<td>Orientation</td>
<td>Species</td>
<td>Sequence</td>
</tr>
<tr>
<td>-----</td>
<td>-------------</td>
<td>-------------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>2340</td>
<td>P1, RP23</td>
<td>F</td>
<td>Mouse</td>
<td>TGGGAGGAACACACTGAAAC</td>
</tr>
<tr>
<td>2341</td>
<td>P2, RP23</td>
<td>R</td>
<td>Mouse</td>
<td>TCTCTGCCAGGTTGACTACA</td>
</tr>
<tr>
<td>2352</td>
<td>P3, Cyp2a4as</td>
<td>F</td>
<td>Mouse</td>
<td>CCTGGGAGGAACACACT</td>
</tr>
<tr>
<td>2353</td>
<td>P4, Cyp2a4as</td>
<td>R</td>
<td>Mouse</td>
<td>ACTACAGGCACTGACTG</td>
</tr>
<tr>
<td>2354/237</td>
<td>P5, Cyp2a5as</td>
<td>F</td>
<td>Mouse</td>
<td>CCTGGGAGGAACACACC</td>
</tr>
<tr>
<td>2355/237</td>
<td>P6, Cyp2a5as</td>
<td>R</td>
<td>Mouse</td>
<td>GCAAGCAGTCACTCCTG</td>
</tr>
<tr>
<td>2400</td>
<td>ApoB1</td>
<td>F</td>
<td>Mouse</td>
<td>GGGCAGCTATATAGTAAGTTTCTGTG</td>
</tr>
<tr>
<td>2401</td>
<td>ApoB1</td>
<td>R</td>
<td>Mouse</td>
<td>TTTGGGTCTTGAAATTCCAGGTGCTTG</td>
</tr>
<tr>
<td>2460</td>
<td>Cyp2s1as</td>
<td>F</td>
<td>Mouse</td>
<td>ATACTCCGGTTTGCTCGTT</td>
</tr>
<tr>
<td>2461</td>
<td>Cyp2s1as</td>
<td>R</td>
<td>Mouse</td>
<td>GGTTCCTTCGGAACATGGG</td>
</tr>
<tr>
<td>2456</td>
<td>RP24</td>
<td>F</td>
<td>Mouse</td>
<td>GCCATATGGATGGGAGACATTTC</td>
</tr>
<tr>
<td>2457</td>
<td>RP24</td>
<td>R</td>
<td>Mouse</td>
<td>TCAGGCAGCAGCATTTCTTGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1130</td>
<td>H19</td>
<td>F</td>
<td>Mouse</td>
<td>GGAGAGGACAGAAGGGCAGT</td>
</tr>
<tr>
<td>1131</td>
<td>H19</td>
<td>R</td>
<td>Mouse</td>
<td>GTTCAAGGTAGGGGGAGGAG</td>
</tr>
<tr>
<td>P891</td>
<td>L30-U</td>
<td>F</td>
<td>Mouse</td>
<td>GGCAAAGCGAAGTTGGTTATC</td>
</tr>
<tr>
<td>P892</td>
<td>L30-U</td>
<td>R</td>
<td>Mouse</td>
<td>CTTCTTCCCAATGCTGGATAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>P579</td>
<td>U6</td>
<td>F</td>
<td>Mouse</td>
<td>GTGCTCGCTTCCGGCAGCACAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ATA</td>
</tr>
<tr>
<td>P580</td>
<td>U6</td>
<td>R</td>
<td>Mouse</td>
<td>CGAATTTGCGTGCATCCTTG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CG</td>
</tr>
</tbody>
</table>

Primers with “P” in front indicate primers from the Peterson Lab
Table 2.2. Oligos used for RT-PCR and PCR.

<table>
<thead>
<tr>
<th>Primer Number</th>
<th>Gene</th>
<th>Primers</th>
<th>Species</th>
<th>Sequence</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>2304</td>
<td>SL Zhx1, Lamp2FA</td>
<td>F</td>
<td>Sea Lamprey</td>
<td>AGTGCAAGCAGT GCCACC</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>2305</td>
<td>SL Zhx1, Lamp2RA</td>
<td>R</td>
<td>Sea Lamprey</td>
<td>ACCAGCTGCGGA TGCGCTC</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>2306</td>
<td>SL Zhx1, Lamp2F2</td>
<td>F</td>
<td>Sea Lamprey</td>
<td>TTCGCGCGCTTC CGGTAC</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>2307</td>
<td>SL Zhx1, Lamp2R2</td>
<td>R</td>
<td>Sea Lamprey</td>
<td>CGAAGCTCGCT TCAGCT</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>2357</td>
<td>SL Zhx1/2, LamF4</td>
<td>F</td>
<td>Sea Lamprey</td>
<td>GCATCGTGGA ACT CTGCTCAATT TGCC CG</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>2358</td>
<td>SL Zhx1/2, LampR4</td>
<td>R</td>
<td>Sea Lamprey</td>
<td>CCAAGGCCCTGC AGT GATCGCGCT C</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>1651</td>
<td>Zhx2 gene</td>
<td>F</td>
<td>Mouse</td>
<td>GGACCGAATCTC ACT ATTTA ACTCA</td>
<td>PCR</td>
</tr>
<tr>
<td>1654</td>
<td>Zhx2 gene</td>
<td>R</td>
<td>Mouse</td>
<td>AGGCTGCGAGTG GGTTAGAA</td>
<td>PCR</td>
</tr>
<tr>
<td>1354</td>
<td>Cre</td>
<td>F</td>
<td>Mouse</td>
<td>ACCTGAAGATGT TCGGATTATCT</td>
<td>PCR</td>
</tr>
<tr>
<td>1355</td>
<td>Cre</td>
<td>R</td>
<td>Mouse</td>
<td>ACCGTCAGTACG TGAGATATCT</td>
<td>PCR</td>
</tr>
<tr>
<td>2185</td>
<td>Slp 1500 promoter</td>
<td>F</td>
<td>Mouse</td>
<td>CTCGAGCTTGGG TTGCCCTGTGTTTCT</td>
<td>PCR</td>
</tr>
<tr>
<td>2227</td>
<td>Slp 2000 promoter</td>
<td>F</td>
<td>Mouse</td>
<td>CTCGAGGTTGG CCAGTCCTGGGA</td>
<td>PCR</td>
</tr>
<tr>
<td>2186</td>
<td>Slp 1500/2000 promoter</td>
<td>R</td>
<td>Mouse</td>
<td>AAGCTTTCTGGT CTGACTGTCTGCT TCC</td>
<td>PCR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>386</td>
<td>pGEMU</td>
<td>F</td>
<td>Plasmid</td>
<td>CACTATAGGGCG AATTGG</td>
<td></td>
</tr>
<tr>
<td>387</td>
<td>PGEML</td>
<td>R</td>
<td>Plasmid</td>
<td>CAAGCTATGCAT CCAACG</td>
<td></td>
</tr>
<tr>
<td>2308</td>
<td>pME18S1F</td>
<td>F</td>
<td>Plasmid</td>
<td>TCTTTGTTTTCGT TTCTGTTCTGC</td>
<td></td>
</tr>
<tr>
<td>2310</td>
<td>Stat5b3F</td>
<td>F</td>
<td>Stat5b</td>
<td>CGAGGGCAGCCT GGACG</td>
<td></td>
</tr>
<tr>
<td>2312</td>
<td>Stat5b5F</td>
<td>F</td>
<td>Stat5b</td>
<td>CTCAGGAGAGAA TGTTTTGGAATC</td>
<td></td>
</tr>
<tr>
<td>2313</td>
<td>PME18S5</td>
<td>R</td>
<td>Plasmid</td>
<td>ATACATTGATGA GTTTGGACAAAC C</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3. Genotyping PCR conditions.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>3 min</td>
<td>1</td>
</tr>
<tr>
<td>95°C</td>
<td>30 sec</td>
<td></td>
</tr>
<tr>
<td>55°C</td>
<td>30 sec</td>
<td>35</td>
</tr>
<tr>
<td>72°C</td>
<td>40 sec</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>5 min</td>
<td>1</td>
</tr>
<tr>
<td>4°C</td>
<td>$\infty$</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.4. Liver perfusion media recipe (pH 7.3-7.4).

<table>
<thead>
<tr>
<th>Ingredient</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.0 g NaCl</td>
<td></td>
</tr>
<tr>
<td>0.4 g KCl</td>
<td></td>
</tr>
<tr>
<td>0.0767 g NaH₂PO₄</td>
<td></td>
</tr>
<tr>
<td>0.1205 g Na₂HPO₄</td>
<td></td>
</tr>
<tr>
<td>2.38 g HEPES</td>
<td></td>
</tr>
<tr>
<td>0.35 g NaHCO₃</td>
<td></td>
</tr>
<tr>
<td>0.19 EGTA</td>
<td></td>
</tr>
<tr>
<td>0.9 g Glucose</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.5. Liver digestion media recipe (pH 7.3-7.4).

<table>
<thead>
<tr>
<th>Ingredient</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.0 g NaCl</td>
<td></td>
</tr>
<tr>
<td>0.4 g KCl</td>
<td></td>
</tr>
<tr>
<td>0.0767 g NaH₂PO₄</td>
<td></td>
</tr>
<tr>
<td>0.1205 g Na₂HPO₄</td>
<td></td>
</tr>
<tr>
<td>2.38 g HEPES</td>
<td></td>
</tr>
<tr>
<td>0.35 g NaHCO₃</td>
<td></td>
</tr>
<tr>
<td>0.560 g CaCl₂ x H₂O</td>
<td></td>
</tr>
<tr>
<td>8.4 uL 1M MgSO₄ x 7 H₂O</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.6. Cyp2a4as and Cyp2a5as oligos used for 5'/3' RACE and RT-PCR.

<table>
<thead>
<tr>
<th>Primer Number</th>
<th>Primer Name</th>
<th>Primer Name</th>
<th>Sequence (5' to 3')</th>
<th>Tm °C</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>2340</td>
<td>P1</td>
<td>Cyp2a4ASF</td>
<td>TGGGAGGAACACTGAAAC</td>
<td>54.9</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>2341</td>
<td>P2</td>
<td>Cyp2a4ASR</td>
<td>TCTCTGCCAGGTGACTACA</td>
<td>55.1</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>2352</td>
<td>P3</td>
<td>Cyp2a4AsFP2</td>
<td>CCTGGGAGGAACACACT</td>
<td>53.9</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>2353</td>
<td>P4</td>
<td>Cyp2a4AsRP2</td>
<td>ACTACAGGCACACTG</td>
<td>51.0</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>2354/2373</td>
<td>P5</td>
<td>Cyp2a5AsFP1</td>
<td>CCTGGGAGGAACACACC</td>
<td>55.3</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>2355/2372</td>
<td>P6</td>
<td>Cyp2a5AsRP1</td>
<td>GCAGGCACGTACCTTG</td>
<td>55.9</td>
<td>RT-qPCR, 5' RACE, RT-PCR</td>
</tr>
<tr>
<td>2380</td>
<td>P7</td>
<td>Cyp2a4as/Cyp2a5asR1</td>
<td>CACATGTCTCTGCCAGGTGACT</td>
<td>59.0</td>
<td>5' RACE</td>
</tr>
<tr>
<td>2377</td>
<td>P8</td>
<td>Cyp2a4asR2</td>
<td>GTGACTACAGGCACACTG</td>
<td>55.1</td>
<td>5' RACE, RT-PCR</td>
</tr>
<tr>
<td>2375</td>
<td>P9</td>
<td>Cyp2a4asR1</td>
<td>ATAGGCCTATAATCCCATGACTCAAGG</td>
<td>59.6</td>
<td>5' RACE</td>
</tr>
<tr>
<td>2370</td>
<td>P10</td>
<td>Cyp2a5asR1</td>
<td>ATAGGCCTATAATCCCATGACTCAAG</td>
<td>58.8</td>
<td>5' RACE</td>
</tr>
<tr>
<td>2395</td>
<td>P11</td>
<td>Cyp2a4asR4</td>
<td>TGGTCCATCAGATTCCAGCAACGCATTAGA</td>
<td>60.0</td>
<td>5' RACE</td>
</tr>
<tr>
<td>2397</td>
<td>P12</td>
<td>Cyp2a4asR5</td>
<td>GTGAGAGGGTGCTTCAGCCAGCATAG</td>
<td>60.6</td>
<td>5' RACE</td>
</tr>
<tr>
<td>2452</td>
<td>P13</td>
<td>Cyp2a5asR6</td>
<td>TGGCCCTTCCTCAGCGTCTGGTTC</td>
<td>65.2</td>
<td>5' RACE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td></td>
</tr>
<tr>
<td>2453</td>
<td>P14</td>
<td>Cyp2a5asR7</td>
<td>GGTCTTGATGTC TGTCAGGTTGGGC</td>
<td>61.6</td>
<td>5' RACE</td>
</tr>
<tr>
<td>2454</td>
<td>P15</td>
<td>Common Primer</td>
<td>TAGTCAAAGCGG TCCCCGAAGACG</td>
<td>62.6</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>2376</td>
<td>P16</td>
<td>Cyp2a4asF1</td>
<td>GACCACAGAGAT CCACCTGTGTCT GCC</td>
<td>64.0</td>
<td>3' RACE</td>
</tr>
<tr>
<td>2371</td>
<td>P17</td>
<td>Cyp2a5asF1</td>
<td>CACAGAGATCCA CCTGTGTCTGCT</td>
<td>60.5</td>
<td>3' RACE</td>
</tr>
<tr>
<td>2378</td>
<td>P18</td>
<td>Cyp2a4asF2</td>
<td>AGTGcctgggag gaacacact</td>
<td>60.8</td>
<td>3' RACE</td>
</tr>
<tr>
<td>2379</td>
<td>P19</td>
<td>Cyp2a4asF3</td>
<td>AGACacagacac atgaacagtc</td>
<td>54.7</td>
<td>3' RACE</td>
</tr>
<tr>
<td>2374</td>
<td>P20</td>
<td>Cyp2a5as5F3</td>
<td>AGACACAGACAC ATGAACCAGG</td>
<td>56.9</td>
<td>3' RACE</td>
</tr>
</tbody>
</table>

Table 2.7. Cyp2a4as and Cyp2a5as 5' RACE PCR conditions.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>3 min</td>
<td>1</td>
</tr>
<tr>
<td>95°C</td>
<td>30 sec</td>
<td>35</td>
</tr>
<tr>
<td>65°C</td>
<td>30 sec</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>1 min</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>7 min</td>
<td>1</td>
</tr>
<tr>
<td>4°C</td>
<td>$\infty$</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.8. Cyp2a4as 5’ RACE PCR conditions.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>3 min</td>
<td>1</td>
</tr>
<tr>
<td>95°C</td>
<td>30 sec</td>
<td>35</td>
</tr>
<tr>
<td>55°C</td>
<td>30 sec</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>1 min 30 sec</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>7 min</td>
<td>1</td>
</tr>
<tr>
<td>4°C</td>
<td>∞</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.9. Cyp2a5as 5’ RACE PCR conditions.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>98°C</td>
<td>30 sec</td>
<td>1</td>
</tr>
<tr>
<td>98°C</td>
<td>10 sec</td>
<td></td>
</tr>
<tr>
<td>55°C</td>
<td>30 sec</td>
<td>35</td>
</tr>
<tr>
<td>72°C</td>
<td>1 min</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>2 min</td>
<td>1</td>
</tr>
<tr>
<td>4°C</td>
<td>∞</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.10. Cyp2a4as 5' RACE nested PCR conditions.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>3 min</td>
<td>1</td>
</tr>
<tr>
<td>95°C</td>
<td>30 sec</td>
<td></td>
</tr>
<tr>
<td>65°C</td>
<td>30 sec</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>1 min 30 sec</td>
<td>35</td>
</tr>
<tr>
<td>72°C</td>
<td>7 min</td>
<td>1</td>
</tr>
<tr>
<td>4°C</td>
<td>∞</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.11. Cyp2a5as 5' RACE nested PCR conditions.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>98°C</td>
<td>30 sec</td>
<td>1</td>
</tr>
<tr>
<td>98°C</td>
<td>10 sec</td>
<td></td>
</tr>
<tr>
<td>63°C</td>
<td>30 sec</td>
<td>35</td>
</tr>
<tr>
<td>72°C</td>
<td>1 min</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>2 min</td>
<td>1</td>
</tr>
<tr>
<td>4°C</td>
<td>∞</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.12. 3’ RACE PCR conditions.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>3 min</td>
<td>1</td>
</tr>
<tr>
<td>95°C</td>
<td>30 sec</td>
<td></td>
</tr>
<tr>
<td>64°C* or 60°C</td>
<td>30 sec</td>
<td>35</td>
</tr>
<tr>
<td>72°C</td>
<td>1 min</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>7 min</td>
<td>1</td>
</tr>
<tr>
<td>4°C</td>
<td>∞</td>
<td></td>
</tr>
</tbody>
</table>

*annealing temperature for P16

Table 2.13. 3’ RACE nested PCR conditions.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>3 min</td>
<td>1</td>
</tr>
<tr>
<td>95°C</td>
<td>30 sec</td>
<td></td>
</tr>
<tr>
<td>60°C</td>
<td>30 sec</td>
<td>35</td>
</tr>
<tr>
<td>72°C</td>
<td>1 min</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>7 min</td>
<td>1</td>
</tr>
<tr>
<td>4°C</td>
<td>∞</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.14. Cyp2a4as and Cyp2a5as nuclear RT-PCR conditions.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>98°C</td>
<td>30 sec</td>
<td>1</td>
</tr>
<tr>
<td>98°C</td>
<td>10 sec</td>
<td></td>
</tr>
<tr>
<td>60, 65, 70°C</td>
<td>30 sec</td>
<td>35</td>
</tr>
<tr>
<td>72°C</td>
<td>1 min</td>
<td></td>
</tr>
<tr>
<td>72°C</td>
<td>2 min</td>
<td>1</td>
</tr>
<tr>
<td>4°C</td>
<td>∞</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER III

A COMPREHENSIVE ANALYSIS OF ZHX GENES AND PROTEINS THROUGHOUT CHORDATE EVOLUTION

The Zinc Fingers and Homeoboxes (Zhx) family is comprised of three members, Zhx1, Zhx2, and Zhx3. Mouse Zhx1 was the first Zhx gene to be characterized and was identified by immunoscreening a mouse bone marrow cDNA library (Barthelemy et al., 1996). Since their initial discovery, studies have shown that Zhx proteins are localized to the nucleus and can heterodimerize and homodimerize with one another, as well as with nuclear factor-Y subunit A (NF-YA) (Hirano et al., 2002; Kawata et al., 2003a; Kawata et al., 2003b; Yamada et al., 2002; Yamada et al., 2003; Yamada et al., 1999a; Yamada et al., 1999b). Due to a natural hypomorphic mutation in Zhx2, studies have largely focused on our understanding of Zhx2 in BALB/cJ mice. Studies in this mouse strain indicated that Zhx2 was responsible for the postnatal silencing of alpha fetoprotein (AFP), glypican 3 (GPC3), and long non-coding RNA H19 (H19) genes in the liver; all of these genes are frequently reactivated in hepatocellular carcinoma (HCC) (Morford et al., 2007; Pachnis et al., 1984; Peterson et al., 2011; Peyton et al., 2000). When placed on a high fat diet, BALB/cJ mice exhibit lower serum lipid levels and reduced atherosclerotic lesions compared to other mouse strains. This trait is also due to the Zhx2 mutation; microarray analysis identified numerous dysregulated hepatic genes involved in lipid metabolism in BALB/cJ mice (Gargalovic et al., 2010). More recently, using BALB/cJ mice and C57BL/6 mice in which the Zhx2 gene has been deleted by cre-mediated deletion, Zhx2 was shown to contribute to sex biased regulation of several liver genes, including those
encoding a number of Cytochrome P450 (Cyp) enzymes and Major urinary proteins (Mups) (Creasy et al., 2016; Jiang et al., 2017). In addition, Zhx2 has been reported as a tumor suppressor in hepatocellular carcinoma, where it represses Cyclin A, Cyclin E, AFP, and multidrug resistance 1 (MDR1) expression (Ma et al., 2015; Shen et al., 2008; Yue et al., 2012). In contrast, Zhx2 is oncogenic in clear cell renal carcinoma (Zhang et al., 2018). Other studies have investigated Zhx1 and Zhx3 expression in various cancers; Zhx1 and Zhx3 mRNA levels are lower in renal cell carcinoma and Zhx1 mRNA levels are lower in human HCC (Kwon et al., 2017; Wang et al., 2013a). Taken together, these studies demonstrate that the expression of Zhx genes is often dysregulated in cancer and that absence of or low Zhx2 expression is associated with the dysregulation of genes involved in many different processes including cancer, lipid metabolism, and sex-biased gene expression.

Zhx genes belong to the Zinc finger (ZF) class of homeobox genes that encode C2-H2 zinc fingers in addition to homeodomains (HD) (Burglin and Affolter, 2016). While ZFs have been shown to interact with DNA, RNA, and proteins, the HD is normally associated with DNA binding, suggesting that Zhx proteins function as transcription factors (Burglin and Affolter, 2016; Cassandri et al., 2017). The ZF class of homeobox proteins contain five families: Adnp, Tshz, Zeb, Zfhx, and Zhx. Specifically, Zhx proteins are predicted to contain two C2-H2 zinc fingers and four or five HDs (Spear et al., 2006). Zhx genes have the same unique gene structure in that they contain two 5’ non-coding exons, an unusually large third exon that encodes the entire Zhx protein, and a non-coding fourth exon (with the exception of Zhx3, in which the last two amino acids are encoded by exon 4) (Spear et al., 2006).
The phyla Chordata consists of three subphyla, Cephalocordata (including lancelet, or *Amphioxus*), Urochordata (including sea squirts, or *Ciona*), and Vertebrata (Satoh et al., 2014) (Figure 3-1). All three groups are characterized by the presence of a notochord, nerve cord, branchial slits, endostyle, myotomes, and a postanal tail. Chordata are part of the superphyla Deuterostomia (deuterostomes) which is separated from subphyla Protostomia (protostomes) that contain both flies (i.e. *Drosophila melanogaster*) and worms (i.e. *Caenorhabditis elegans*). Phylogenomic data places lancelets as the most basal chordate and sea squirt as being closer to the vertebrate lineage (Figure 3-1) (Delsuc et al., 2006). Lampreys are considered the most basal living vertebrate and are characterized by both lack of jaws (agnathostomes) and paired appendages (Lara-Ramirez et al., 2017). Furthermore, lampreys exhibit unique features that are not observed in jawed vertebrates (gnathostomes) and may have ancestral vertebrate biology that has been lost in gnathostomes (Smith et al., 2018). These include the ability to fully recover after spinal cord transection, the independent evolution of adaptive immune receptors, and physical rearrangement of the genome during development (Smith et al., 2018).

Because of their placement in vertebrate evolution, lamprey are an important model system for understanding both morphological and molecular vertebrate evolution (Lara-Ramirez et al., 2017). Specifically, lamprey are an important model for understanding the evolution of vertebrate features that are associated with gene duplication events (Green et al., 2015). Gene duplication results in redundancy and as a result, duplicated genes can acquire mutations that result in the loss, subfunctionalization or neofunctionalization of proteins. In comparison to single gene duplications, whole genome duplications (WGDs) result in an organism’s entire genome being able to acquire
new function. Genes that are duplicated as a result of WGD duplication are termed ohnologs. WGDs allow for duplicated genes to co-evolve with one another, leading to the formation of new biological networks (Kasahara, 2007).

It is generally accepted that the ancestral lineage of all or most vertebrates experienced two separate WGD’s, known as the “1R-2R hypothesis” (Sacerdot et al., 2018). Furthermore, genomic sequencing has revealed the occurrence of an additional teleost-specific WGD (Glasauer and Neuhauss, 2014). It has been suggested that these large-scale duplication events coincide with vertebrate diversification (Canestro et al., 2013). Support for the 2R hypothesis comes from analysis of gene families and synteny across Drosphila, Ciona, Branchiostoma, fish, and mammals (Canestro et al., 2009; Larsson et al., 2008; Smith and Keinath, 2015). Recent studies have relied upon large-scale in silico identification of ohnologous genes in order to reconstruct vertebrate ancestral genome states, and as a result, have proposed that two WGD events preceded the divergence of the sea lamprey lineage (Sacerdot et al., 2018). Alternatively, it has been proposed that only one WGD event preceded the divergence of sea lampreys; many ohnologous groups compared between the sea lamprey scaffold and chicken chromosomes show a 1:2 ratio (Smith et al., 2018). Although these studies find contradicting evidence for the timing of WGD events, both studies rely upon correct annotation and phylogenetic relationships of genes in order to correctly decipher the relationship of genes across chordate genomes (i.e. the relationship of genes within a species, as well as among species is important for correctly inferring WGD duplication events). For example, Smith et al. recently demonstrated through both synteny and phylogenetic analysis that sea lamprey and japanese lamprey Hox-ε and Hox-β predicted
proteins may have recently duplicated within the sea lamprey lineage. In contrast, other Hox genes show clustering between humans and japanese lamprey, suggesting that these genes were duplicated by WGD and share a common origin (Smith et al., 2018). These results suggest the need for continued rigorous evaluation of gene annotation and phylogenetic relationships of specific gene families within early chordates to ensure proper evolutionary inferences for larger scale studies.

It is known that protostome species such as *C. elegans* and *D. melanogaster* contain about 100 homeobox genes (Burglin and Affolter, 2016). In the deuterostome branch, the basal chordate, lancelet (*Branchiostoma floridae*) contains 133 homeobox genes (Burglin and Affolter, 2016; Delsuc et al., 2006). *Drosophila* contains two zinc finger homeodomain proteins (Zfhx1 and Zfhx2) and lancelets contain four ZF homeodomain proteins (Tshz, Zeb, Zfhx, and Zhx/Homez) (Takatori et al., 2008; Zhong and Holland, 2011b). This data suggests that Zhx genes first appeared with the emergence of the chordate lineage.

With the availability of high-throughput genome assemblies, we are now able to comprehensively investigate Zhx gene structure and determine phylogenetic relationships of Zhx genes throughout chordate evolution. We have identified the most conserved regions across vertebrate Zhx proteins as well as specific features that may be unique to Zhx1, Zhx2, or Zhx3. We show that Zhx proteins contain a unique, highly conserved amino-terminal region. We also investigated gnathostome homeodomain (HD) homology by multiple sequence alignment (MSA) and find that gnathostome HD1 is the most conserved HD among Zhx proteins. Furthermore, comparison of ZF regions by MSA
show that higher amino acid sequence homology exists between gnathostome Zhx1 and Zhx2, whereas the zinc finger region of Zhx3 is more divergent across species. In addition, previous analyses from our lab identified an unannotated Zhx gene within the 2013 sea lamprey assembly (Smith et al., 2013). Further investigation shows that this sea lamprey Zhx coding region, in contrast to the Zhx genes of gnathostomes, contains introns and is expressed as two alternatively spliced transcripts that are present in multiple adult lamprey tissues. Using BLASTP analysis, we also confirm the presence of single Zhx genes in both lancelet and sea squirt. Comparisons of Zhx coding regions between human, mouse, and earlier chordate lineages revealed dynamic intron gains and losses throughout chordate evolution. Using our sequence data, we explored the relationship of chordate Zhx full-length proteins, zinc finger regions, and homeodomains using both distance-based and Bayesian phylogenetic tree-building analyses. Our results indicate that Zhx genes emerged within the chordate lineage and we identify Zhx3 as the primordial Zhx gene. Our data demonstrate the importance of gene family characterization that can be applied towards determining WGD duplication events and correctly annotating early vertebrate genome assemblies.

RESULTS

Identification of gnathostome and early chordate Zhx genes. Previous BLAST analyses in our lab indicated that Zhx genes were not in C. elegans, D. melanogaster or yeast but were present in humans, dogs, rodents, and zebrafish (Spear et al., 2006), suggesting that Zhx genes were restricted to the vertebrate lineage. To expand upon our previous analyses and gain further insight into Zhx proteins within the vertebrate lineage,
we utilized BLASTP to analyze the NCBI database. To ensure representation of Zhx proteins throughout the vertebrate lineage, we included taxa from the major clades of mammals, birds, amphibians, lobe-finned fish, ray-finned fish, and sharks. Our analysis of 22 vertebrate species confirmed that all these species contained all three Zhx proteins (Appendix Table I).

**Sea lamprey contains a single functional Zhx-like gene with multiple coding exons.** Our initial BLAST analysis of the sea lamprey genome (2013 assembly) (Smith et al., 2013) identified a potential Zhx-like gene. Using a series of RT-PCR, 3’-RACE and 5’ RACE experiments with adult lamprey liver RNA (provided by Dr. Jeramiah Smith), previous members of our lab characterized the full-length sea lamprey Zhx-like transcript that encodes a 1079 amino acid protein that contains two ZFs and five HDs. In contrast to gnathostome Zhx genes, in which the entire Zhx coding region was contiguous on a single large exon (except for the last two amino acids of Zhx3), the sea lamprey mRNA was comprised of 7 coding exons and a 5’ non-coding exon (Figure 3-2A). Further studies (see below) indicate that the Zhx protein encoded by this gene is most similar to gnathostome Zhx3.

**Sea lamprey Zhx has two spliced isoforms present in multiple tissues.** Our initial cloning revealed that the sea lamprey Zhx gene encoded two spliced isoforms in adult liver due to the usage of 2 splice donor sites at the end of exon 4 that resulted in the presence or absence of 75 nucleotides (encoding 25 amino acids). To identify whether
these spliced isoforms exist in multiple tissues, we performed both RT-PCR and RT-qPCR using RNA from liver, kidney, muscle, brain, and eyes. Primers were used that span several exons, including those that lie within exons 3 and 5 and therefore span alternatively-spliced exon 4 (Figure 3-2A). We also used primers for elongation factor 1-alpha (EF1A), which has been previously used as a housekeeping gene to normalize RT-qPCR data from sea lamprey tissues (Kasamatsu et al., 2010). However, our results using tissues from two different lamprey demonstrated that EF1A is not expressed equally across tissues (Figure 3-2B and C). Our largest variability came from liver (~16 vs 19 cycles between Lampreys 1 and 2) and kidney (~21 vs ~15 cycles between Lampreys 1 and 2). Among those that did not show as much variability between tissues in the two lamprey samples were muscle (~16 vs ~14.4), blood (~14 in both) and eye (~16 and ~16.2). This result was surprising because large variability in housekeeping gene levels has not been seen before with RNA prepared from mouse using the same RNA isolation and cDNA synthesis protocols; tissue differences between mouse and lamprey, tissue storage, RNA preparation, or cDNA synthesis efficiency may have played a role in cycle number variability. This variability may have been caused by time in captivity, where one sea lamprey was sacrificed at an earlier timepoint in captivity. Studies in additional lamprey are needed in order to compare Zhx RNA levels across multiple tissues. Given this inherent variability between tissues types, we are unable to quantitate Zhx levels across all lamprey tissues, but can compare across animals for both blood and eye, which show relatively equal amplification of Zhx transcript isoforms in both tissue types. Regardless of this variability, this data indicates that both Zhx mRNA spliced isoforms are present in all tissues.
Characterization of the sea lamprey *Zhx* pseudogene. After our characterization of the sea lamprey *Zhx* gene, the 2018 sea lamprey assembly became available (Smith et al., 2018). Analysis of this updated sequence identified what initially appeared to be two additional *Zhx* genes that were annotated as “*Zhx2*” and “*Zhx1*” (Figure 3-3A). A closer analysis of the annotated “*Zhx2/Zhx1*” region revealed that these annotated genes were separated by a small 27 bp gap between the translation stop codon of “*Zhx2*” and initiating methionine in “*Zhx1*” (Figure 3-3B). It seems highly unlikely that two distinct genes are separate by such a small region. Furthermore, the upstream “*Zhx2*” gene contains two ZFs and three HDs, whereas the downstream “*Zhx1*” gene encodes a single HD. It should also be noted that the predicted *Zhx* proteins encoded by these two genes indicate the presence of multiple introns (Figure 3-3C). To test whether either of these genes were expressed, individually or as a single transcript, RT-qPCR was performed with lamprey RNA from multiple tissues using a series of primers (shown in Fig. 3-3C). Despite numerous efforts using different conditions, we were unable to detect any transcripts (data not shown). Furthermore, PCR with these primers using genomic DNA failed to generate any amplicons (data not shown). Based on these data, we propose that this region encodes a single *Zhx* pseudogene (which we will refer to as *ZhxP*), although we cannot rule out the possibility that this region has not been annotated correctly or may be expressed in other tissues.

*Zhx* gene structures are dynamic throughout chordate evolution. The most striking aspect of sea lamprey *Zhx*, as well as *ZhxP*, is the presence of multiple introns (Figure 3-4), which is in contrast to the three *Zhx* genes in all gnathostomes analyzed. This led us
to ask whether Zhx genes exist in other lower chordates and, if so, whether these genes contain introns. We used BLASTP with the elephant shark Zhx proteins to search for Zhx genes in both sea squirt (Ciona intestinalis) and lancelet (Branchiostoma floridae) genome assemblies. This search identified a single Zhx gene in both species (Figure 3-4). The lancelet Zhx gene is predicted to contain 5 coding exons and encodes a protein that contains 2 ZFs and 5 or 6 HDs. In contrast, the coding region of the sea squirt Zhx gene contains no introns and encodes a protein that contains a single ZF and two HDs.

One question is whether the primordial Zhx gene contained multiple introns which were lost in sea squirt and gnathostomes, or whether introns were introduced in lancelet and lamprey. The fact that the introns in lancelet Zhx, lamprey Zhx and lamprey ZhxP are in different locations (Figure 3-4) would argue that these introns were introduced after the lancelet and lamprey split off from a common chordate ancestor. A second question is when duplication events occurred to give rise to Zhx and ZhxP in lamprey and the three Zhx genes in gnathostomes. A model to account for these duplications is described below (Fig. 3-12).

Genome annotations for the sea lamprey assembly were completed using MAKER software (Campbell et al., 2014). MAKER annotates and masks repetitive elements in the genome, aligns protein and RNA evidence to the assembly in a splice-aware fashion to identify splice sites, and compares all predicted gene models to RNA and protein alignments to make ab initio gene models. Importantly, MAKER is only as good as the protein databases that it is supplied and partly relies on homology, RNA-sequencing data, and open reading frame predictions. Although this program can be very useful for initial annotations of new genomes, verifying these gene predictions are crucial.
for future annotations of genome assemblies which rely on previous builds, or for applications such as *in situ* hybridization to visualize mRNA localization, or gene knockdown (Boorman and Shimeld, 2002; McCauley and Bronner-Fraser, 2006). Therefore, new assembly gene annotations should be validated by other means to ensure high success of future studies. Having identified Zhx genes in lower chordates, we investigated the phylogenetic relationship of sea lamprey, sea squirt, lancelet, and gnathostome Zhx predicted proteins.

**Annotation of ZF and HD regions in Zhx proteins.** We next sought to determine the relationship of chordate Zhx proteins. Mouse and human Zhx proteins were predicted to have two C$_2$H$_2$ ZF and four or five HDs (Spear et al., 2006). To determine whether these conserved regions were common across the chordate lineage, Zhx proteins were manually curated using Geneious 11.1.5 software. Specifically, the predicted 60 amino acid HDs were anchored by highly conserved F/Y/W48 and F/Y/W49, followed by identification of highly conserved positions L/F/M/A16 and R/L53; those with conserved amino acids at these positions were designated as bonified HDs (Burglin and Affolter, 2016). The ZFs were anchored by two amino terminal cysteines and two carboxy terminal histidines and were based on the classical zinc finger (CX$_{2-4}$CX$_{12}$HX$_{2-6}$H)(Iuchi, 2001). As a secondary test to validate ZFs and HDs, Zhx proteins were queried against the NCBI conserved domain database (CDD). Interestingly, from cross-checking our manual curation against the CDD database, we found that gnathostome Zhx1 and Zhx3 proteins were often predicted to contain a fifth homeodomain region, although conserved residues in this fifth HD were often lacking. Therefore, we have decided to name Zhx1 and Zhx3 HDs that do
not contain the conserved amino acid residues described above as remnant HDs (Holland et al., 2007; Zhong and Holland, 2011a). It should be noted, however, that none of the HDs, including those judged to be functional or remnant, have been tested functionally.

Alignment of gnathostome Zhx proteins reveals the highest conservation within the amino terminus. To accurately assess the relationship of early chordate Zhx proteins to gnathostome proteins, we first needed to determine the relationship of gnathostome Zhx proteins to each other. To first gain insight about the conservation of Zhx proteins throughout gnathostome evolution, MUSCLE multiple sequence alignment (MSA) within Geneious software was used to align Zhx predicted proteins from all 22 species (Edgar, 2004; Edgar and Batzoglou, 2006). MSA revealed that nine highly conserved regions exist within gnathostome proteins, the amino-terminus, the zinc finger region (ZFR, containing both ZF1 and ZF2), two regions between ZF2 and HD1, and the homeodomain regions (Figure 3-5A). Currently, we are unsure of the significance for the region between ZF2 and HD1 having high homology (57.9%, 65.1%). Of the remaining regions, the most conserved region was the amino terminus followed by HD1, HD3, HD2, ZFR, HD4, and HD5. Interestingly, first 12-13 amino acids of the amino terminus were extremely conserved (>93%) across all Zhx proteins (Figure 3-5B). Furthermore, BLASTP of this sequence against the multiple protein databases indicated that this sequence is not found in other proteins but is unique to Zhx proteins.

Although the number of publicly available sequenced animal genomes continues to grow, many are poorly annotated and rely solely upon annotation software to identify homology between a newly assembled genome and other thoroughly annotated genomes
such as mouse (mm10) and human (hg38). For example, a simple search of Zhx1 in the UCSC database against the elephant shark 2013 assembly reveals multiple hits for Zhx1. In addition, Refseq gene predictions show homology to all three Zhx proteins from multiple species. This type of issue is pervasive throughout genome assemblies and can confound efforts to identify homology of genes or proteins across multiple species, especially in genomes that remain poorly annotated and assembled. To verify that our predicted gnathostome Zhx proteins were homologous to mouse and human Zhx1, Zhx2, and Zhx3, we performed phylogenetic analysis using distance-based tree building. Specifically, we used neighbor-joining tree building with Jukes-Cantor genetic distance modeling to calculate pairwise distances from our MUSCLE MSA. Using this analysis, distances between taxa are first calculated based on the number of amino acid differences between two proteins, while assuming equal probabilities of encountering any given amino acid within a peptide sequence (.05), and equal rates of substitution for any amino acid over time (Yang and Rannala, 2012). After the distances have been calculated, all sequences are placed into a star tree and sequences that are most alike are clustered together until a final tree is resolved that minimizes tree length. Using this method, we were able to determine that all annotated Zhx1, Zhx2, and Zhx3 proteins across species correctly clustered as predicted (Figure 3-5C).

Alignment of early chordate and gnathostome Zhx predicted proteins. Using the criteria described for ZF and HD annotation described above, we aligned lancelet, sea squirt, and sea lamprey Zhx ZFs and HDs to those found in gnathostomes (Figure 3-6A). We chose not to include lamprey ZhxP because our previous studies indicate it is likely a
pseudogene. The ZFR aligns within all chordates; the single sea squirt ZF aligns with ZF1 from all other species. We found that lancelet and gnathostome HD1, HD2 and HD3 align. Interestingly, lancelet HD4 does not align with any gnathostome HDs, whereas lancelet HD5 and HD6 align with gnathostome HD4 and remnant HD5, respectively. Sea squirt HD1 aligns with HD2 from other species and sea squirt HD2 aligns with lancelet HD6 and gnathostome remnant HD5. Sea lamprey HD1, HD2, and HD3 all align with their gnathostome counterparts. However, sea lamprey HD4 did not align with any other HDs and sea lamprey HD5 aligns with gnathostome HD4 and lancelet remnant HD5. MUSCLE MSA revealed that sea lamprey Zhx contains partial sequence identity to the amino terminus of gnathostome Zhx proteins, although not to the extent of conservation among gnathostome Zhx proteins. (Figure 3-6B). These data confirm the evolutionary relationship of gnathostome Zhx proteins with the Zhx proteins found in lower chordates.

**Evolutionary relationships of sea lamprey and gnathostome Zhx proteins.** Results from MAKER analysis (described previously) predicted that sea lamprey Zhx would cluster with gnathostome Zhx3 proteins in a phylogenetic tree. To test this, we needed to determine the best outgroup for phylogenetic analyses. Previous studies have shown that sea squirts have undergone significant gene loss, high amounts of amino substitution, and exhibit a marked reduction in synteny compared to vertebrates (Alexandra Louis, 2012; Putnam et al., 2008). Sea squirts appear to have lost one ZF and several HDs, whereas lancelet still have both ZFs and show similarity across many of their HDs to both gnathostomes and sea lamprey. We therefore used lancelet as an outgroup for phylogenetic analysis.
Our initial phylogenetic analysis used the neighbor-joining distance-based tree method as previously described in Figure 3-5 for determining the relationship of gnathostome proteins. This analysis could not definitively determine the position of sea lamprey Zhx compared to gnathostome Zhx proteins when using lancelet as an outgroup (Figure 3-7A). An important aspect of distance based trees is that they often underestimate the true distance because some sites may have experienced multiple substitution events or different rates of substitution events, especially in the case of divergent sequences. Because we are dealing with divergent sequences in our tree, neighbor-joining methods are not optimal.

As an alternative, we utilized Bayesian phylogenetic analysis. In contrast to distance-based trees, Bayesian phylogenetics determines the probability of data given certain parameters. These parameters are based on the structure of the substitution model and prior probability parameters. For proteins, the substitution model is based on the rates of substitution of amino acids across the proteins; rates of substitution can be equal, have variability across all sites, have invarianility across a portion of sites with equal substitution for the remaining sites, have invarianility across a portion of sites with unequal substitution for the rest of sites, or have correlated rates where adjacent sites are correlated to those amino acids that are variable. The prior probability parameters are based on the topology, the branch lengths, the stationary frequencies of amino acids, the proportion of invariable sites, and the shape parameter of the gamma distribution of rate variation. The prior probability default for topology is uniform, where there is an equal probability on all distinct resolved topologies, or trees. The alternative is to constrain trees so that certain nodes are always present in all trees discovered. Branch lengths can
either be clock constrained or unconstrained. Molecular clocks are used to estimate divergence of species across a tree. However, using a molecular clock parameter within a tree constrains every species (or branches) in a tree to a certain mutation rate. Therefore, using unconstrained branch lengths allow for mutations rates to vary within lineages, or branches in a tree. Stationary frequencies are essentially the composition of amino acids across all proteins. Many assume that the stationary frequencies of amino acids are uniform. Changing this distribution assumes that the expected proportion of a given amino acid will be different across all species during evolution, which in many cases is unknown. The prior for proportion of invariable sites is usually set to have a distribution between 0 and 1. This means that there is no prior knowledge about how proportionate or disproportionate amino acid substitutions may be across our dataset. The shape parameter of the gamma distribution is determined by the number of variable amino acids among proteins that are compared within the tree. Based on the number of variable amino acids, each tree is assigned a probability of occurring. Since we do not have any prior knowledge about this variation, the shape parameter default is set to uniform, and will therefore encompass a wide range of possible variable sites as our analyses proceeds.

Bayesian phylogenetics allows us to calculate a posterior distribution by multiplying a prior distribution (set by parameters described above) by a likelihood function (based on our observations, or different subtrees made). From this posterior distribution, the probability that a given tree will lie within the posterior distribution can be calculated. MrBayes software uses the Metropolis-coupled Markov chain Monte Carlo model to determine the probability distribution of trees, branch lengths, and substitution parameters (Ronquist et al., 2012). This model selects which posterior probability ratio is
greater given two subtrees. If the posterior probability ratio is greater for a previous tree versus the subsequent tree, then the model builds from the previous tree. MrBayes software runs two separate analyses from one dataset starting from two random trees. As the runs continue, the trees made from the separate runs should show convergence and this can be measured by the standard deviation of split frequencies.

Using the same MUSCLE MSA alignment for distance tree building which included all gnathostome Zhx proteins, sea lamprey Zhx and lancelet Zhx as an outgroup, we first determined which best-fit model to use for MrBayes by PROTTEST3. The JTT + I + G + F model, which relies on the Jones Taylor Thornton substitution matrix and incorporates a proportion of invariable sites (I), with rate variation among sites (G), and equal amino acid frequencies (F), was selected (Darriba et al., 2011). We incorporated these parameters into MrBayes software, and ran our analyses for 1 million generations (standard deviation of split frequencies = .0025). From this analysis, a consensus tree was generated and loaded into FigTree1.4.4 for visualization. Interestingly, Bayesian analysis, like distance tree building, was unable to resolve the position of sea lamprey Zhx (Figure 3-7B). However, in agreement with distance tree building, gnathostome Zhx1 and Zhx2 show higher similarity to each another than to Zhx3. Taken together, this data suggest that gnathostome and sea lamprey Zhx proteins are highly divergent across their entire length.

**Lancelet and sea lamprey Zhx proteins are most highly related to Zhx3 based on ZF region alignments.** Since we could show consistent alignment of ZFs and ZFRs across all chordate taxa, we considered whether ZFs or ZFR could deduce the relationship of sea
lamprey Zhx to gnathostome Zhx proteins. We began by determining the percent identity across gnathostome ZFs and ZFRs by alignment of ZF1, ZF2, and ZFR for gnathostome Zhx proteins using MUSCLE MSA. Both Zhx1 and Zhx2 contain the highest identities across ZF1, ZF2, and ZFR compared to Zhx3 (Table 3.1). To determine the conservation of ZFR residues between Zhx1, Zhx2, and Zhx3, we compared consensus sequences (>90% identity across all species) for each region. Both Zhx1 and Zhx2 shared the largest number of identical amino acid residues for ZF1, ZF2, and ZFR, the second most common residues was between Zhx1 and Zhx3 (Figure 3-8, Table 3.2). These data indicate that Zhx1 and Zhx2 ZFR are more highly related to each other than to Zhx3, and therefore have likely diverged more recently.

To identify the gnathostome ZFR most highly related to the sea lamprey ZFR, we used both distance-based and Bayesian phylogenetics. ZFRs for sea lamprey and gnathostome proteins were first aligned by MUSCLE MSA and then analyzed by Geneious Tree Builder or MrBayes software to produce consensus trees. For all trees, lancelet was used as an outgroup. Interestingly, distance-based methods clearly indicated that the sea lamprey ZFR was most closely related to gnathostome Zhx3 ZFRs (Figure 3-9). However, Bayesian analysis of this same alignment produced low percent probability for our trees and could not correctly resolve the relationship of Zhx proteins to one another, including Zhx1 and Zhx2 groups (data not shown). Therefore, further analysis will be needed to optimize Bayesian analysis to produce higher confidence trees, perhaps with the inclusion of a greater diversity of species. Furthermore, distance based trees may be suitable to deduce meaningful relationships because there is less divergence for many sequences used within the ZFR tree.
**HD1 is the most conserved gnathostome Zhx HD.** Since our previous analyses identified HD regions among the most conserved regions across gnathostome Zhx proteins, we further investigated the relationship of HDs to each other to identify regions that might be specific to Zhx1, Zhx2, or Zhx3. Furthermore, we anticipated that this analysis would identify conserved residues across Zhx HDs that could help in deducing the relationship of early chordate Zhx proteins to gnathostome Zhx proteins.

We first evaluated gnathostome HDs. Annotated gnathostome HD regions were aligned by MUSCLE MSA for all gnathostome Zhx proteins (Table. 3.3). HD1 showed the highest conservation, followed in order by HD3, HD2, HD4, and HD5 (Table 3.3). To find conserved residues within Zhx HDs, we performed MUSCLE MSAs for each HD within Zhx1, Zhx2, or Zhx3. Consensus sequences from each alignment were then filtered for having at least 90% conservation across all residues and compared against each other (Figure 3-10). Because a bonified HD5 region is only present in three species (not including remnant HD5s), we chose not to analyze conservation of this HD. From this analysis, we found that the highest amino acid conservation occurs in alpha helix 3 for all four HDs compared to alpha helix 1 and alpha helix 2. Previous research demonstrated that each alpha helix, as well as the amino terminal region of HDs, are important for contact with DNA bases (Gehring et al., 1994a; Gehring et al., 1994b).

Comparatively across Zhx HDs, the highest amount of conservation occurs in alpha helix 3 of HD1 (81.9%). This data suggests that, due to the highest conservation of HD1 and its alpha helix 3, these amino acid residues may be important functionally for Zhx1, Zhx2, and Zhx3. Based on the number of matches between amino acids within Zhx1, Zhx2, and Zhx3 HDs, Zhx1 HDs have the highest homology to Zhx3 HDs (Table 3.4). Further
evaluation across each HD revealed that HD4 contains the least amount of conservation across all Zhx proteins, and contains the highest amount of conservation between Zhx1 and Zhx2 (Figure 3-10, Table 3.4). However, the evolutionary relationship of gnathostome Zhx proteins were not readily deduced based on the total number of amino acids shared among Zhx HDs.

**Lancelet and sea lamprey Zhx HDs are most highly related to Zhx3.** Although our full length Zhx alignments could not identify the gnathostome Zhx protein most closely related to the sea lamprey Zhx protein, ZFR analysis linked Zhx3 to sea lamprey Zhx. We further tested the relationship of gnathostome Zhx proteins with those of lower chordates by HD phylogenetic analyses. MUSCLE MSA was used to align 313 HDs from lancelet, sea squirt, sea lamprey Zhx and Zhx, and all gnathostomes, including those assigned as remnant HDs. We included both remnant and ZhxP HDs in this analysis because they might provide information about the relationship of early chordate HDs and might be useful in discovering possible duplication events within the sea lamprey lineage. Bayesian phylogenetic analysis identified all sea lamprey Zhx HDs to be most closely related to Zhx3 (Figure 3-11), consistent with our ZFR analysis. This analysis also indicates that all lancelet HDs are most closely related to Zhx3. We also find that sea lamprey ZhxP HD1 and HD2 group with gnathostome Zhx2 for HD1 and HD2, and ZhxP HD3 groups best with Zhx3 HD4. Given that ZhxP HD3 is most closely related to Zhx3 HD4, this could indicate that Zhx3 HD4 gave rise to ZhxP HD3. This region is currently annotated as Zhx1 in the sea lamprey assembly, and may suggest that Zhx3 gave rise to Zhx1.
Curiously, we also find that sea squirt HD1 groups most closely with Zhx1 HD1 suggesting that sea squirt Zhx contains domains similar to Zhx1 and Zhx3. Within this same lineage, the elephant shark Zhx1 HD1 branch lies between Zhx1 and Zhx3, indicating that it contains attributes similar to both Zhx1 and Zhx3, further suggesting that Zhx3 gave rise to Zhx1.

Interestingly, sea lamprey HD4 and HD5 showed monophyletic grouping with lancelet HD4, lancelet HD6, and sea squirt HD2 (smaller panel, Figure 3-11). This suggests that lancelet HD4 may have given rise to lancelet HD6, sea squirt HD2, and sea lamprey HD4. Furthermore, the strong relationship of sea lamprey HD4 and HD5 suggests that sea lamprey HD4 gave rise to HD5. Further up the same branch, lancelet HD5 and HD2 show grouping with gnathostome Zhx3 HD4, suggesting that lancelet HD4 also gave rise to HD2 and HD5. However, given limitations within this dataset, we cannot determine whether HD1, HD3 or HD4 from lancelet was the original Zhx HD.

DISCUSSION

Zhx proteins are found throughout all vertebrates analyzed to date. In the present study, we identify and characterize Zhx gene structures and proteins from early chordate species, extending our current knowledge about the emergence of Zhx genes. We find that sea lamprey Zhx3L (Zhx3-like gene) is expressed across five different tissues as two alternatively spliced transcripts. Alternative splicing has been proposed to contribute to the evolution of novel phenotypes by providing alternative protein isoforms and opportunity for the evolution of new functions (Bush et al., 2017). Transcript isoforms
can be viewed as “internal paralogs” in the same gene, and these paralogs can have different functions similar to neofunctionalization of duplicated genes (Bush et al., 2017). Three different models relating evolution of alternative splicing and gene duplication have been proposed. The independent model proposes that transcript isoforms between paralogous genes and non-duplicated genes are similar (i.e. there is no relationship between gene duplication and alternative splicing) (Bush et al., 2017). The functional sharing model establishes that subfunctionalization of paralogous genes occur where one paralog expresses alternative transcript isoforms that perform a first set of ancestral functions, whereas the other paralog expresses alternative transcript isoforms that perform a second set of ancestral functions. The accelerated alternative splicing model proposes that alternative splicing is decreased in large families but increased in small families compared to singleton genes (Reddy et al., 2013). This model suggests that singleton genes that exhibit alternative splicing are likely to be duplicated. In addition, it is proposed that large families within the accelerated alternative splicing model would have no need for alternative splicing because new functions could be acquired by different family members. Based on the accelerated alternative splicing model, alternative splicing may precede gene duplication. To date, no studies have reported alternatively spliced isoforms across gnathostome Zhx coding regions. It is tempting to consider whether Zhx3L alternative splicing is indicative of a duplication event in early vertebrates (via an ancestral Zhx3-1L gene giving rise to Zhx3L and Zhx1L). A previous study by Chen et al. investigated the relationship of alternative splicing across 47 eukaryotic species and found that alternative splicing has increased over 1,400 million years; mammals and birds were found to have the highest proportion of genes that were
alternatively spliced (Chen et al., 2014). However, within the Zhx gene family, alternative splicing has not increased based on data from mouse and sea lamprey. This suggests that a closer examination of alternative splicing across multiple species is needed. In addition, Chen et al. did not include sea lamprey, or early chordates within their analysis. Perhaps inclusion of the new sea lamprey assembly within a comprehensive analysis of alternative splicing, combined with gene duplication across gene families would reveal nuances to whole genome duplication events.

Utilizing multiple sequence alignments and distance tree building, we identified the high conservation of the amino terminus, homeodomain, and zinc finger regions across gnathostome Zhx proteins. The importance of these conserved domains for protein function, homodimerization or heterodimerization, or post-translational modifications remains unknown. Zhx proteins are unique in that they contain both zinc fingers and homeodomains, both of which are important for DNA binding transcription factors. In addition, they contain multiple homeodomains which is also unique among homeodomain proteins. A central issue is how homeodomain transcription factors select their target genes in the genome using a DNA binding domain with limited sequence specificity; they recognize short and similar AT rich sequences in vitro (Bobola and Merabet, 2017). Although Zhx binding motifs across the genome remain to be identified, it is plausible that multiple Zhx homeodomains allow for higher specificity of DNA binding sites within the genome. Furthermore, Zhx proteins have been shown to homo- and hetero-dimerize with one another and NF-YA, which may add greater specificity for gene targeting. Alternatively, not all homeodomains may be needed in order to effectively localize Zhx proteins to gene targets. Analysis of Zhx homeodomain regions
revealed the highest conservation of HD1 alpha helix 3 across all gnathostome Zhx proteins. This suggests that this region is important for all Zhx protein function. Furthermore, the variability among other HD regions both within and among Zhx proteins suggests further fine tuning of gene targeting by all of these proteins, at least in regards to their homeodomains.

Interestingly, we also found that Zhx1 and Zhx2 share high conservation of their zinc finger regions compared to Zhx3. The zinc finger domain is the most frequently utilized DNA binding motif found in eukaryotic transcription factors (Cassandri et al., 2017). Binding of the zinc finger domain relies upon close contact of a few amino acids on the surface of the alpha helix with three base pairs in the major groove of DNA (Peng et al., 2014). The high conservation of Zhx1 and Zhx2 ZFs, this suggests that Zhx1 and Zhx2 may regulate similar gene targets. Taken together, specificity of Zhx proteins are most likely enhanced by a combination of ZF and HD specificities, however, future studies will need to address the importance of these domains for potential DNA binding and gene target regulation.

Based on phylogenetic analyses within this chapter, we propose a model of Zhx evolution (Figure 3-12). We propose that an ancestral Zhx3-1L (combined Zhx3L and Zhx1L gene) is the original gene that arose in the common chordate ancestor. Additionally, because our Bayesian analysis showed that lancelet HD4 grouped with lancelet HD6, HD2, and HD5, we propose that HD4 is the primordial HD that gave rise to all other Zhx HDs through a series of duplication events. We suggest that lancelet HD1 and HD3 diverged from other HDs early in chordate evolution, which explains why these HDs do not group with other HDs in our Bayesian HD tree. Alternatively, it remains
possible that lancelet HD1 or HD3 gave rise to HD4 which in turn gave rise to other HDs. Because sea squirt HD2 groups with lancelet HD6, and sea squirt HD1 groups with Zhx1 HD1, we propose that all other HDs and ZF2 were lost in this lineage. Our data show that sea lamprey Zhx contains two ZFs and five HDs. Bayesian phylogenetic analyses showed that sea lamprey HD4 and HD5 group with one another compared to other sea lamprey HDs, suggesting that sea lamprey HD4 gave rise to HD5, or that HD4 and HD5 duplicated after sea lamprey divergence. Furthermore, lancelet HD5 is considered a remnant HD5, due to its lack of conserved residues and contains a unique amino acid that is not present in other remnant HDs (L16I). This data suggests that sea lamprey HD5 arose independently after lancelet and ciona lineages, or that lancelet HD5 has acquired significant mutations. In addition, our MSA of full length Zhx did not show alignment of sea lamprey HD5 with lancelet HD5, and lancelet HD5 does not group with sea lamprey HD5. This data could also suggest that HD5 had been lost in Zhx1L and Zhx2L genes before Zhx1L and Zhx2L fusion in the sea lamprey lineage, or that HD4 and HD5 in sea lamprey acquired significant mutations. Additional analyses of the Zhx1L and Zhx2L region (ZhxP) is needed to verify that these current annotations are accurate. Zhx3-1L duplication and subsequent evolution to ancestral Zhx3L and Zhx2-1L genes, followed by Zhx2-1L duplication and evolution to Zhx2L and Zhx1L may have occurred prior to the divergence of the sea lamprey lineage; ZhxP HD3 is most related to Zhx3 HD4, and ZhxP HD2 and HD1 are most related to Zhx2. This suggests that Zhx1, Zhx2, and Zhx3 may have been, at least at one time, present in sea lamprey. Sea lamprey Zhx2L and Zhx1L gene fusion resulted in loss of HD3 and HD4 from Zhx2L and loss of HD1, HD2, and HD3 from Zhx1L.
Zhx1 has undergone changes in both HD3 and HD5 in the mammalian lineage that have resulted in a change in the conserved HD position 53. Previous to this, Zhx1 HD5 shows loss of conservation from the elephant shark lineage forward, albeit some shared and different changes to HD5 position 53 have occurred in different lineages. In contrast to Zhx1, Zhx2 HD5 has been completely lost from the bird lineage onward, whereas Zhx3 has only undergone changes to Zhx3 HD5 throughout gnathostome evolution. These data suggest that progressive loss of HD5 may also be occurring for Zhx1 and Zhx3. In addition, all of these homeodomains may be entirely functional within these species, or changes within these homeodomains may have no impact whatsoever on the overall functionality of Zhx proteins. Future experiments will need to address the functionality of Zhx proteins in binding specificity to gene targets.
Figure 3-1. A phylogenetic tree of chordate evolution within the context of geologic time. Timing of chordate radiation events based on divergence time estimates using amino acid substitution rates (Blair and Hedges, 2005; Kumar and Hedges, 1998). Time estimates are shown for each species branch (on dotted lines). Divergence times are indicated in millions of years. Groupings of gnathostomes (jawed vertebrates) and all vertebrates are shown to the far right. A geologic timescale is also included and shows the Precambrian eon, the Cambrian explosion (CE), and Paleozoic, Mesozoic and Cenozoic (CZ) eras.
Figure 3-2. Characterization of the sea lamprey Zhx gene. A. To determine the complete Zhx gene structure, 5′/3′ RACE and RT-PCR were performed using adult lamprey liver RNA, followed by sequencing of amplicons. The resulting Zhx gene spans 22,340 nucleotides and contains 8 exons (gray boxes). Sea lamprey alternatively spliced transcripts are shown below the Zhx gene region and total transcript lengths are indicated on the far right (3491 and 3416 nts). Primers (red triangles) were used to amplify a region of Zhx transcripts that had been previously shown to have two spliced isoforms in liver. Sizes of the RT-PCR amplicons generated from each primer set are shown (bp). An alternatively spliced isoform of Zhx removes the 3′ end of exon 4. Diagram is to scale. B. Sea lamprey Zhx has alternatively spliced transcripts in multiple sea lamprey tissues. RNA from two adult sea lamprey liver (L), kidney (K), muscle (M), blood (B), and eye (E) was prepared and RT-PCR using primers for EFA1 (top panel) or that span the alternatively processed Zhx exon 4 (lower panel) was performed. Migration of the 100 bp
ladder is shown at left. C. Log valued cycle number (left) and melting peaks (right) across two adult lamprey tissues for elongation factor 1-alpha (EF1A) shows differences in abundance of EF1A in different tissues based on the log linear phase of amplification, and a strong, singular melting peak at 88°C indicating specificity of primers. Liver (red), blue (kidney), green (muscle), black (blood), eye (purple).
Figure 3-3. The 2018 sea lamprey genome assembly reveals a predicted Zhx pseudogene. A. Screenshot of the annotated “Zhx2” and “Zhx1” genes from the sea lamprey \textit{(Petromyzon marinus)} SIMR database genome browser (genomes.stowers.org). RNA-sequencing track selected from st24b stage of lamprey development (late stage embryo development, neural crest migration) shows alignment of RNA-sequencing peaks aligned to the current Zhx2 and Zhx1 annotations (Smith et al., 2018) B. Zoomed in view from the \textit{P. marinus} SIMR database genome browser between the 3’ end of the annotated “Zhx2” gene, and the 5’ end of the annotated “Zhx1” gene shows a 27 bp gap between the “Zhx2” and “Zhx1”. C. Manual annotation the Zhx2/1 region as a single predicted Zhx pseudogene. The 27 bp gap between “Zhx2” and “Zhx1” was included within exon 6 of the predicted single pseudogene, with the “Zhx2” stop codon shown. With this manual annotation, the sea lamprey pseudogene contains 9 exons and spans 6,269 bp. Primers for RT-PCR and PCR were designed against both annotated Zhx2 and Zhx1 gene regions. Three primer sets (red, blue, and orange arrows) were designed against the Zhx2/1
annotated region using NCBI Primer-Blast and manually checked against the sea lamprey SIMR database Blastn feature for non-specific binding and used for RT-PCR and PCR analysis. Numbers next to primer pairs indicate predicted amplicon length from PCR of genomic DNA and RT-PCR of mRNA, respectively; for the orange primer pair set that spans the 27 bp gap junction, both PCR and RT-PCR sizes are predicted to be the same. Diagram is to scale.
Figure 3-4. Zhx genes in lower chordates show dramatically different structures.

The complete sea lamprey Zhx gene was determined based on RT-PCR and 3’/5’-RACE and contains a single 5’ non-coding exon (light grey box) and seven coding exons (gray boxes). The location of the two ZFs (dark gray boxes) and five HDs (black boxes with numerals) in the sea lamprey protein are shown. Amino acid number in the Zhx protein are shown as numbers above exons starting in Exon 2. Intron lengths (bp) are depicted below the gene. Alternative splicing of exon 4 (numbered 639 above exon 4) results in loss of the first predicted six amino acids within HD3. The sea lamprey ZhxP gene is shown based on alignment with other Zhx proteins and contains two ZFs, 3 HDs, and a stop codon in exon 6 after HD2. Dotted lines before and after the depicted 5’ and 3’ exons indicates that we have not determined whether non-coding upstream or downstream exons exist. Intron lengths are denoted beneath the gene, with asterisks (*) indicating the presence of multiple stop codons. Protein aa’s are shown above each exon. Numbering of amino acids is as shown for the lamprey Zhx protein. BLASTn against the
current NCBI lancelet (*Branchiostoma floridae*) and sea squirt (*Ciona intestinalis*) assemblies revealed the presence of a single *Zhx* gene. The lancelet *Zhx* protein contains five homeodomains and one HD that does not contain conserved residues (yellow box 5). The predicted lancelet protein is 1394 aa in length and contains most of its HD’s within one exon. In contrast to both sea lamprey and lancelet, the sea squirt *Zhx* coding region appears to be on a single exon, similar to *Zhx* gene structure seen in gnathostomes. However, the sea squirt *Zhx* gene contains one zinc finger and two HDs.
Figure 3-5. Amino terminus, zinc finger, and homeodomain regions are most conserved across gnathostome Zhx proteins. A. MUSCLE multiple sequence alignment (MSA) of Zhx proteins from 22 gnathostome species using Geneious software reveals several regions of high homology (Kearse et al., 2012). The highest homology is found within the first 12 amino acids (93.3%), followed by HD1 (69.5%). Conserved regions are marked and labeled above; N-terminal conserved aa’s (MAS), zinc finger region (ZFR, includes ZF1, ZF2 and intervening sequence), HD1, HD2, HD3, HD4, and HD5. Percent identity is shown as green, yellow, and red regions. Green indicates 100% identity, yellow represents identity between 99% and 30%, red indicates less than 30% identity. Length of protein alignment is shown above the percent identity and includes sequence gaps. B. N-terminus sequence showing high conservation among all gnathostome Zhx1, Zhx2, and Zhx3 proteins. MSA shows high conservation of the first 12 amino acids for all Zhx proteins (MASxRKSTTPCM). C. Neighbor-joining distance phylogenetic tree of Zhx full-length predicted proteins using MUSCLE MSA of 66
vertebrate sequences as described in methods and materials. Zhx1: red; Zhx2: blue; Zhx3: green. Alligator (Al), Anole (An), Chicken (Chic), Chimpanzee (Chim), Clawed Frog (Claw), Coelacanth (Coe), Cow (Cow), Dog (Dog), Elephant Shark (ES), Guinea Pig (GP), Horse (Hor), Human (Hum), Mouse (Mou), Minke Whale (MW), Opossum (Opp), Pig (Pig), Rabbit (Rab), Rat (Rat), Saker Falcon (SF), Spotted Gar (SG), Starling (Star), Turkey (T).
A.

Figure 3-6. MUSCLE MSA reveals alignment of lancelet, sea squirt, and sea lamprey ZFs and HDs to gnathostome ZFs and HDs. Multiple sequence alignment (MSA) with Zhx predicted proteins from 68 species (see appendix) was generated in Geneious 11.1.5 using MUSCLE. Zhx proteins were separately and manually annotated prior to MSA. The consensus sequence region is depicted above, as well as percent identity. Within the consensus sequence, darker gray regions on light gray boxes indicate higher consensus across all species. For simplicity, elephant shark (ES_Zhx) Zhx proteins are aligned to sea lamprey (SL_Zhx), ciona (CI_Zhx) and lancelet (BF_Zhx) due
to their placement within the chordate lineage. ZF regions (pink; include ZF1, ZF2 (both blue) and intervening region) show consistent alignment across all species. HD regions (green boxes; remnant HD regions are in yellow) are mainly aligned with several exceptions from lancelet and sea lamprey as noted in the text. Furthermore, although all are not shown here, four species (elephant shark, spotted gar, coelacanth, and clawed frog) may contain a remnant Zhx2 HD5 that shows conservation with all HD aa’s except R53. This suggests a progressive loss of Zhx2 HD5 throughout vertebrate evolution. B. MSA of sea lamprey Zhx protein to gnathostome Zhx proteins reveals partial conservation of the N-terminus Zhx signature.
Figure 3-7. **Determining the relationship of sea lamprey Zhx to gnathostome Zhx proteins.**

A. Zhx proteins were aligned using MUSCLE MSA followed by phylogenetic distance tree analysis (Geneious Tree Builder) with Jukes-Cantor genetic distance model and neighbor joining tree building. The consensus tree was built using 100 replicates for bootstrapping and majority greedy clustering, and with lancelet Zhx (BF_Zhx) used as an outgroup. Consensus support based on bootstrapping is shown at each branch base. Amino acid substitution length is shown at the bottom of each tree with a scale bar (0.3).

Lancelet used as an outgroup does not resolve the relationship of sea lamprey Zhx (SL_Zhx) to gnathostome Zhx proteins, but shows high consensus support (100%).
Monophyletic groupings of Zhx1 (red), Zhx2 (blue), and Zhx3 (green) are shown across gnathostomes, and Zhx1 and Zhx2 predicted proteins are more closely related to one another compared to Zhx3 based on groupings, which supports ZFR analysis. Both lancelet and lamprey show significant divergence of their Zhx proteins compared to gnathostome Zhx proteins. B. Bayesian phylogenetic analysis of MUSCLE MSA was also evaluated by distance tree analysis. Trees were run for 1,000,000 generations, where standard deviation of split frequency <.01 indicating that two separate, simultaneously run trees from the same dataset have converged. Shown is the consensus tree from combined simultaneous runs that were visualized in FigTree1.4.4 software. Posterior probability percentage is shown at each node with lancelet Zhx (BF_Zhx) used as an outgroup. Zhx1 and Zhx2 show closer association to each other than to Zhx3, in agreement with distance tree building. Alligator (Al), Anole (An), Chicken (Chic), Chimpanzee (Chim), Clawed Frog (Claw), Coelacanth (Coe), Cow (Cow), Dog (Dog), Elephant Shark (ES), Guinea Pig (GP), Horse (Hor), Human (Hum), Lancelet (BF), Mouse (Mou), Minke Whale (MW), Opossum (Opp), Pig (Pig), Rabbit (Rab), Rat (Rat), Saker Falcon (SF), Spotted Gar (SG), Sea Lamprey (SL), Starling (Star), Turkey (T).
Table 3.1. Percent pairwise identity of Zhx predicted protein zinc fingers and zinc finger regions.

<table>
<thead>
<tr>
<th>Zhx predicted protein</th>
<th>Zinc Finger 1</th>
<th>Zinc Finger 2</th>
<th>Zinc Finger Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhx1</td>
<td>94.9</td>
<td>93.3</td>
<td>94.3</td>
</tr>
<tr>
<td>Zhx2</td>
<td>92.4</td>
<td>88.1</td>
<td>91.9</td>
</tr>
<tr>
<td>Zhx3</td>
<td>65.7</td>
<td>66.8</td>
<td>65.8</td>
</tr>
</tbody>
</table>
Figure 3-8. Zinc finger MSA shows high divergence of Zhx3 ZFs compared to those of Zhx1 and Zhx2. MUSCLE MSA of Zhx zinc finger 1 (ZF1; 22 aa), zinc finger 2 (ZF2; 22 aa), zinc finger region (54 aa, entire region), and consensus sequences (90% consensus threshold) for each Zhx predicted protein reveals high similarity of Zhx1 and Zhx2 ZFRs compared to Zhx3 ZFR, suggesting a more recent divergence of Zhx1 and Zhx2 ZFRs. Black circles signify conserved ZF cysteines and histidines. Red circles signify additional conserved residues that span all Zhx proteins and may be unique to Zhx ZFRs.
Table 3.2. Amino acid matches between Zhx predicted protein ZF regions.

<table>
<thead>
<tr>
<th>Zhx proteins</th>
<th>ZF1</th>
<th>ZF2</th>
<th>ZF region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3</td>
<td>5</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>1,2</td>
<td>8</td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>1,3</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2,3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 3-9. Neighbor-distance joining phylogeny using zinc finger regions (ZFRs) places sea lamprey within the gnathostome Zhx3 lineage. ZFR regions across 22 gnathostome species, lancelet, and sea lamprey, were aligned using MUSCLE MSA followed by phylogenetic distance tree analysis (Geneious Tree Builder) with Jukes-Cantor genetic distance model and neighbor joining tree building. Consensus tree was built using 100 replicates for bootstrapping and majority greedy clustering, with lancelet Zhx (BF_Zhx) used as an outgroup. Consensus support based on bootstrapping is shown.
at each branch base. Amino acid substitution length is shown at the bottom of each tree with a scale bar (0.3). Sea lamprey Zhx ZFR is most closely related to gnathostome Zhx3 ZFRs. Zhx1 and Zhx2 ZFRs are most closely related to each other compared to Zhx3 ZFR, which also shows a higher sequence divergence across taxa. Alligator (Al), Anole (An), Chicken (Chic), Chimpanzee (Chim), Clawed Frog (Claw), Coelacanth (Coe), Cow (Cow), Dog (Dog), Elephant Shark (ES), Guinea Pig (GP), Horse (Hor), Human (Hum), Lancelet (BF), Mouse (Mou), Minke Whale (MW), Opossum (Opp), Pig (Pig), Rabbit (Rab), Rat (Rat), Saker Falcon (SF), Spotted Gar (SG), Sea Lamprey (SL), Starling (Star), Turkey (T).
Table 3.3. Percent pairwise identity of homeodomains across Zhx predicted proteins.

<table>
<thead>
<tr>
<th>Homeodomain</th>
<th>Percent Pairwise Identity</th>
<th>Percent Identical Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>69.5</td>
<td>36.7</td>
</tr>
<tr>
<td>2</td>
<td>60.6</td>
<td>26.7</td>
</tr>
<tr>
<td>3</td>
<td>63.3</td>
<td>16.7</td>
</tr>
<tr>
<td>4</td>
<td>58.0</td>
<td>11.7</td>
</tr>
<tr>
<td>5*</td>
<td>33.9</td>
<td>20.0</td>
</tr>
</tbody>
</table>

* Only includes Zhx3 from 3 species.
Figure 3-10. Gnathostome HD1 helix 3 is the most conserved region in Zhx homeodomains. MUSCLE MSA was conducted Zhx1 HD1, HD2, HD3 and HD4 for Zhx1, Zhx2 and Zhx3 [22 species in all cases except Zhx1 HD3 (10 species), Zhx2 HD4 (21 species) and Zhx3 HD4 (21 species)]. Consensus HDs were filtered for 90% identity across alignments. Amino acid number within the homeodomain is shown across the top, along with annotated regions for conserved alpha helices that are important for homeodomain function (Alpha1-3). Conservation between HD regions within different Zhx proteins is shown by highlighted residues (gray; conserved for all Zhx proteins, magenta; conserved between Zhx1 and Zhx2, cyan; conserved between Zhx2 and Zhx3, and yellow; conserved between Zhx1 and Zhx3). At the bottom, the percent conservation for each homeodomain (HD1, HD2, HD3 and HD4) within alpha helices across Zhx proteins is presented. Residues that are conserved among all HDs across all Zhx proteins are indicated by black circles and were filtered based on 90% identity by MSA of all consensus HDs. Amino acid residues that are critical conserved residues for
homeodomain function are indicated by red circles. Alpha helix 3 within HD1 is the most conserved HD region across all HDs of all Zhx proteins. Zhx remnant HD5 or HD5 is not included and only contained 5 matches between Zhx2 and Zhx3, and no conservation across all Zhx predicted proteins or similarity between Zhx1 and Zhx3.
Table 3.4. Amino acid matches between Zhx predicted protein HD regions.

<table>
<thead>
<tr>
<th>Zhx proteins</th>
<th>HD1</th>
<th>HD2</th>
<th>HD3</th>
<th>HD4</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3</td>
<td>27</td>
<td>21</td>
<td>20</td>
<td>15</td>
<td>83</td>
</tr>
<tr>
<td>1,2</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>1,3</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>2,3</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>Totals</td>
<td>44</td>
<td>35</td>
<td>34</td>
<td>36</td>
<td>--</td>
</tr>
</tbody>
</table>
Figure 3-11. Bayesian phylogeny of all Zhx homeodomains confirms Zhx3 as the primordial chordate Zhx gene. Bayesian phylogenetic analysis of MUSCLE MSA sequence alignment from 313 HDs across all Zhx proteins. Trees were run for 30,000,000 generations (standard deviation of split frequency = 0.008, see methods for more details). No outgroups were used for this analysis. Shown is the consensus tree from combined simultaneous runs that were visualized in FigTree1.4.4 software. All Zhx HDs are shown. Asterisks indicate lancelet HDs, black circles indicate sea squirt HDs, green circles indicate sea lamprey Zhx HDs, and blue circles indicate sea lamprey ZhxP HDs. Panel on the left is zoomed in to show the relationship of lancelet HD4 to sea lamprey HD4 and HD5, lancelet HD6, and sea squirt HD2. Percent probabilities are shown at the branch bases.
Figure 3-12. Proposed model of Zhx and HDs throughout chordate evolution. Taxon clades containing Zhx proteins are shown with their representative Zhx protein structures. Blue boxes represent zinc fingers, whereas green boxes indicate homeodomains. Starting with the common chordate ancestor, a common ancestral Zhx3-1L HD4 is proposed to have given rise to all current six HDs (green, numbered) that are present in lancelet. The yellow HD in lancelet represents a remnant HD that does not contain the conserved leucine at position 16 (L16I). Sea squirt Zhx3-1L contains two HDs, the first of which (HD1) is most related to Zhx1 HD1 and the second (HD1) which is most closely related to lancelet HD6. The second ZF (not shown) and other HDs (black striped) were lost in
this lineage. The 1R-2R hypothesis suggests that two whole genome duplication events happened prior to sea lamprey divergence. An ancestral \textit{Zhx3-1L} containing two zinc fingers and five homeodomains, underwent duplication, via whole genome duplication, to give rise to another \textit{Zhx3-1L} gene. This secondary \textit{Zhx3-1L} then acquired mutations which allowed it to become an ancestral \textit{Zhx2-1L} gene. \textit{Zhx2-1L} is proposed to have duplicated locally to give rise to two \textit{Zhx2-1L} genes. These genes then acquired mutations and became \textit{Zhx1L} and \textit{Zhx2L} genes. Sea lamprey contain \textit{Zhx3L} and a pseudogene (\textit{ZhxP}) that could be the result of a \textit{Zhx2L-Zhx1L} gene fusion. Sea lamprey \textit{Zhx3L} HD4 and HD5 are most related to each other, and are not related to lancelet HD5 or HD6, suggesting that either an independent duplication event in the sea lamprey lineage may have given rise to these related HDs, or they acquired significant mutations in this lineage making them more similar. Early gnathostome \textit{Zhx} proteins, including fish and amphibians, contain five HDs, including remnant HDs, for all \textit{Zhx} proteins.
CHAPTER IV

ZHXY2 REGULATES SEX-LIMITED PROTEIN (Slp) GENE EXPRESSION IN
MALE LIVER

The mammalian liver is a sexually dimorphic organ and performs essential functions such as metabolic regulation, hormone secretion, cholesterol homeostasis and drug metabolism (Si-Tayeb et al., 2010). Sex-biased liver gene expression occurs in rat, mice, and humans; approximately ~1000 genes in rat, ~1600 genes in mouse, and ~1200 genes in human are sex-biased and contribute to gender differences in metabolism as well as disease (Wauthier et al., 2010; Wauthier and Waxman, 2008; Zhang et al., 2011). For example, males are more than two-fold likely to develop liver cancer compared to females in both rodents and humans (Zheng et al., 2018). The combined gender disparity in liver cancer development and significant sex-biased liver gene expression differences between adult males and females prompt the need to investigate the molecular mechanisms governing the control of sex-biased genes. Importantly, these studies may identify factors that predispose males to liver cancer and lead to new therapeutic targets for this and other sex-biased cancers as well as other diseases.

Liver sexual dimorphism is determined by sex differences in circulating growth hormone (GH) profiles (Wauthier et al., 2010). In humans and rodents, GH secretion from the anterior pituitary occurs in bursts. In mice, GH is secreted in episodic bursts with peak plasma concentrations reaching almost 100 ng/ml in both sexes (MacLeod et al., 1991). Females have frequent bursts occurring every 15 to 45 minutes, leading to
relatively consistent and higher average plasma GH levels compared to males. In males, GH bursts occur every 2-2.5 hours, leading to greater variation of and lower average plasma GH levels compared to females (Adams et al., 2015; MacLeod et al., 1991).

In male mice, GH pulses activate Stat5b, a transcription factor required for sex-biased liver gene expression (Waxman and Holloway, 2009). Stat5b knockout (Stat5b\(^{-/-}\)) male mice lose male-specific liver expression; male-specific liver RNAs, including Major urinary proteins (Mups) and sex-limited protein (Slp) are decreased whereas female-specific RNAs, including cytochrome p450 2b9 (Cyp2b9) and Cyp2a4, are increased compared to wild-type mice (Holloway et al., 2006). In contrast, male-specific RNAs and female-specific RNAs remain largely unaffected in Stat5b\(^{-/-}\) females (Clodfelter et al., 2006). In addition to Stat5b, liver-specific deletion of the transcription factor HNF4\(\alpha\) causes dysregulation of 646 genes that are normally sex-biased between males and female mice (in total 864 sex-specific genes) (Holloway et al., 2008). By comparing genes whose expression changes when Stat5b and HNF4\(\alpha\) are deleted in male liver (expression ratio >1.25, \(p<.05\)), 176 sex-biased genes were found to be dependent on both regulators, including Elolv3, Slp, Mups, Cyp2b9, Cyp2b13, Cyp4a14 (Holloway et al., 2006; Holloway et al., 2008).

*Zinc finger and homeoboxes 2 (Zhx2)* is a member of a small gene family that contains Zhx1 and Zhx3. Zhx2 contains two amino-terminal C\(_2\)-H\(_2\) zinc fingers and four carboxy-terminal homeodomains; these domains suggest that Zhx proteins are involved in nucleic acid binding (Burglin and Affolter, 2016; Spear et al., 2006). Zhx2 loss leads to dysregulated expression of sex-biased genes in male liver. Specifically, Zhx2 deletion leads to upregulation of female-biased cytochrome p450 (Cyp) mRNA expression in
male liver including Cyp2a4, Cyp2b9, and Cyp2b13, and downregulation of male-biased expression including Mups and Elovl3 (Creasy, 2015; Creasy et al., 2016; Jiang et al., 2017).

The overlap of Stat5b, HNF4α, and Zhx2 sex-biased targets led us to consider whether these factors act to co-regulate sex-biased genes within male liver. We demonstrate that Zhx2 regulates Slp, a highly male-biased gene also known to be regulated by both Stat5b and HNF4α. We also characterize Slp expression in the postnatal male and female liver. Furthermore, we find that the absence of Zhx2 does not alter mRNA levels of the known sex-biased transcription factors Stat5b, Stat5a, HNF4α, and Cux2. We also show that a known Stat5b direct target, Igf-1, is unaffected in Zhx2<sup>DHep</sup> livers. Using transient co-transfections, we find that the Slp promoter is regulated by Zhx2 and Stat5b. We also expand our view of genes regulated by Zhx2, Stat5b, and HNF4α by aligning previously completed microarrays to identify overlapping targets. We find that stringent sex-biased genes are often significantly dysregulated in both Stat5b and HNF4α knockout/knockdown male livers, whereas a much smaller proportion of sex-biased genes are dysregulated between Zhx2 and Stat5b, or Zhx2 and HNF4α male livers. Surprisingly, we also discovered that the majority of stringently sex-biased genes that are dysregulated in BALB/cJ mice, which have a natural hypomorphic mutation in their Zhx2 gene, are also dysregulated in Stat5b and HNF4α knockout/knockdown livers. Taken together, these data indicate that the majority of genes dysregulated in the absence of Zhx2 are sex-independent genes, and that sex-biased genes dysregulated in BALB/cJ livers, including Slp, may be co-regulated by Stat5b and HNF4α.
RESULTS

*Slp and C4 genes are highly similar.* Previous results from our lab have demonstrated loss of male-biased expression patterns in Zhx2^{ΔHep} male liver.

For example, mRNA levels of male-biased genes such as Elovl3 and Mups are significantly lower in Zhx2^{ΔHep} compared to floxed Zhx2^{fl/fl} male liver. In addition, female-biased Cyps including Cyp2a4, Cyp2b9, Cyp2b13, Cyp4a10, and Cyp39a1 exhibit significantly elevated mRNA levels in male Zhx2^{ΔHep} compared to male Zhx2^{fl/fl} liver.

To expand upon our data showing dysregulated expression of sex-biased genes in male Zhx2-deficient liver, we examined whether the absence of Zhx2 affected Slp expression. Slp, which arose from a duplication of the C4 gene (Robins and Samuelson, 1992), is highly expressed in male liver but not female liver of numerous mouse strains, including C57BL/6J (Passmore and Shreffler, 1970). Interestingly, C4 does not exhibit sex-biased expression in the liver (Robins and Samuelson, 1992). Furthermore, livers of male mice that lack Stat5b or HNF4α have significantly lower Slp mRNA levels compared to normal male liver (Holloway et al., 2006).

To design specific primers for Slp mRNA, we first analyzed the sequence similarity of Slp (otherwise known as C4a) and C4 (otherwise known as C4b) using Geneious software pairwise alignments. Alignment of Slp and C4, including 10 kB upstream and downstream of each gene, revealed that roughly 1.8 kB upstream of the transcription start sites and ~1 kB downstream of transcriptional stop sites are highly similar; sequence homology drops off considerably outside these regions (Figure 4-1A). Our alignments identified several gaps exist upstream, within, and downstream of genes.
The mRNAs from these two genes are very similar, exhibiting 92.6% pairwise identity (Figure 4-1B). Using this information, we designed gene-specific primers to quantitate Slp and C4 mRNA levels.

**Slp mRNA levels are significantly reduced in Zhx2 liver-specific knockout male liver.** To test the impact of Zhx2 loss on Slp and C4 expression, RT-qPCR was used to analyze RNA from Zhx2fl/fl and Zhx2ΔHep male and female liver. (Figure 4-2A). Hepatic Zhx2 levels are dramatically lower in male and female Zhx2ΔHep mice compared to Zhx2fl/fl mice; low Zhx2 levels Zhx2ΔHep mice are due to continued Zhx2 expression in non-parenchymal liver cells (Jiang et al., 2017). We next measured Slp and C4 expression. (Figure 4-2B and C). In Zhx2ΔHep male liver, Slp mRNA levels were reduced to almost undetectable levels and remained nearly undetectable in both Zhx2ΔHep and Zhx2fl/fl female liver (Figure 4-2B). In contrast, C4 mRNA levels were not significantly affected by loss of Zhx2 in either male or female mice (Figure 4-2C). Thus, Zhx2 contributes to high Slp expression in adult male liver.

**Sex-biased Slp expression is established in the postnatal liver.** Sex differences in mouse liver gene expression are established primarily after four weeks of age, although sex-biased expression of several genes is established prior to this (Conforto and Waxman, 2012). To determine the developmental profile of Slp in male and female mice, we measured Slp mRNA levels beginning at embryonic day 17.5 by RT-qPCR (e17.5) (Figure 4-3A). Slp expression does not significantly differ between males and females
until after postnatal day 28 (p28). In male liver between the ages of four and eight weeks, Slp mRNA levels increase roughly 100-fold. In contrast, in female liver, Slp mRNA levels are low and do not significantly change during maturation. Furthermore, Zhx2 mRNA levels do not significantly differ between males and females from birth to 8 weeks of age (Figure 4-3B). This data suggests that male-biased Slp expression is established between four and eight weeks after birth.

**Sex-biased transcription factor gene expression is unaltered in Zhx2 liver-specific knockout mouse liver.** Since Zhx2 expression is not sex-biased, yet the absence of Zhx2 results in significantly lower male Slp mRNA levels, we explored whether the Zhx2 loss might affect mRNA levels of other transcription factors that are known to regulate sex-biased liver gene expression. To test this, we measured Stat5b, Stat5a, HNF4α, Bcl6, and Cux2 mRNA levels in Zhx2^{fl/fl} and Zhx2^{A^Hep} male and female liver. Previous results from our lab showed a modest, but significant reduction of Stat5b mRNA levels and lower, but not significant decrease in Stat5a mRNA levels in Zhx2^{A^Hep} compared to Zhx2^{fl/fl} male liver (Creasy, 2015). My analysis with the primers used in these experiments, resulted in multiple non-specific products, despite efforts to optimize RT-qPCR conditions (data not shown). Therefore, new Stat5b and Stat5a RT-qPCR primers were designed. Using these new primers, Stat5b and Stat5a mRNA levels were found to remain the same regardless of Zhx2 status (Figure 4-4A and B). Additionally, HNF4α levels were unaffected by Zhx2 loss (Figure-4C). We found that Stat5b, Stat5a, and HNF4α mRNA levels were the same in male and female mice. HNF4α mRNA levels were not significantly different between Zhx2^{fl/fl} male and females consistent with previous reports (Holloway et al.,
2008). Stat5b expression was also higher compared to Stat5a expression across all datasets, which is consistent with previous work (Stat5a levels >10 fold difference compared to Stat5b levels in mouse liver) (Dr. David Waxman, personal communication). However, microarray datasets have shown that Stat5a mRNA levels are roughly two fold higher in females versus males which is inconsistent with our data presented here (Clodfelter et al., 2006). Comparisons between the Stat5a probe used to measure Stat5a mRNA levels by Clodfelter et al. and our Stat5a primers using BLAT search within the UCSC Genome Browser did not identify any differences between gene targets (Dr. David Waxman, personal communication). Strain-specific differences, including Stat5a expression levels or transcript variants, may contribute to these observed differences. In contrast to Stat5b, Stat5a, and HNF4α, Bcl6 and Cux2 show male-biased and female-biased gene expression in Zhx2^ΔHep^ mice, respectively (Figure 4-4D and E). However, in Zhx2^ΔHep^ mice compared to the Zhx2^fl/fl^ controls, there are no significant differences in Bcl6 or Cux2 mRNA levels. Thus, we conclude that Zhx2 does not regulate expression of known sex-biased regulators in the liver.

**Zhx2 positively regulates the Slp promote.** Previous studies have shown that Slp expression is positively regulated by GH in male mice. Furthermore, Slp expression is drastically reduced in male livers deficient in Stat5b, the downstream effector for GH signaling. Consistent with this data, Stat5b binding sites have been identified in the Slp promoter region (Varin-Blank et al., 1998). Because Slp expression is also greatly reduced in the absence of Zhx2, even though Stat5b mRNA levels are unchanged, we
hypothesized that Zhx2 may independently positively regulate Slp expression, potentially through the promoter region.

To determine whether Zhx2 directly regulates Slp gene expression, we cloned fragments of the Slp promoter upstream of the firefly luciferase in the pGL14.4 expression vector. These fragment extended 1500 bp or 2000 bp upstream of the Slp transcription start site (+1); both ended 20 bp downstream of +1 (Figure 4-5). The Slp 2000 promoter contains two gamma interferon-activated sequence (GAS) elements that have been shown to bind activated Stat5b prepared from liver nuclear extracts (Varin-Blank et al., 1998). In addition, we analyzed the Slp promoter region using JASPAR transcription factor binding prediction software for other potential Stat5b:Stat5a binding sites (Khan et al., 2018). In addition to the two experimentally-determined sites, two additional high-scoring candidate Stat5b:Stat5a binding sites were predicted at about -600 bp, relative to +1, within the Slp promoter.

To test whether Zhx2 regulates the Slp promoter, the mouse hepatoma cell line Hepa 1.6 was transiently co-transfected with the Slp promoter-containing luciferase plasmids and a Zhx2 expression vector. Plasmids containing the herpes simplex virus thymidine kinase (HSVtk) minimal promoter fused to Renilla luciferase were also transfected to normalize for variation in transfection efficiency. The control promoterless firefly luciferase vector (pGL4.14) shows minimal luciferase activity when transfected with either control pcDNA3.1 empty vector (EV) or the Zhx2 expression plasmid (Figure 4-6). The Slp promoter plasmids (1500 and 2000) both had significantly higher basal luciferase activity compared to pGL4.14. When Zhx2 was co-expressed with the Slp promoter-luciferase plasmids, the Slp 2000 plasmid was activated whereas the Slp 1500
plasmid did not respond. Although statistically significant, Zhx2 only activated the Slp 2000 plasmid 1.75 fold compared to the EV control. These data suggest that Zhx2 may positively regulate Slp gene expression through upstream elements that are not present in the 1500 Slp promoter.

**Zhx2 and Stat5b positively regulate the Slp 2000 promoter.** Because the Slp 2000 promoter plasmid is activated by Zhx2 and contains experimentally identified Stat5b GAS binding sites, we wanted to determine whether Zhx2 might cooperate with Stat5b to activate Slp expression. We obtained the Stat5b-activateable reporter plasmid 4XNTCP-Luc (gift from Dr. Young-Kwon Hong). The 4XNTCP-Luc plasmid contains four GAS elements derived from the rat Na+/taurocholate cotransporting poly-peptide (NTCP) gene upstream of a HSVtk promoter in the pT109Luc firefly luciferase plasmid (Figure 4-7A) (Ganguly et al., 1997). We also received two Stat5b expression plasmids (gifts from Dr. David Waxman); Stat5b pME18S containing normal mouse Stat5b that requires GH stimulation for nuclear translocation and subsequent gene activation, whereas Stat5b* pCI contains mouse Stat5b with two site-specific mutations, H299R and S711F, that render it constitutively phosphorylated on tyrosine 699 and transcriptionally active in the absence of GH (Shipley and Waxman, 2003). To ensure these proteins would be expressed to similar levels, we cloned the Stat5b* cDNA from the pCI plasmid backbone into the pME18S backbone. Both Stat5b and Stat5b* pME18S were subsequently sequenced to verify site specific mutation differences between normal and constitutively active Stat5b. Without exogenous GH stimulation, we expect Stat5b* but not Stat5b to activate the 4XNTCP plasmid.
As predicted, when 4XNTCP was co-transfected with Stat5b into Hepa 1.6 cells, luciferase activity was not increased compared to 4XNTCP co-transfected with the empty control pcDNA vector (EV). In contrast, co-transfection with the constitutively active Stat5b* increased 4XNTCP luciferase activity 3.8-fold compared to cells transfected with EV (Figure 4-7B).

To test whether Zhx2 and Stat5 cooperatively regulate the Slp2000 promoter, co-transfections with different combinations of EV, Stat5b, Stat5b* and Zhx2 were used as drivers of luciferase activation. Transfection of the promoterless pGL4.14 with any driver plasmid did not result in significant differences in luciferase activity (Figure 4-8). However, co-transfections of the Slp2000 promoter with Stat5b or Stat5b* resulted in significantly higher luciferase activity, albeit fold changes were not impressive (1.7- and 1.3-fold, respectively). Surprisingly, Stat5b co-transfections showed higher reporter luciferase activity compared to Stat5b*. In contrast to our earlier results, co-transfection with Zhx2 did not result in significantly higher Slp luciferase activity and there was a lower-fold difference between Zhx2 and EV control transfections (~1.2 fold). Although our previous results were significant, we note that this previous activation was modest (~1.7 fold). Additional experiments are needed in order to reduce possible error as these current datasets were only completed at most three times. Co-transfection with Stat5b and Zhx2 or Stat5b* and Zhx2 with Slp2000 increased luciferase activity 2.75- or 2.1-fold, respectively, compared to EV. These data suggest that Stat5b and Zhx2 cooperatively activate the Slp2000 promoter. However, we note that the fold activation are modest and are concerned by the fact that Stat5b was more potent than Stat5b* in these co-transfections.
Igf-1 mRNA levels are not altered in Zhx2 liver-specific knockout mouse liver. To test whether loss of Zhx2 affects another Stat5b target in mouse liver, RNA was isolated and RT-qPCR was performed using Igf-1-specific primers. Igf-1 is a direct target of Stat5b and liver-specific knockout of Stat5b/Stat5a in males reduces Igf-1 gene expression (Cui et al., 2007; Woelfle and Rotwein, 2004). Igf-1 mRNA levels were unaffected by loss of Zhx2 in male or female liver (Figure 4-9). This data suggests that Stat5b function is not altered in Zhx2<sup>ΔHep</sup> male liver.

**Zhx2, Stat5b, and HNF4α overlapping regulated genes.** Because Slp is known to be regulated by Stat5b (Holloway et al., 2006) and Zhx2 (Figure 4-2), and since our previous studies demonstrated that many sex-biased genes are dysregulated in the livers of male mice in the absence of Stat5b, HNF4α or Zhx2, we performed a more global analysis of sex-biased genes that could be coordinately regulated by these three factors. To accomplish this, we first downloaded microarray datasets for Zhx2 (Gargalovic et al., 2010), Stat5b (Clodfelter et al., 2006), and HNF4α (Holloway et al., 2008). The Zhx2 microarray compared hepatic differences in gene expression between male Sub13 (Zhx2 WT) and BALB/cJ mice (mutated Zhx2). The Stat5b microarray compared differences in gene expression between male and female 129 x BALB/c floxed mice livers and those containing a disrupted Stat5b gene (Stat5b<sup>−/+</sup>). The HNF4α microarray compared differences in gene expression between male and female C57BL/6 mice containing floxed HNF4α or liver-specific HNF4α KO mice livers (HNF4α<sup>ΔHep</sup>). Stat5b<sup>−/−</sup> mice are
shown to be deficient for Stat5b at both the mRNA and protein level for all tissues examined (Udy et al., 1997). HNF4α mice contain LoxP sites surrounding HNF4α exon 4 and exon 5; breeding with albumin-Cre mice removes exon 4 and exon 5 within liver, resulting in a frame shift, and premature stop codon which eliminates HNF4α function (Hayhurst et al., 2001; Wiwi et al., 2004).

We first identified sex-biased genes common between both Stat5b and HNF4α datasets by comparing male floxed and female floxed mice. Between both datasets, we identified 105 genes that were sex-biased (Figure 4-10). Importantly, these genes represent stringent sex-biased genes and were found using strict criteria (male^{fl/fl}: female^{fl/fl} ratio >2.0, p-value <0.005). We used this criteria to identify genes that have more sex-bias overall, and are more likely to be consistently sex-biased across datasets. Even using these criteria, we identified 11 genes (from our original 116) that were inconsistent between the two datasets for sex-biased expression; multiple probes did not show consistent sex-bias, or sex-bias was switched between the two microarrays. We then used this list of common sex-biased genes to identify sex-biased genes that were analyzed in the Zhx2 microarray; 29 genes were identified. We also used the list of common sex-biased genes to re-identify sex-independent genes within Stat5b and HNF4α microarrays. Both sex-independent and sex-biased genes were then filtered for stringent (wildtype:knockdown/knockout ratio >2.0, p<0.005) or relaxed differential expression (wildtype:knockdown/knockout ratio >2.0, p<0.05) between Sub13/BALB/cJ, Stat5b^{β/β}/Stat5b^{−/−}, and HNF4α^{β/β}/HNF4α^{ΔHep} male mice.

Using these sets of criteria, we discovered that the majority of genes that are dysregulated by absence of Zhx2 are sex-independent (Figure 4-11A and B). Two sex-
biased genes that are dysregulated only in the absence of Zhx2 are Carbonic anhydrase 3 (Car3) and SET domain-containing 4 (Setd4). Car3 is a male-biased gene that is upregulated (2.1 fold) whereas Setd4 is a female-biased gene that is upregulated (2.05 fold) in BALB/cJ male livers. Previous studies have demonstrated that stable transfection of the hepatocellular carcinoma (HCC) cell line, SK-Hep1, with a Car3 expression vector enhances anchorage independent growth and invasiveness, both of which are important for cancer metastasis (Dai et al., 2008). Setd4 is a newly discovered methyltransferase which has been shown to have elevated mRNA expression in several breast cancer cell lines, including estrogen-receptor negative breast cancer (Faria et al., 2013).

Only one female-biased gene was found to be commonly dysregulated in the absence of both Zhx2 and Stat5b; microsomal glutathione S-transferase 3 (Mgst3). Mgst3 is upregulated (2.15, 1.84 fold by different probes) in BALB/cJ mice, and upregulated (2.3 fold) in Stat5b<sup>-/-</sup> males. Although previous studies have shown that Mgst3 is not the most abundant Mgst isoform within the liver, human MGST3 has been shown to detoxify foreign compounds within the liver and to synthesize leukotriene-C4, an inflammatory mediator (Hayes et al., 2005). Two male-biased genes were found to be commonly dysregulated by both Zhx2 and HNF4α in our relaxed differential dataset; melatonin receptor 1 alpha (Mtnr1a) and Trehalase (Treh) (Figure 4-11B). Both genes are downregulated in BALB/cJ (Mtnr1a, 3.46 fold; Treh, 211.9 fold) and HNF4α<sup>Hep</sup> (Mtnr1a, 9 fold; Treh, 5.75 fold) male livers. However, both transcripts show low mRNA abundance in mouse liver based on EMBL-EBI Expression Atlas data.

The majority of sex-biased genes that are differentially regulated in BALB/cJ male liver compared to Sub13 were also found to be dysregulated in both Stat5b<sup>-/-</sup> and
HNF4α<sup>ΔHep</sup> males (18 genes) (Figure 4-11B, Table 4.1). We have previously validated downregulation of Elovl3, Mup4, and Mup1 by RT-qPCR in Zhx2<sup>ΔHep</sup> and Zhx2<sup>−/−</sup> male liver, all of which are male-biased and concordantly downregulated in the absence of Zhx2, Stat5b, or HNF4α. Furthermore, previous results from our lab have demonstrated significant upregulation of the female-biased genes, Cyp2b13 and Cyp2b9, in Zhx<sup>ΔHep</sup> males, which also show concordant upregulation between all three microarrays. Initial experiments from our lab have shown that Cyp17a1 is downregulated in Zhx<sup>ΔHep</sup> males (data not shown), and studies from this chapter demonstrate that Cux2 is not upregulated in Zhx<sup>ΔHep</sup> male liver. These results highlight the importance of validating microarray datasets. Although the majority of sex-biased genes that are dysregulated in the absence of Zhx2 are also dysregulated in Sta5b and HNF4α deficient livers, a much larger population of sex-biased genes are commonly dysregulated between Stat5b- and HNF4α-deficient male livers but not in Zhx2-deficient livers (55 genes). These data suggest that Zhx2 may contribute to some sex-biased regulation, however, a larger portion of sex-biased regulation in the liver is dictated by Stat5b and HNF4α. Furthermore, a larger proportion of genes that are dysregulated in Stat5b<sup>−/−</sup> male liver show sex-bias gene expression compared to HNF4α.
DISCUSSION

*Slp* and *C4* are duplicated genes that are located within the S region of the mouse H-2 locus along with two other genes that function in the complement pathway, *C2* and factor B (*Bf*) (Robins, 2004). Evidence for *Slp* and *C4* gene duplication comes from their extensive homology in flanking as well as coding regions. This high amount of homology was reflected from our sequence analysis (Figure 4-1), where *Slp* and *C4* transcripts contain roughly 92% sequence identity over ~15 kB of sequence; this high homology was found to decrease substantially past ~1.8 kB upstream and about 1 kB downstream of the aligned genes.

Previous studies showed that two DNAse I-hypersensitive sites mapped specifically to *Slp* at 2 kB and 2.3 kB upstream of the transcriptional start site, and that these sites were inducible in females treated with testosterone (Hemenway and Robins, 1987). It was discovered that this upstream region of *Slp* contained an ancient endogenous provirus sequence which imparts testosterone regulation (Stavenhagen and Robins, 1988). Transgenic mice harboring a small 120 bp sequence (C’Δ9) from this region revealed that it contained hormonal response elements which imparted testosterone-mediated activation of *Slp* expression through the androgen receptor in tissues such as the kidney, testis, and prostate (Nelson and Robins, 1997a). However, in liver, this region was not sufficient to induce expression of the reporter gene. Although testosterone is required to indirectly stimulate expression of *Slp* in liver by chromatin remodeling, direct regulation of *Slp* in liver is mediated by growth hormone (Nelson and Robins, 1997b; Robins, 2004). Electrophoretic mobility shift assays later demonstrated that two sequences upstream of *Slp* (gamma interferon-activated sequences, or GAS) can
bind Stat5b from liver extracts of male mice (Varin-Blank et al., 1998). These results demonstrate differential regulation of the Slp promoter between kidney and liver.

RT-qPCR analysis of Slp and C4 mRNA levels in Zhx2$$^{β/β}$$ and Zhx2$$^{ΔHep}$$ adult mouse liver revealed that Slp is highly expressed in Zhx2$$^{β/β}$$ males compared to females. Zhx2$$^{ΔHep}$$ males had a significant reduction in hepatic Slp mRNA levels approaching the limit of detection, whereas C4 mRNA levels were unaffected by the Zhx loss. Both Slp and C4 mRNA levels were unaffected in Zhx2$$^{ΔHep}$$ female liver. We also demonstrate that Slp mRNA levels increase substantially in males between four and eight weeks old. Because we found that Zhx2 expression is sex-independent in the postnatal liver, we hypothesized that Zhx2 may dysregulate expression of known sex-biased regulators. RT-qPCR for Stat5a, Stat5b, HNF4α, Bcl6, and Cux2 show no differences in mRNA levels of these genes between Zhx2$$^{β/β}$$ and Zhx2$$^{ΔHep}$$ livers. We do note, however, that male Zhx2$$^{ΔHep}$$ Bcl6 mRNA levels are not significantly different compared to Zhx2$$^{β/β}$$ females, whereas Zhx2$$^{β/β}$$ males and females show a significant difference in their Bcl6 mRNA levels. Furthermore, although mRNA levels of these transcription factors were unaffected by the absence of Z hx2, we have not tested whether Zhx2 may act to regulate the function of these sex-biased regulators via protein-protein interactions.

To directly test Slp promoter regulation by Zhx2, we used luciferase constructs that contained promoter regions of the Slp gene that extend 1500 bp or 2000 bp upstream of Slp exon 1. The 2000 bp region contains two Stat5b binding sites identified by Varin-Blank and colleagues. Although Slp 2000 bp and Zhx2 co-transfection conferred a significant activation of luciferase activity compared to controls and the 1500 bp upstream region, the effect was modest. This effect of Zhx2 in vitro does not correlate to
the dramatic decrease in Slp mRNA levels observed with the absence of Zhx2 in vivo. These data suggest that Zhx2 acts through other mechanisms, either indirectly or directly and possibly in concert with other factors, to influence Slp mRNA expression in male liver. Since our expression vectors only contain 2000 bp of the Slp promoter, it is also possible that Zhx2 controls Slp expression through binding sites elsewhere in the Slp gene. Alternatively, Zhx2 may require additional factors that are present in vivo but absent from tissue culture systems tested here.

To identify whether additional cofactors were needed to induce additional activation of the Slp promoter by Zhx2 in vitro, we co-transfected with wild-type or constitutively active Stat5b. To our knowledge, these are the first results that have tested Slp promoter activation by Stat5b in vitro. Additionally, our experiments tested sites that were demonstrated to bind Stat5b by Varin-Blank and colleagues. To our surprise, although we were able to demonstrate activation of a well-known target (4xNTCP) by constitutively active Stat5b, co-transfections with Slp promoters and Stat5b showed mixed results. Slp co-transfections with constitutively active Stat5b versus inactive Stat5b, which requires exogenous GH activation for phosphodimerization and subsequent regulation of gene targets, yielded lower luciferase activity. These results may indicate the need to understand better the underlying growth hormone pathways that function within the Hepa 1.6 cell line.

Studies of Stat5b-deficient and HNF4α liver-deficient mice have identified a striking codependence of both transcription factors for sex-specific liver gene expression. Microarrays have demonstrated global loss of sex-specific liver gene expression in Stat5b−/− and HNF4ΔHep male mice (Clodfelter et al., 2006; Holloway et al., 2006;
Holloway et al., 2008). Recently, we have discovered striking similarities to sex-biased genes that are dysregulated in Zhx\textsuperscript{D Hep} males, whereas sex-biased gene expression in females remain largely unaffected by Zhx2 loss (Creasy et al., 2016; Jiang et al., 2017). Using previously published microarray datasets, we identified common dysregulated genes between Zhx2, Stat5b, and HNF4\textalpha. These data revealed that a smaller proportion of sex-biased genes are regulated by Zhx2 compared to Stat5b and HNF4\textalpha. However, these results may be a reflection of the microarray dataset sizes; two times and four times the number of genes were analyzed for the Stat5b and HNF4\textalpha microarrays compared to the Zhx2 microarray, respectively. We also noticed some gene inconsistencies during the identification of common sex-biased genes within both Stat5b and HNF4\textalpha datasets. For example, some genes that were measured by multiple probes showed sex-biased gene expression for one probe and not the other. These results are somewhat concerning because we identified common sex-biased genes from both microarrays by stringent parameters. These observations justify the need to incorporate larger sample sizes, as well as newer technologies such as RNA-sequencing, as more rigorous methods that negate the need for multiple probe design. Additionally, each microarray dataset contained different mouse strains and this could lead to variability; strains may have different transcription factor networks that influence sex-biased expression patterns. Lastly, RNA-sequencing for Zhx\textsuperscript{D Hep} or Zhx2\textsuperscript{KO} mice have not yet been completed and would provide insight on the plethora of genes that are affected by Zhx2 loss.
Figure 4-1. Slp and C4 mRNA are highly similar. A. Sequence of Slp and C4 genes, including 10 kB upstream and 10 kB downstream, were obtained from UCSC genome browser mm10 and aligned using Geneious. C4a spans 14,363 bp and contains 40 exons, whereas C4b spans 15,518 bp and contains 41 exons. Alignment of Slp and C4 revealed the presence of multiple gaps upstream, within, and downstream of genes and showed roughly 88% pairwise identity across the two genes. Percent pairwise identity significantly dropped roughly 1.8 kB upstream of the transcriptional start sites, and 1 kB downstream of the transcriptional stop sites. The blue box represents the gene regions and the transcriptional start site is denoted by +1. B. Pairwise alignment of Slp and C4
mRNA identified nucleotide differences that were used in the design of gene-specific primers for RT-qPCR analysis. C4 and Slp sequences are labeled with 5’ UTR (light blue) on the left and 3’ UTR (pink) on right; exons are labeled by white boxes; green regions indicate identical nucleotides, yellow denotes nucleotide differences. Slp-specific primers spanning exon 13 and 14 are labeled by dark blue arrows and C4-specific primers spanning exons 21 and 24 are labeled by dark pink arrows. Slp mRNA does not contain C4 exon 21.
Figure 4-2. Slp is not expressed in Zhx2 liver-specific knockout male mice. Liver RNA from 8-week old Zhx2^{fl/fl} or Zhx2^{ΔHep} male (M) and female (F) littermates was analyzed by RT-qPCR for Zhx2 (panel A), Slp (panel B) or C4 (panel C). The mRNA levels were normalized to ribosomal protein L30 (L30) mRNA levels. For Zhx2 RT-qPCR, 10 male and 7 female Zhx2^{fl/fl} mice and 12 male and 10 female Zhx2^{ΔHep} livers were analyzed. For Slp and C4 RT-qPCR, 5 male and 2 female Zhx2^{fl/fl} mice and 5 male and 4 female Zhx2^{ΔHep} livers were analyzed. Ordinary one-way ANOVA was used for statistical comparisons. ***p<.001.
Figure 4-3. Male-biased *Slp* expression is established after birth. Liver RNA was prepared from male (M, blue) and female (F, red) C3B6F1 livers at embryonic day 17.5 (M n= 4, F n= 5), postnatal days p1 (M and F, n= 5), p7 (M n= 5, F n= 8), p14 (M n= 5, F n= 4), p21 (M and F, n= 4), p28 (M n= 5, F n= 4), and p56 (M n= 4, F n= 3) and analyzed by RT-qPCR. *Slp* and *Zhx2* mRNA levels were normalized to ribosomal protein L30 levels. Hash marks indicate longer than 1 week period between timepoints and apply to both graphs (males and females) between p28 and p56. **A.** Hepatic *Slp* mRNA levels are not significantly different between male and female livers until p56 where male *Slp* mRNA levels are dramatically increased. **B.** Hepatic *Zhx2* mRNA levels increase gradually in the postnatal liver in both sexes and are not sex-biased. Two-way ANOVA was used for female to male statistical comparisons at each timepoint. ***p<.001.
**Figure 4-4.** Known sex-biased regulators are unaffected by the absence of Zhx2 in adult liver. Liver RNA was prepared from Zhx2<sup>fl/fl</sup> or Zhx2<sup>ΔHep</sup> male (M) and female (F) littermates and analyzed by RT-qPCR. 

A. Stat5b: Zhx2<sup>fl/fl</sup> (M and F, n=5) and Zhx2<sup>ΔHep</sup> (M n=4, F n=5).
B. Stat5a: Zhx2<sup>fl/fl</sup> (M n=6, F n=5) and Zhx2<sup>ΔHep</sup> (M n=5, F n=4).
C. HNF4α: Zhx2<sup>fl/fl</sup> (M n=10, F n=5) and Zhx2<sup>ΔHep</sup> (M n=7, F n=8).
D. Bcl6: Zhx2<sup>fl/fl</sup> (M n=10, F n=7) and Zhx2<sup>ΔHep</sup> (M n=12, F n=11).
E. Cux2: Zhx2<sup>fl/fl</sup> (M n=10, F n=7) and Zhx2<sup>ΔHep</sup> (M n=12, F n=11). None of these factors are significantly different in livers of Zhx2<sup>ΔHep</sup> and Zhx2<sup>fl/fl</sup> mice. Consistent with previous studies, Bcl6 is significantly higher in male Zhx2<sup>fl/fl</sup> compared to female Zhx2<sup>fl/fl</sup> liver and Cux2 is significantly higher in female Zhx2<sup>fl/fl</sup> compared to male Zhx2<sup>fl/fl</sup> liver. HNF4α, Stat5b, and Stat5a mRNA levels are not sex-biased. One-way ANOVA was used for statistical comparisons. *p<.05, ***p<.001.
Figure 4-5. Stat5b binding sites are present within the *Slp* promoter region. *Slp* upstream promoter regions (-2000 and -1500), including 20 bp downstream of the transcriptional start site were cloned upstream of the firefly luciferase gene in the pGL4.14 plasmid. Experimentally validated (red balls) and JASPAR predicted (black balls) are present in several locations with the promoter region that extends 2000 bp upstream of the *Slp* gene (Varin-Blank et al., 1998).
Figure 4-6. *Zhx2 positively regulates the Slp promoter*. Hepa 1.6 cells were transfected with reporter vectors pGL4.14 with no promoter (pGL4.1.4) or fused to a region of the *Slp* promoter extending 1500 bp (1500) or 2000 bp (2000) upstream of *Slp* exon 1 along with the pcDNA3.1 empty vector control (EV) or pcDNA 3.1 vector expressing *Zhx2* (*Zhx2*) at a 2:1 driver to reporter ratio (350 ng driver, 175 ng reporter); Renilla luciferase was included (10 ng) as an internal control and used for normalization. pGL4.14 was not responsive to EV or *Zhx2*. The 1500 and 2000 *Slp* promoters showed a higher luciferase activity compared to pGL4.14. *Zhx2* significantly activated the *Slp* 2000 promoter, suggesting that upstream sequence between 1500 and 2000 of *Slp* are activated by *Zhx2*. Experiments were completed in three duplicate sets (n=6). One-way ANOVA was used for statistical comparisons. ***p<.001.
Figure 4-7. Constitutively active Stat5b positively regulates NTCP GAS elements *in vitro*. **A.** Structure of the reporter NTCP (Na+/taurocholate cotransporting polypeptide) vector that contains four IFN-γ activated sequence (GAS)-like elements (GLEs) fused to a minimal thymidine kinase promoter in front of the luciferase gene. The NTCP luciferase vector has been shown previously to respond to Stat5b and GH co-transfections in HepG2 cells (Park, Wiwi, & Waxman, 2006). **B.** Hepa 1.6 cells were transfected with the NTCP plasmid along with driver vectors pcDNA3.1, Stat5b (Stat5b, containing a normal Stat5b), or Stat5b* (Stat5b*, containing a constitutively active Stat5b); Renilla luciferase (10 ng) was included as an internal control and used for normalization. Experiments were completed in two duplicate sets (n=4). Stat5b* transfections showed significantly higher luciferase activity indicating positive regulation of the NTCP promoter. Ordinary one-way ANOVA was used for statistical comparisons. ***p<.001.
Figure 4-8. Zhx2 and Stat5b coordinately regulate the Slp promoter. Hepa 1.6 cells were transfected with reporter vectors pGL4.14 (pGL4.14) or pGL4.14-2000 (2000; 100 ng) driver plasmids pcDNA 3.1 (EV), wild-type Stat5b (Stat5b), constitutively active Stat5b (Stat5b*) and Zhx2 (Zhx2), alone or in different combinations as shown. Renilla luciferase (10 ng) was included as an internal control and used for normalization. Driver to reporter ratio for single driver transfections was 4:1 (400 ng driver) and ratio for co-transfection of drivers to reporter was 2:1 (200 ng of each driver). pGL4.14 was unresponsive to transfection with any driver plasmid. The 2000 Slp promoter had statistically higher luciferase activity compared to pGL4.14 in response to all driver vectors. The highest luciferase activity was obtained when Zhx2 and Stat5b or Zhx2 and Stat5b* were co-transfected. Experiments were completed in two duplicate sets (n=4) for pGL4.14 + EV and Slp 2000 (EV, Stat5b, Stat5b*, Zhx2, Stat5b/Zhx2, and Stat5b*/Zhx2); pGL4.14 (Stat5b, Stat5b*, Zhx2, Stat5b/Zhx2, and Stat5b*/Zhx2) were completed in one duplicate set (n=2). One-way ANOVA was used for statistical
comparisons. Significance is reported relative to 2000 S/p promoter EV controls.

Ordinary one-way ANOVA was used for statistical comparisons. **p<.01, ***p<.001, ****p<.0001.
Figure 4-9. Igf-1 mRNA levels are the same in the livers of Zhx2 liver-specific knockout and Zhx2 floxed adult mice. Liver RNA from 8-week old Zhx2^{fl/fl} or Zhx2^{ΔHep} male and female littermates were analyzed by RT-qPCR for Igf-1 gene expression. The mRNA levels were normalized to Ribosomal protein L30 (L30) mRNA levels. One-way ANOVA was used for statistical comparisons.
Figure 4-10. Methods for identifying common genes between Zhx2, Stat5b, and HNF4α microarray datasets. Microarray datasets were downloaded for Zhx2, Stat5b, and HNF4α. Microarrays were first filtered for matching Unigene IDs between datasets. Sex-biased and sex-independent genes were initially identified from both Stat5b and HNF4α datasets as these contained gene data for both males and females, while the Zhx2 microarray was limited to male liver only. To identify sex-biased genes, we used stringent criteria (male wildtype to female wildtype ratio > 2.0, and p-value <.005). This stringency was set to identify genes that are heavily and consistently sex-biased, whereas relaxed parameters may show more variability between datasets. Sex-biased genes that were common between Stat5b and HNF4α datasets were then identified by Unigene IDs. The amount of common sex-biased genes may be limited both by stringency, but also by the number of genes examined within the Stat5b microarray. Furthermore, we identified some genes that were sex-biased by one probe measurement, or oppositely sex-biased between the two datasets (11 genes). These genes were not counted in the final common
sex-biased gene set. Common sex-biased genes were then used to re-filter Stat5b or HNF4α microarrays, and filter the Zhx2 microarray for sex-biased and sex-independent genes. Following identification of sex-biased and sex-independent genes, genes were then filtered for differential expression between male wildtype and male knockout/knockdown within each microarray. We completed two separate filters, one that was stringent (male wildtype to male knockout ratio >2.0, p-value < 0.005) or relaxed (male wildtype to male knockout ratio >2.0, p-value <0.005). Common differentially regulated sex-biased and sex-independent genes were found between Zhx2, Stat5b, and HNF4α microarrays.
**Figure 4-11. Common differentially regulated sex-biased and sex-independent genes between Zhx2, Stat5b, and HNF4α microarrays.** Sex-independent and sex-biased (underlined) common genes that are differentially regulated in all datasets are shown. Sex-biased genes are further broken down into male-biased (blue) or female-biased (red) genes that are affected in each dataset or between datasets. **A.** Stringent differentially regulated genes (male wildtype to male knockout ratio >2.0, p-value < 0.005). **B.** Relaxed differentially regulated genes (male wildtype to male knockout ratio >2.0, p-value <0.05).
Table 4.1. Sex-biased genes significantly affected (FC>2, p<0.05) in BALB/cJ, Stat5b knockout, and HNF4α liver-specific knockout compared to wildtype male mice (Knockout or knockdown / wildtype).

<table>
<thead>
<tr>
<th>Gene Identifier</th>
<th>Gene Name</th>
<th>Gene ID</th>
<th>BALB/cJ Fold Change</th>
<th>Stat5b&lt;sup&gt;−/−&lt;/sup&gt; Fold Change</th>
<th>HNF4α&lt;sup&gt;ΔHep&lt;/sup&gt; Fold Change</th>
<th>Sex-Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_007703</td>
<td>Elongation of very long chain fatty acids 3</td>
<td>Elov13</td>
<td>-5.5</td>
<td>-3.9</td>
<td>-17.8</td>
<td>Male-Biased</td>
</tr>
<tr>
<td>NM_008295</td>
<td>Hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 5</td>
<td>Hsd3b5</td>
<td>6.3</td>
<td>-27.6</td>
<td>-6.6</td>
<td>Male-Biased</td>
</tr>
<tr>
<td>NM_008579</td>
<td>Meiosis expressed gene 1</td>
<td>Meig1</td>
<td>3.3</td>
<td>-2.8</td>
<td>-2.6</td>
<td>Male-Biased</td>
</tr>
<tr>
<td>NM_008648</td>
<td>Major urinary protein 4</td>
<td>Mup4</td>
<td>-2.6</td>
<td>-5.3</td>
<td>-4.9</td>
<td>Male-Biased</td>
</tr>
<tr>
<td>NM_012050</td>
<td>Osteomodulin</td>
<td>Omd</td>
<td>-4.8</td>
<td>-3.7</td>
<td>-4.1</td>
<td>Male-Biased</td>
</tr>
<tr>
<td>NM_013884</td>
<td>Chondroitin sulfate proteoglycan 5</td>
<td>Csgp5</td>
<td>3.9</td>
<td>-2.1</td>
<td>-2.7, -8.3</td>
<td>Male-Biased</td>
</tr>
<tr>
<td>NM_031188</td>
<td>Major urinary protein 1</td>
<td>Mup1</td>
<td>-19.4</td>
<td>-3</td>
<td>-3.3</td>
<td>Male-Biased</td>
</tr>
<tr>
<td>NM_144796</td>
<td>Sushi domain containing 4</td>
<td>Susd4</td>
<td>-17.6</td>
<td>-8.8</td>
<td>-2.9</td>
<td>Male-Biased</td>
</tr>
<tr>
<td>NM_153558</td>
<td>Lipocalin 13</td>
<td>Lcn13</td>
<td>-59</td>
<td>-2.6</td>
<td>-16.1</td>
<td>Male-Biased</td>
</tr>
<tr>
<td>NM_007804</td>
<td>Cut-like 2</td>
<td>Cutl2 (Cux2)</td>
<td>10.1</td>
<td>3.2</td>
<td>7.9</td>
<td>Female-Biased</td>
</tr>
<tr>
<td>NM_007809</td>
<td>Cytochrome P450, family 17, subfamily a, polypeptide 1</td>
<td>Cyp17a1</td>
<td>-8.3</td>
<td>9.1</td>
<td>4.5</td>
<td>Female-Biased</td>
</tr>
<tr>
<td>NM_007813</td>
<td>Cytochrome P450, family 2, subfamily b, polypeptide 13</td>
<td>Cyp2b13</td>
<td>125</td>
<td>100</td>
<td>5.3</td>
<td>Female-Biased</td>
</tr>
<tr>
<td>NM_010000</td>
<td>Cytochrome P450, family 2, subfamily b, polypeptide 9</td>
<td>Cyp2b9</td>
<td>255</td>
<td>50</td>
<td>9.3</td>
<td>Female-Biased</td>
</tr>
<tr>
<td>NM_011994</td>
<td>ATP-binding cassette, sub-family D (ALD), member 2</td>
<td>Abcd2</td>
<td>4.9, 4.2</td>
<td>12.5</td>
<td>5.3</td>
<td>Female-Biased</td>
</tr>
<tr>
<td>NM_134246</td>
<td>Acyl-CoA thioesterase 3</td>
<td>Acot3</td>
<td>3.1</td>
<td>7.7</td>
<td>6.2</td>
<td>Female-Biased</td>
</tr>
<tr>
<td>NM_145368</td>
<td>Acyl-coenzyme A amino N-acetyltransferase 2</td>
<td>Acnat2</td>
<td>-2.76</td>
<td>7.7</td>
<td>4.8</td>
<td>Female-Biased</td>
</tr>
<tr>
<td>NM_145424</td>
<td>RDH16 family member 2</td>
<td>Rdh16f2</td>
<td>-3.2</td>
<td>2.7</td>
<td>2.9</td>
<td>Female-Biased</td>
</tr>
<tr>
<td>XM_129214</td>
<td>Proprotein convertase subtilisin/kexin type 5</td>
<td>Pcsk5</td>
<td>2</td>
<td>2.5</td>
<td>2.1</td>
<td>Female-Biased</td>
</tr>
</tbody>
</table>

Microarrays were completed using 8-week old Sub13 (Zhx2-wildtype) and BALB/cJ (Zhx2-low) male mice (n=4 per genotype), 8-9 week old 129 x BALB/c Stat5b floxed and Stat5b<sup>−/−</sup> mice (n=6 per genotype), or 7-week old HNF4α floxed and HNF4α<sup>ΔHep</sup> male mice (n=4 per genotype). Common sex-biased genes that are differentially regulated within all microarrays were found using methods described in Figure 4-10.
CHAPTER V

DIFFERENTIAL EXPRESSION OF HIGHLY HOMOLOGOUS MOUSE CYP2A4 AND CYP2A5 GENES

The mammalian Cytochrome P450 (Cyp) gene superfamily encodes enzymes involved in a wide range of processes in adult liver, including metabolism of pharmaceuticals, foreign chemical and pollutants; cholesterol, sterol and bile acid biosynthesis; and steroid synthesis (Nelson et al., 2004). Humans and mice contain 58 and 105 Cyp genes, respectively (Hrycay and Bandiera, 2009). The expansion of Cyp gene number and function is hypothesized to coincide with the interaction of animals and plants; plants developed metabolites to make themselves less palatable or more toxic to animals and animals in response evolved new Cyp genes to metabolize plant toxins (Nebert and Dieter, 2000). Consequently, Cyps have the ability to metabolize modern xenobiotics including pollutants, carcinogens, and drugs. Many Cyps are expressed in a variety of tissues. However, the majority of mouse Cyps are expressed highly in the liver, the major detoxification organ (Renaud et al., 2011). The products of xenobiotic metabolism are often more toxic and/or carcinogenic than the original substrate; these metabolites can lead to human disease, including cancer (Nebert and Dieter, 2000; Nebert and Russell, 2002). Thus, understanding differential Cyp gene regulation can provide valuable information towards understanding molecular pathways that predispose individuals to disease, including cancer.
In mammals, the liver must perform a number of new metabolic functions at birth that coincide with new sources of nutrition. This functional shift is governed by gene expression changes during the perinatal period. Levels of many Cyp mRNAs increase dramatically during postnatal liver development and different Cyp genes exhibit distinct developmental expression patterns (Cui et al., 2012). These distinct changes in Cyp gene expression are governed by a variety of factors, including different transcription factors and hormones.

Previous studies demonstrated that multiple mouse and human Cyp genes exhibit sex-biased expression (Waxman and Holloway, 2009; Zhang et al., 2011) and that some mouse Cyp genes are zonally regulated (Braeuning et al., 2006; Loeppen et al., 2005; Sekine et al., 2006). In mice, sex-biased gene expression differences are established during the first few months after birth (Conforto and Waxman, 2012) whereas zonal expression initiates shortly before birth (Burke et al., 2018).

The mouse Cyp2 family contains 53 individual genes that can be further divided into subfamilies, including the Cyp2a subfamily comprised of 4 members: Cyp2a4, Cyp2a5, Cyp2a12, and Cyp2a22. The Cyp2a4 and Cyp2a5 genes are expressed in the liver, kidney, and olfactory mucosa and exhibit female-biased expression in liver (Lindberg et al., 1989; Piras et al., 2003; Squires and Negishi, 1988; Su et al., 1996). Hepatic Cyp2a4 mRNA levels are nearly 1000-fold higher in female compared to male mice whereas Cyp2a5 mRNA levels in the liver are only about 3-fold higher in females compared to males (Creasy et al., 2016).

Cyp2a4 and Cyp2a5 are highly similar at the DNA sequence level; the proteins encoded by these two genes differ by only 11 amino acids (Lindberg and Negishi, 1989).
Despite their high homology, the substrate targets of Cyp2a4 and Cyp2a5 differ significantly. Cyp2a4 catalyzes testosterone hydroxylation (testosterone 15α-hydroxylase), whereas Cyp2a5 catalyzes coumarin hydroxylation (coumarin 7-hydroxylase). Lindberg and colleagues demonstrated through site-directed mutation of Cyp2a5, that changing Phenylalanine at position 209 to Leucine, the amino acid at this position in Cyp2a4, significantly reduces coumarin hydroxylation but increases testosterone hydroxylation.

Cyp2a5 is the mouse ortholog of human CYP2A6, based on shared catalytic hydroxylation of coumarin and a variety of procarcinogens (Hrycay and Bandiera, 2009). Although coumarin 7-hydroxylase is present in mouse, rats, and humans, Cyp2a4 is only found in mice. Further examination revealed that Cyp2a5 and Cyp2a4 are present in many domestic mouse strains whereas a wild mouse strain, Mus spretus, only contains Cyp2a5 (Aida et al., 1994). This indicates that Cyp2a5 gave rise to Cyp2a4 via gene duplication within the last 2 million years (Lindberg et al., 1992).

Given the high sequence homology but dramatic differences in sex-biased expression of mouse Cyp2a4 and Cyp2a5 genes, we sought to understand how these two genes are differentially regulated. We found that the extent of sex-biased expression is due to postnatal differences in male mice. We also demonstrate that Cyp2a4 and Cyp2a5 have different zonal expression patterns; Cyp2a4 and Cyp2a5 mRNA levels are higher in periportal (PP) and pericentral (PC) hepatocytes, respectively. Bioinformatic analysis revealed potential differences in transcription factor binding sites that could contribute to distinct aspects of Cyp2a4 and Cyp2a5 expression. The presence of an antisense expressed sequence tag (EST) overlapping Cyp2a4 led us to hypothesize that differential
sex-biased and/or zonal regulation of \textit{Cyp2a4} and \textit{Cyp2a5} may be due to antisense regulation of Cyp2a4. We have fully characterized this Cyp2a4 antisense transcript and also discovered a similar Cyp2a5 antisense transcript. We found that both Cyp2a4 and Cyp2a5 antisense transcripts are developmentally regulated in the liver but are present at significantly lower levels than their sense counterparts. Cyp2a4 antisense (Cyp2a4as) transcripts are present as multiple spliced isoforms whereas Cyp2a5as has only a single major spliced isoform. Taken together, these studies identify potential transcriptional and post-transcriptional mechanisms governing different Cyp2a4 and Cyp2a5 mRNA expression patterns in mouse liver.

\textbf{RESULTS}

\textit{Cyp2a4 and Cyp2a5 loci are highly similar.} Previous studies suggest that \textit{Cyp2a4} arose from a duplication of \textit{Cyp2a5} in domesticated mice roughly 2 million years ago (Lindberg et al., 1992). Past studies have also recognized that sequence differences exist between \textit{Cyp2a4} and \textit{Cyp2a5} in 129/J mice (Squires and Negishi, 1988). To compare these two genes in C56BL/6 mice, matrix analysis was performed using sequence information of C57BL/6 genomic data. This indicates that a region of strong homology extends roughly 15 kbp upstream and 48 kbp downstream of these two genes (Fig. 5-1A). The similarity within this 70 kbp region is 96.3%, but completely disappears outside this region. A more focused analysis of C57BL/6 \textit{Cyp2a5} (7,926 bp) and \textit{Cyp2a4} (7,897 bp) genes indicated that the exons were 98.2% identical; transcript lengths for Cyp2a4 and Cyp2a5 are 1715 and 1716 nt, respectively. Within the exons there are 31 nucleotide differences, 23 of which are in coding regions, resulting in 12 amino acid differences.
between the two proteins (Fig. 5-1B). In addition, the greatest variation between Cyp2a4 and Cyp2a5 genes was found in the 3’ UTR; with 6 nt differences existing across 205 bp sequences. Thus, sequence similarity over the 70 kbp of DNA surrounding the Cyp2a4 and Cyp2a5 genes was nearly as high as the exonic regions of these genes, consistent with a recent duplication event.

**Cyp2a4 and Cyp2a5 mRNA levels are higher in female liver.** Previous studies demonstrated that Cyp2a4 and Cyp2a5 had female-biased liver expression patterns in 5-11 week old 129/J mice (Harada and Negishi, 1984; Negishi et al., 1989; Squires and Negishi, 1988). To measure Cyp2a4 and Cyp2a5 liver sex-biased mRNA levels in C57BL/6 and C3H/HeJ mice, Cyp2a4 and Cyp2a5 specific primers were used, and mRNA levels were quantitated by RT-qPCR (Figures 5-2A and B). We designed primers that spanned sequence differences in Exon 4; primers were designed against the exon 3 and 4 splice junction and within exon 4. Significant female-biased mRNA levels for Cyp2a4/a5 were seen in both mouse strains. Importantly, female-biased Cyp2a4 gene expression is much more pronounced, whereas Cyp2a5 female-biased expression is more modest in 8-week old mice. These results demonstrate that Cyp2a4 and Cyp2a5 exhibit female-biased gene expression independent of mouse strain.

**Cyp2a4 and Cyp2a5 expression increases after birth.** Many Cyps and other genes active in drug metabolism show sex differences beginning at puberty (Conforto and Waxman, 2012; Hart et al., 2009). Previous studies demonstrated gradually increased
Cyp2a4 mRNA levels during male and female mouse liver development until postnatal day 28 (p28). However, by p56, male gene expression had decreased, while female expression increased dramatically (Figure 5-3A) (Creasy et al., 2016). To compare developmental expression patterns of Cyp2a4 and Cyp2a5, we measured Cyp2a4 and Cyp2a5 female and male liver mRNA levels, beginning at embryonic day 17.5. We find that Cyp2a5 mRNA levels increase similarly in male and female mouse liver until p28; expression is then markedly increased from p28 to p56 specifically in females whereas expression remains constant in males (Figure 5-3B). Thus, in 8-week old C3H female mice, Cyp2a4 and Cyp2a5 mRNA levels are at most 1.5-fold different, whereas in 8-week old C3H male mice, Cyp2a5 is predominantly expressed, albeit at levels about 3-fold lower than in females.

**Cyp2a4 and Cyp2a5 display different zonal expression patterns.** To directly determine the zonal expression of Cyp2a4 and Cyp2a5, we isolated pericentral (PC) and periportal (PP) hepatocytes from E3-βgl-Dd transgenic mice. These mice contain the alpha-fetoprotein enhancer element 3 (E3) fused to the human β-globin promoter to drive transcription of the mouse major histocompatibility complex H-2Dd class I structural gene (Dd mice). E3 has been shown to restrict Dd expression to PC cells (Clinkenbeard et al., 2012; Ramesh et al., 1995). Utilizing this PC-restricted Dd cell surface expression, we separated Dd positive (PC hepatocytes) from Dd negative (PP hepatocytes) populations by fluorescence-activated cell sorting (FACS) (Figure 5-4A). RNA was prepared from male and female separated populations and analyzed using RT-qPCR. A known PP gene, serine dehydratase (Sds), and a known PC gene, ornithine aminotransferase (OAT), were
used to confirm separation of the hepatocyte populations (Figure 5-4B). As expected, Sds showed enrichment in PP hepatocytes and OAT showed enrichment in PC hepatocytes. We next measured \( \text{Cyp2a4} \) and \( \text{Cyp2a5} \) expression in these cell populations. Consistent with \( \text{Cyp2a4} \) expression being very low in males, \( \text{Cyp2a4} \) expression was low in both sorted male populations compared to females (Figure 5-4C). However, in female mice, \( \text{Cyp2a4} \) shows an enrichment in the PP hepatocyte populations. In contrast, \( \text{Cyp2a5} \) expression is enriched in PC populations in both male and female mice (Figure 5-4D). However, these differences are not significant (males; \( p \)-value = 0.09; females; \( p \)-value = 0.34).

To examine zonal gene expression using another protocol, we chose carbon tetrachloride (CCl\(_4\)) injected mice. CCl\(_4\)-induced liver damage is a well-characterized experimental model for studying liver regeneration. A single injection of CCl\(_4\) into adult male mice causes acute pericentral necrosis that can be visualized three days after administration. Proliferation of hepatocytes bordering pericentral necrosis restores normal liver cellularity within 7-10 days. C3H/HeJ mice were injected with a single i.p. dose of mineral oil (MO) or CCl\(_4\) and livers were collected three days post injection, a time when most PC hepatocytes have been killed so the majority of cells should have a PP gene expression pattern. \( \beta \)-cat mRNA levels, which are not zonally expressed, do not change with this treatment (Figure 5-5). Sds mRNA levels were significantly higher in CCl\(_4\) treated mice, consistent with its known PP expression pattern. OAT mRNA levels were lower in CCl\(_4\) treated mice, consistent with its known PC expression pattern. Cyp2a4 mRNA levels were higher whereas Cyp2a5 mRNA levels were lower after CCl\(_4\)
treatment. These data demonstrate that *Cyp2a4* and *Cyp2a5* have contrasting zonal expression patterns; *Cyp2a4* shows PP expression and *Cyp2a5* shows PC expression.

**Cyp2a4 and Cyp2a5 expression differences may be regulated by transcriptional or post-transcriptional mechanisms.** Combined differences in sex-biased and zonal Cyp2a4 and Cyp2a5 mRNA levels patterns led us to consider how the *Cyp2a4* and *Cyp2a5* genes may be differentially regulated in mouse liver. We reasoned that analysis of the limited sequence differences between these genes may provide clues to any transcriptional or post-transcriptional mechanisms that contribute to their differential regulation.

We initially examined potential transcription factor binding differences between the *Cyp2a4* and *Cyp2a5* regions 2000 bp upstream from their transcriptional start sites. Aligning these sequences revealed only 22 nt differences between them. These *Cyp2a4* and *Cyp2a5* sequences were then scanned for predicted transcription factor binding (PTFB) sites using *Mus musculus* JASPAR CORE software for 176 *Mus musculus* transcription factors (Khan et al., 2018; Sandelin et al., 2004). Those PTFB sites that were altered by the nucleotide differences between *Cyp2a4* and *Cyp2a5* were then identified. To obtain greater stringency, only PTFB sites with a relative score of 0.80 and overall score above 7.5 were considered as more likely to bind *Cyp2a4* or *Cyp2a5* upstream regions. In total, of the 22 nucleotide differences between the genes, 12 resulted in changes to PTFB sites (Figure 5-6). Unique PTFB for *Cyp2a4* included Jun:Fos, Rxra, and Zfx, whereas unique PTFB for *Cyp2a5* included NR3C1, NFYA, and FoxA2. In addition, some PTFB sites were shared between the *Cyp2a4* and *Cyp2a5* upstream
regions but had different PTFB scores (Egr2, Sp1, Hic1, HNF4a). This provides a short list of potential transcription factors that may contribute to different *Cyp2a4* and *Cyp2a5* expression patterns.

While we were investigating potential transcription factor binding site differences between *Cyp2a4* and *Cyp2a5*, we came across an expressed sequence tag (EST), RP23-267I20.3 (RP23). RP23 is annotated in the UCSC Genome Database as an antisense RNA that overlaps with the 5' end of the *Cyp2a4* gene. No such annotation is present for the *Cyp2a5* gene (Figure 5-7). To determine whether this transcript existed in the mouse liver and whether it may contribute to the differential expression patterns of *Cyp2a4* and *Cyp2a5*, PCR primers were designed against RP23 and RT-PCR was performed using RNA from the sorted Dd hepatocyte populations (Figure 5-8A). To identify specific amplicons, products were resolved by electrophoresis using a 1% agarose gel. The migration of amplicons from the male and female hepatocyte populations appeared slightly different so, to discriminate these products further, they were resolved using a 6% acrylamide gel (Figure 5-8B). The acrylamide gel showed that three band sizes were present, an upper band found in both sexes, one lower band in males, and a third band in females. A closer look at our primer sequences revealed that, if an antisense transcript was also synthesized from the *Cyp2a5* gene, the RP23 primers could not distinguish between that RNA and one from the *Cyp2a4* upstream region (Figure 5-8C). Therefore, specific primers were designed to distinguish between Cyp2a4 and Cyp2a5 antisense transcripts (Cyp2a4as and Cyp2a5as), taking advantage of a four nucleotide gap sequence difference for one primer (Figure 5-9A; P4, P6) and a single nucleotide difference between the sequences for the other primer (Figure 5-9A; P3, P5). Each of these
differences were placed at the 3’ ends of the PCR primers. RT-PCR from male and female whole liver RNA, and analysis of melting curves associated with each primer pair for Cyp2a4as (P1, P2; P3, P4) and Cyp2a5as (P5, P6) revealed that a minor shift was present between males and females for Cyp2a4as (Figure 5-9B). These slight shifts prompted us to analyze amplicons by gel electrophoresis on both agarose and acrylamide gels. Gel electrophoresis revealed that Cyp2a5as primers amplified a single amplicon from both sexes (Figure 5-9C). In contrast, Cyp2a4as-specific primers (P3, P4) amplified two products in the female samples and one in the male samples. The male and female Cyp2a4as (P3, P4) and Cyp2a5as amplicons were cloned and sequenced, along with male Cyp2a4as (P1, P2) and lower band of female (P1, P2). Sequencing of Cyp2a4as (P1, P2) bands from both males and females resulted in mixed Cyp2a4as and Cyp2a5as species; some amplicons contained both the 4 n.t. gap in Cyp2a5as and single nucleotide polymorphisms present in Cyp2a4as. In contrast, the lower band from female Cyp2a4as (P3, P4) matched only Cyp2a4as sequence, whereas both male and female bands from the Cyp2a5as reactions matched only Cyp2a5as. Surprisingly, the upper band from Cyp2a4as (P3, P4) male and female mice was identified to be from the unrelated Apolipoprotein B (ApoB) transcript. Our initial analysis of Cyp2a4as (P1, P2 and P3, P4) primers by BLASTN, linked from PrimerQuest, did not identify potential cross reactivity of our primers with the ApoB transcripts. However, we could manually align the Cyp2a4as-2 (P3, P4) primers to the amplified portion of the ApoB mRNA exon 26 (Figure 5-9D). In contrast, the Cyp2a5as primers showed less similarity to the ApoB transcript, especially at the 3’ ends of the primers. Temperature gradient PCR reactions were performed using annealing temperatures between 60°C and 70°C with Cyp2a4as (P3 & P4) primers
At 70°C annealing, the ApoB amplicon is absent, only the Cyp2a4 amplicon is present and it is more abundant in females compared to males. All future RT-qPCR reactions for Cyp2a4as were therefore carried out with a 70°C annealing temperature to ensure specific Cyp2a4as amplification and quantitation, whereas all Cyp2a5as RT-qPCR reactions were carried out with a 60°C annealing temperature, which this data show, result in a single specific amplicon. ApoB abundance was also analyzed and is much more abundant than Cyp2a4as in C3H mice livers and this higher expression is thought to have played a role in off-target amplification (Table 5.1).

**Cyp2a4as and Cyp2a5as transcripts extend into Cyp2a4 and Cyp2a5 gene regions.**

To characterize the 5’ ends of Cyp2a4as and Cyp2a5as RNAs, 5’ RACE was performed using C57BL/6 female liver RNA. For initial reactions, a single RT primer (P7) was used to synthesize cDNA for Cyp2a4as and Cyp2a5as simultaneously (Figure 5-10A). Two gene-specific primers were then used for the first PCR reactions (P6, P8) followed by a second set of gene-specific primers for the nested PCR reaction (P9, P10). Amplified products were cloned and sequenced, which identified additional upstream sequence for both Cyp2a4as and Cyp2a5as (Figures 5-10B and C). All three bands cloned and sequenced for Cyp2a4as identified upstream sequences that overlapped with each other, whereas only a single band from Cyp2a5as identified upstream sequence. The upstream Cyp2a5as region identified aligned with the first exon of RP23. Additional 5’ RACE reactions were performed using gene-specific RT and PCR primers for Cyp2a4as (P9, P11, P12) and Cyp2a5as (P6, P10, P13, P14) and further upstream sequence for Cyp2a4as, but not Cyp2a5as were identified (Figures 5-10C and D).
To characterize the 3’ end of the Cyp2a4as and Cyp2a5as RNAs, 3’ RACE was performed using oligo dT-anchor primer for the RT reaction and three gene specific PCR primers for Cyp2a4as (P16, P18, P19) and Cyp2a5as (P17, P5, P20) with anchor primer for the first PCR reactions (Figure 5-11A and B). Aliquots of Cyp2a4as (P18) and Cyp2a5as (P5) reactions were then used in nested PCR reactions with P19 and P20 and anchor primer. Nested PCR reactions yielded a single amplicon of the same size for Cyp2a4as and Cyp2a5as (Figure 5-11C). These 3’ RACE amplicons were cloned and sequenced, and were found to match the 3’ end of RP23. Furthermore, a cleavage-polyadenylation signal, AAUAAA, was conserved between Cyp2a4as and Cyp2a5as at an appropriate distance from the 3’ end that had been primed with oligo dT, suggesting that both transcripts use the same signal for 3’ end processing.

**Cyp2a4as transcripts are present as multiple spliced isoforms.** To fully characterize Cyp2a4as, primers were designed to amplify full length Cyp2a4as transcripts from the ends identified by the 5’ and 3’ RACE experiments. Curiously, our initial RT-PCR followed by cloning and sequencing from C57BL/6 female whole liver RNA resulted in mixed populations of Cyp2a5as and Cyp2a4as. This result was surprising because the Cyp2a5 5’ RACE experiments had not identified additional upstream Cyp2a5as sequence. To address this issue, we used gene-specific 3’ primers that incorporate the 4 nt gap in the last exon of the transcripts (P8, P6) and one 5’ primer that could bind both transcripts (P15) (Figure 5-12A). RT-PCR was carried out in male and female nuclear fractions to enrich for 2a4/2a5 antisense transcripts. Additionally, males were used as a negative control for Cyp2a4as-specific amplification because its expression was so low.
Cloning and sequencing identified five female Cyp2a4as spliced isoforms and only one Cyp2a5as spliced isoform in males and females (Figure 5-13). Within Cyp2a4as isoforms, the region that overlaps Cyp2a4 intron 1 and exon 2 varied, with the other antisense exons being constant.

Cyp2a4as transcript lengths ranged between 1098 nt to 1333 nt; the Cyp2a5as transcript length was 1235 nt. Using the Coding Potential Assessment Tool (CPAT), predicted open reading frames (ORF) were identified and the probability that they coded for proteins was analyzed. H19 is a known long non-coding RNA (lncRNA) that was used as a negative control for CPAT analysis (Bartolomei et al., 1991). Because the ORFs identified for H19, Cyp2a4as, and Cyp2a5as transcripts were relatively short, with non-optimal hexamer use, they are all judged to be non-coding (Table 5.2) (Wang et al., 2013b). As expected, ORFs predicted to be coding were identified for the Cyp2a4 and Cyp2a5 mRNA sequences. These data suggest that Cyp2a4as and Cyp2a5as are long non-coding transcripts, and that Cyp2a4as is comprised of multiple spliced isoforms in females.

**Cyp2a4as and Cyp2a5as expression patterns.** Once we had characterized the extent of the Cyp2a4as and Cyp2a5as transcripts, we then turned to characterizing their expression patterns for comparison to the mRNAs from their overlapping genes. To ensure that we detected all the RNA isoforms and specifically distinguished Cyp2a4as from Cyp2a5as, we used primers P3 and P4 with a 70°C annealing temperature for Cyp2a4as and primers P5 and P6 with a 60°C annealing temperature for Cyp2a5as in the RT-qPCR reactions. We quantitated the Cyp2a4as and Cyp2a5as transcripts in the same cDNA samples used
for the Cyp2a4 and Cyp2a5 mRNA analyses to assess sex-biased, developmental and zonal gene expression patterns in adult mice. Cyp2a4as and Cyp2a5as RNA levels are considerably lower than their sense mRNA counterparts in adults of both sexes, ranging from about 100- to 800-fold lower in C57BL/6 mice (Figure 5-14A, Table 5.3). In addition, both antisense transcripts are female-biased with the ratio of female to male expression of the antisense transcripts being similar to those of their corresponding mRNAs. That is, Cyp2a4 and Cyp2a4as are both highly female biased whereas Cyp2a5 and Cyp2a5as are only modestly biased (Table 5.4). In males, Cyp2a5as is predominantly expressed compared to Cyp2a4as, whereas in females, both antisense are expressed roughly equally (Table 5.3). Again, this is similar to the expression ratios of Cyp2a5 and Cyp2a4 in males and females. During development, Cyp2a4as expression is significantly different in males and females by p21 (Figure 5-14B). Thus, sex-biased expression for Cyp2a4as occurs earlier in development compared to Cyp2a4 mRNA, which was not seen until p56 (see Figure 5-3A). In contrast, Cyp2a5as sex-biased expression is delayed until p28 (Figure 5-14B), but this is still earlier than the sex-biased expression of Cyp2a5, which is not evident until p56 (see Figure 5-3B). In addition to sex-biased expression, both antisense transcripts exhibit zonal regulation that mirrors their sense counterparts (Figure 5-14C compared to Figures 5-4C and D). Cyp2a4as is PP enriched, whereas Cyp2a5as is PC enriched. Taken together, our data show that Cyp2a4as and Cyp2a5as expression is considerably lower than the mRNAs that they overlap but their patterns of expression positively correlate with Cyp2a4 and Cyp2a5 mRNA expression under all conditions we have analyzed.
Potential transcription factor binding differences between \textit{Cyp2a4as} and \textit{Cyp2a5as} promoters. Given that \textit{Cyp2a4as} and \textit{Cyp2a5as} exhibit sex-biased expression patterns that precede \textit{Cyp2a4} and \textit{Cyp2a5} during liver development, we questioned whether differences in the promoter regions of the two antisense transcripts existed that may affect their transcription. The sequence upstream of the 5’RACE-identified 5’ ends of \textit{Cyp2a4as} and \textit{Cyp2a5as} were analyzed using JASPAR and the same criteria were applied for prediction of transcription factor binding as described for \textit{Cyp2a4} and \textit{Cyp2a5} promoter analysis. Compared to the \textit{Cyp2a4} and \textit{Cyp2a5} promoter regions, many more single nucleotide polymorphisms (SNPs) and gaps exist between the two regions of these promoters (Figure 5-15). In total, these two regions exhibit 95.6% similarity and contain 90 differences over 2000 bp, including all nucleotide gaps and SNPs. Interestingly, among one of the highest scoring predicted transcription factor binding sites was Stat5a:Stat5b for \textit{Cyp2a4as}. Stat5 is a known sex-biased transcriptional regulator that acts to repress transcription of female-biased genes in male liver.

\textbf{Cyp2a4as and Cyp2a5as are nuclear-enriched transcripts.} Functional studies of lncRNAs have shown that many act as repressors or activators of gene expression through various mechanisms including modifying the epigenetic state of surrounding target genes, regulating alternative splicing of target mRNAs, and deterring RNA polymerase binding efficiency to target genes (Clark et al., 2012). In the cytoplasm, lncRNAs can regulate transcript stability by preventing transcription factor binding to target genes, reducing translational efficiency of sense RNAs, and sequestering of miRNAs preventing mRNA target degradation (Kung et al., 2013). To identify a potential
functional role for Cyp2a4as and Cyp2a5as, we determined the subcellular localization of these transcripts. RNA from liver nuclei and cytoplasmic fractions was isolated and analyzed by RT-qPCR for known cytoplasmically-enriched spliced mRNAs (ribosomal protein L30, Cyp2a4, Cyp2a5) and known nuclear-enriched RNAs, including unspliced pre-mRNA (L30-U) and U6 small nuclear RNA (U6). We first determined whether cytoplasmic to nuclear ratios were different between two males and two females for all genes and found that no differences existed. As a result, cytoplasmic to nuclear ratios were combined for both sexes and each ratio was normalized to the cytoplasmic to nuclear RNA ratio for L30 mRNA. Cytoplasmically enriched L30, Cyp2a4, and Cyp2a5 mRNAs were significantly different from nuclear enriched unspliced L30 pre-mRNA (L30-U) and U6 (Figure 5-16). By comparing the cytoplasmic to nuclear abundance ratio of the Cyp2a4as and Cyp2a5as transcripts to the control cytoplasmic and nuclear RNAs, both are clearly localized to the nucleus. Thus, any functional role they may play in contributing to Cyp2a4 and Cyp2a5 regulation would require this to occur in the nucleus.

**Cyp2a4 and Cyp2a4as are upregulated in a mouse model of HCC.** We have previously shown that *Cyp2a4* expression, but not *Cyp2a5* expression, is increased in tumors in a mouse model of HCC (Creasy et al., 2016). To validate these results and test whether Cyp2a4as and Cyp2a5as RNA levels are altered in liver tumors, we analyzed their expression in this mouse tumor model. C57BL/6 male mice were injected with PBS or DEN and, after 34 weeks, liver tissue was collected; from DEN-injected mice, non-tumor and tumor tissue was collected. RNA was isolated from these liver tissue samples. Similar to the previous results, Cyp2a4 mRNA was significantly higher in tumor tissue.
compared to non-tumor and PBS controls whereas Cyp2a5 mRNA was not affected in these tumors. Cyp2a4as, like Cyp2a4, showed significantly higher expression in DEN-induced liver tumors compared to PBS and non-tumor liver tissue (Figure 5-17). Cyp2a5as also mirrored Cyp2a5, and was not differentially expressed in tumors. Overall, Cyp2a4as and Cyp2a5as mirror expression patterns of Cyp2a4 and Cyp2a5 in DEN-induced liver tumors.

**Additional Cyp2 family antisense RNAs identified in male and female C57BL/6 liver.** Because we found that both Cyp2a4 and Cyp2a5 genes had overlapping antisense transcripts that were expressed in a similar pattern to their mRNA, we wondered whether other members of the Cyp2 gene family or other Cyp genes also contained antisense transcripts. Therefore, we analyzed the UCSC Genome Database Gencode tracks for overlapping, antisense annotated EST’s. We identified five potential Cyp antisense transcripts: RP23.105M23.5 overlapping Cyp8a1 (PTGIS), RP23-346B17.4 overlapping Cyp5a1, AC157553.6 overlapping Cyp2f2, RP23-86I19.1 overlapping Cyp2u1, and RP24-191c15.3 overlapping Cyp2s1. Additionally, we identified an annotated EST long non-coding RNA (lncRNA) between Cyp2b10 and Cyp2s1 named RP24-191C15.5. We chose Cyp8a1 antisense (Cyp8a1as), Cyp2u1 antisense (Cyp2u1as), Cyp2s1 antisense (Cyp2s1as), and lncRNA RP24-191C15.5 (RP24) transcripts to determine whether they were expressed in mouse liver because we could design intron-spanning primers specific to these transcripts. Amplicons from Cyp8a1as RT-qPCR reactions did not match predicted amplicon size despite attempted optimization of RT-qPCR conditions. Additionally, RT-qPCR for Cyp2u1as failed to produce any detectable transcript. Only
specific amplicons with predicted sizes for lncRNA RP24 and Cyp2s1as were detected in adult male and female livers (Figure 5-18). Unlike *Cyp2a4as* and *Cyp2a5as*, these genes do not exhibit sex-biased expression and they are expressed at much lower levels than *Cyp2a4as* and *Cyp2a5as*. As expected, H19 mRNA, a well-known long non-coding RNA, also showed no sex-biased expression. These data demonstrate that *Cyp2a4as* and *Cyp2a5as* are expressed at comparable levels to H19 in normal adult livers and show unique sex-biased patterns compared to lncRNA RP24 and Cyp2s1as described here.

Previous studies have analyzed Cyp2b10 and Cyp2s1 mRNA levels in mouse liver (Renaud et al., 2011). Cyp2s1 is primarily expressed in extra-hepatic tissues whereas Cyp2b10 is a female-biased mRNA. Further characterization of lncRNA RP24 and Cyp2s1 is needed to determine their full transcript lengths and any potential regulatory roles in *Cyp2s1* and Cyp2b10 gene expression patterns.

**DISCUSSION**

*Cyp2a4* and *Cyp2a5* are highly homologous genes that are differentially regulated in male and female mice. In the present study, we used pairwise alignment and discovered that *Cyp2a4* and *Cyp2a5* genes are highly homologous up to ~15 kB upstream and ~48 kB downstream of both genes. These results are consistent with the proposal that *Cyp2a5* gave rise to *Cyp2a4* by gene duplication (Wang et al., 2003). Interestingly, our analysis revealed that the highest number of differences between Cyp2a4 and Cyp2a5 transcripts are within their 3’ UTR. Recently, Hao and colleagues have identified sex-biased microRNAs (miRNAs) that are regulated by growth hormone patterns in males and females (Hao and Waxman, 2018). Specifically, miR-1948-3p is a male-biased
miRNA, whereas miR-802 is a female-biased miRNA. Furthermore, both miRNAs show sex-differential expression in livers of 4 week old mice and are able to bind the Argonaute complex. Interestingly, inhibition of miR-1948-3p in male mice lead to increased Cyp2a4 mRNA levels (~3 fold compared to negative controls), whereas inhibition of miR-802 in female mice led to decreased Cyp2a4 and Cyp2a5 mRNA levels (~2 fold). Our own analysis using miRNA prediction programs TargetScan, miRANDA, and DIANA tools-Tarbase predicted no binding of these miRNAs to either Cyp2a4 and Cyp2a5 3’ UTRs. This may suggest that regulation of Cyp2a4 and Cyp2a5 by these miRNAs may be indirect.

RT-qPCR for Cyp2a4 and Cyp2a5 mRNA levels revealed that both RNAs are present at relatively low levels until p28. Interestingly, in females, Cyp2a4 and Cyp2a5 mRNA levels substantially increase from p28 to p56. In males, Cyp2a4 mRNA levels decrease from p28 to p56 and Cyp2a5 mRNA levels are only slightly increased from p28 to p56, resulting in female-biased expression of both Cyp2a4 and Cyp2a5 in adult mouse liver. These results suggest that female-biased mechanisms important for positive regulation of Cyp2a4 and Cyp2a5 gene expression are activated between p28 and p56 in female mice. Furthermore, these female-biased positive regulators of Cyp2a4 and Cyp2a5 gene expression are either not activated in males, or are repressed by other mechanisms in males.

Previous studies showed that growth hormone (GH) secretion patterns differences between males and female mice manifest between 4 and 8 weeks of age, at the onset of puberty (Conforto and Waxman, 2012). In males, GH release from the anterior pituitary is pulsatile, characterized by periods of little to no GH in circulation (Waxman and
O'Connor, 2006). In females, GH release from the anterior pituitary is nearly continuous and GH is always present in circulation. GH is known to be the major hormonal regulator of sex-specific genes in liver. GH binding to its cell surface receptor activates Jak2, a GH-receptor tyrosine kinase, which in turn phosphorylates and activates Stat5b that translocates into the nucleus to regulate Stat5b target genes. In male mice, Stat5b is activated in response to each GH pulse and activates male-biased transcriptional regulators such as Bcl6. In females, continuous GH stimulation leads to reduced Stat5b nuclear translocation and activation of female-biased transcriptional regulators such as Cux2 (Clodfelter et al., 2006; Waxman et al., 1995). Furthermore, it has been previously shown that loss of Stat5b or HNF4α results in significant upregulation of Cyp2a4 in male liver (Holloway et al., 2006). In addition, loss of HNF4α results in significant upregulation of Cyp2a5 in male liver whereas Cyp2a5 gene expression has not been not tested in Stat5b−/− mice. (Clodfelter et al., 2006; Holloway et al., 2008). Interestingly, loss of HNF4α in female liver does not significantly affect Cyp2a4 gene expression (Holloway et al., 2006). These results suggest that sex-biased gene expression of Cyp2a4 and Cyp2a5 relies on integration of many factors in both males and females. Our analysis using JASPAR software for prediction of transcription factor binding sites to the Cyp2a4 or Cyp2a5 promoter identified differences in HNF4α binding; the Cyp2a4 promoter contained a slightly higher binding site prediction score (9.69) compared to Cyp2a5 (9.19). Perhaps the slightly increased binding score of HNF4α to the Cyp2a4 promoter region results in higher repression of Cyp2a4 compared to Cyp2a5 in male liver. Interestingly, in addition to these findings, FoxA2 was predicted to bind only within the Cyp2a5 promoter region. Recent data shows that FoxA2, along with FoxA1, are
associated with the enhancer chromatin mark, H3K4me1, and are enriched at male-biased DNase hypersensitivity (DHS) sites in male liver (Sugathan and Waxman, 2013). Furthermore, it has been shown that FoxA1, FoxA2, and HNF4α co-occupy H3K4me1 loci in mouse liver (Hoffman et al., 2010). Given these findings, perhaps FoxA2 and HNF4α act to positively regulate expression of Cyp2a5 in male liver, whereas binding of only HNF4α acts to repress Cyp2a4. However, ChIP-Seq analysis from C57BL/6 liver has shown that FoxA2 binds to both Cyp2a4 and Cyp2a5 regions, questioning the specific contributions of FoxA2 to Cyp2a5 gene regulation (Bochkis et al., 2008). Future studies will need to address whether HNF4α is able to directly bind and differentially regulate Cyp2a4 and Cyp2a5 promoters to govern sex-biased gene expression patterns. Furthermore, the specific activation of Cyp2a4 and Cyp2a5 by binding of differential transcription factors during development in male and female liver may provide insight about global postnatal sex-biased activation that is known to occur for other mouse Cyp genes (Conforto and Waxman, 2012).

In addition to being sex-biased, we found that Cyp2a4 and Cyp2a5 are zonally regulated; Cyp2a4 is periportally enriched whereas Cyp2a5 is pericentrally enriched. A previous study by Braeuning and colleagues demonstrated that the majority of Cyp transcripts are pericentrally enriched in adult C3H mice (Braeuning et al., 2006). However, these studies failed to separate Cyp2a4 and Cyp2a5 mRNA and measured Cyp2a4/2a5 transcript simultaneously due to microarray probe design. As a result, both Cyp2a4 and Cyp2a5 transcripts were thought to be enriched in pericentral hepatocytes. In addition, these studies relied upon digitonin/collagenase perfusion of the liver to isolate periportal and pericentral hepatocyte populations; this method can lead to significant
cross contamination when trying to effectively separate hepatocytes for subsequent analysis. More recently, a combined approach using single cell RNA sequencing and single molecular fluorescence in situ expression measurements of landmark genes has allowed for inferred spatial reconstruction of global zonation within the liver (Halpern et al., 2017). In agreement with our data, Cyp2a4 mRNA was found to be periportally enriched, whereas Cyp2a5 mRNA levels were found to be pericentrally enriched.

Promoter analysis of Cyp2a4 and Cyp2a5 did identify one known transcription factor associated with zonal regulation that was predicted to differentially bind Cyp2a4 or Cyp2a5; HNF4α. Recently, HNF4α was demonstrated to exhibit signaling cross-talk between TCF4 and β-catenin (β-cat), an essential transcriptional regulator of zonal gene expression (Gougelet et al., 2014). These studies showed that TCF4 is a mandatory partner of β-cat. In addition, TCF4-occupied HNF4α motifs and some β-cat repressed genes (periportal) are activated by HNF4α. JASPAR analysis completed within this chapter, identified a stronger predicted binding site for the Cyp2a4 promoter compared to the Cyp2a5 promoter. Future studies are needed to address whether HNF4α binds differentially to Cyp2a4 and Cyp2a5, and whether β-cat directly regulates Cyp2a4 and Cyp2a5 gene expression in male and female mice.

In this chapter, we identify two antisense RNAs, Cyp2a4as and Cyp2a5as that are nuclear enriched and polyadenylated, positively correlated with Cyp2a4 and Cyp2a5 gene expression patterns, exhibit sex-biased expression earlier in development than their sense counterparts, and are expressed at significantly lower levels than their sense counterparts. Furthermore, Cyp2a4as can exist as five alternatively spliced transcripts whereas Cyp2a5as has one major spliced isoform.
Evaluation of the Cyp2a4as and Cyp2a5as promoter regions identified a significantly higher number of differences that might contribute to differential regulation compared to the Cyp2a4 and Cyp2a5 promoter regions. Among these differences, Stat5α:Stat5β was predicted to bind to the Cyp2a4as promoter, whereas Bcl6 was predicted to bind to both Cyp2a4as and Cyp2a5as promoters, with a slightly higher predicted binding score for the Cyp2a4as promoter. Recent investigations of sex-biased lincRNAs promoter regions identified enrichment for GH regulated TF’s including Stat5, HNF6, FoxA1, FoxA2, Bcl6 and Cux2 (Melia et al., 2016). In male mice, and because Cyp2a4as and Cyp2a5as are female-biased, Stat5 and Bcl6 may preferentially repress Cyp2a4as expression compared to Cyp2a5as.

There are a number of mechanisms by which antisense RNAs could regulate sense mRNA: mechanisms related to transcription, RNA-DNA interactions, RNA-RNA interactions in the nucleus and RNA-RNA interactions in the cytoplasm (Faghihi and Wahlestedt, 2009). Given their nuclear enrichment, Cyp2a4as and Cyp2a5as are most likely not acting within the cytoplasm. Although many of the functional mechanisms of antisense RNAs remain unknown, evidence suggests that some ncRNAs may regulate transcription by serving as “ligands” for transcription factors. One example, involving ncRNAccDN1s transcripts, causes repression of the C CDN1 (Cyclin D1) gene by binding of the TLS (translocated in liposarcoma) protein that inhibits p300 histone acetyltransferase (HAT) activity (Wang et al., 2008). Moreover, ncRNAccDN1s transcripts are present at remarkably low numbers (~2 copies per cell). Although this example illustrates negative regulation of a gene by ncRNA, perhaps Cyp2a4as and Cyp2a5as are able to regulate their sense counterparts by recruitment of transcription
factors that activate gene expression. Furthermore, this example illustrates the functionality of low abundance ncRNA transcripts.

An example of positive regulation by RNA-DNA interaction involves the antisense RNA, HOTTIP. In human primary fibroblasts, HOTTIP activates expression of HOXA genes by recruitment of WDR5 that interacts with histone methyltransferase MLL1, which is required for methylation of histone 3 on lysine 4 (K3K4me3). Furthermore, HOTTIP RNA levels are very low (~0.3 copies per cell) (Wang et al., 2011). Future studies will need to address whether Cyp2a4as and Cyp2a5as are able to recruit complexes that modify Cyp2a4 and Cyp2a5 histone patterns. Given that Cyp2a4as and Cyp2a5as RNA levels are significantly sex-biased preceding sex-biased differential gene expression of Cyp2a4 and Cyp2a5, a model where recruitment of histone modifiers to these regions to enhance sex-biased modifications is plausible.

Lastly, RNA-RNA duplex formation may result in several outcomes such as alternative splicing of mRNA transcript, inhibition of sense mRNA processing, and RNA editing. Although alternatively spliced transcripts are not present for Cyp2a4 and Cyp2a5 based on spliced EST data from the UCSC Genome Browser, future studies will need to determine whether processing of Cyp2a4 and Cyp2a5 mRNAs, including nuclear transport, are affected with loss of Cyp2a4as or Cyp2a5as.

In a study by Ling and colleagues, genome-wide chromatin accessibility between male and female mice was determined by DNAseI hypersensitivity (DHS) assays and high-throughput sequencing (Ling et al., 2010). In addition, males were given continuous GH administration, as a way to mimic female GH patterns, and the affect of continuous GH on chromatin remodeling was measured by DHS. Mapping the DHS regions
surrounding Cyp2a4 and Cyp2a5 shows overlap of DHS sites with three regions of Cyp2a4as transcripts that exhibit strong responses to female GH patterns (Figure 5-19A). In contrast to this, the Cyp2a5 region only shows one strong response DHS, and shows overlap with two regions of Cyp2a5as. Furthermore, males contain only one DHS site upstream of the Cyp2a4 gene (blue) and one DHS site within the Cyp2a4 gene. Because Cyp2a4 gene expression is substantially lower in males and high in females, these data suggest that additional female DHS sites (pink) may be required in order to activate Cyp2a4 gene expression. Likewise, because Cyp2a5 gene expression is higher in females relative to males, upstream DHS sites may be required in order to highly express Cyp2a5 in liver. Additionally, the DHS regions overlapping the promoter region of Cyp2a4as may suggest that Cyp2a4as is strongly regulated by female GH patterns. Interestingly, our analysis of the Cyp2a4as promoter region showed predicted binding for Stat5a:Stat5b in this same region, and this site was not predicted for Cyp2a5as. Future studies will need to address DHS regions during development of mice and whether the DHS region overlapping the start of the Cyp2a4as promoter region correlates with establishment of Cyp2a4as sex-biased patterns at p21. Given this information, there are several possible models that could account for Cyp2a4as and Cyp2a5as regulation of Cyp2a5 and Cyp2a5 gene expression. One possibility is that early developmental Cyp2a4as and Cyp2a5as gene expression acts to positively regulate Cyp2a4 and Cyp2a5 gene expression by the binding of transcriptional factors that activate or repress transcription through enhanced binding of the promoter regions or putative enhancer regions upstream identified by DHS sites (Figure 5-19B). However, the identity of these early developmental transcription factors that are able to activate Cyp2a4as and Cyp2a5as gene expression remains to be
determined. Furthermore, it remains unknown whether DHS sites within \textit{Cyp2a4} and \textit{Cyp2a5} regions dynamically change during development and whether the binding of certain sex-biased factors and remodeling of chromatin may enhance transcription of \textit{Cyp2a4} and \textit{Cyp2a5} in females compared to males.
Figure 5-1. The sequences of the mouse Cyp2a4 and Cyp2a5 genes are highly similar.

A. Dot matrix alignment of the region of DNA extending 20 kbp upstream (-20 kB) and 50 kbp (+60 kB) downstream of the C57BL/6 Cyp2a4 (x-axis) and Cyp2a5 (y-axis) genes identified a ~70 kbp region of strong similarity. The ~8 kbp Cyp genes from the
beginning of exon 1 (+1) to the end of the 3' exon are highlighted by the gray box. **B.** The transcribed regions of the *Cyp2a4* and *Cyp2a5* genes are aligned; the exons are 98.2% identical. Boxes represent exons; protein coding regions are in black. Specific nucleotide differences in the exons are shown in black letters above and below the genes and joined by gray lines. Nucleotide differences that result in amino acid differences are shown along the top; the first amino acid represents those found in C57BL/6 *Cyp2a4*. Those amino acid changes found in the comparison of the *Cyp2a4* and *Cyp2a5* genes in the C57BL/6 strain but not in the 129 strain are designated by asterisks. Differences in position 209 and 365 are critical for the different enzymatic activities of *Cyp2a4* and *Cyp2a5* and are boxed in green.
Figure 5-2. Cyp2a4 and Cyp2a5 are more highly expressed in female liver compared to male liver. A. RT-qPCR for Cyp2a4 or B. Cyp2a5 mRNA in male (blue) and female (red) liver. Male C57BL/6 n=8, male C3H/HeJ n=4, female C57BL/6 n=7, female C3H/HeJ n=4. mRNA abundance was normalized to Serine and Arginine Rich Splicing Factor 4 (SFRS4) mRNA levels. Statistics: Males were compared to females for each strain by Student’s unpaired t-test. *=p<.05, **=p<.01, ***p<.001.
Figure 5-3. Cyp2a4 and Cyp2a5 mRNA levels increase in postnatal mouse liver. A. Cyp2a4 and B. Cyp2a5 mRNA levels in C3H/HeJ male (M, blue) and female (F, red) livers from embryonic day 17.5 (M and F, n=3) postnatal days p1 (M and F n=3), p7 (M and F n=3), p14 (M n=6, F n=3), p21 (M n=5, F n=3), p28 (M n=5, F n=3) and p56 (M and F n=4). Both Cyp2a4 and Cyp2a5 mRNA levels increase in the postnatal liver but only become significantly different between males and females at day 56. Statistics: Two-way ANOVA was used for female to male statistical comparisons. Hash marks indicate longer than 1 week period between timepoints and apply to both graphs even when not indicated between p28 and p56. ***p<.001.
Figure 5-4. Cyp2a4 and Cyp2a5 mRNA display different zonal expression patterns. Periportal (PP) and pericentral (PC) hepatocytes were separated, based on H2-Dd expression, from E3-βgl-Dd transgenic mouse livers by flow activated cell sorting (FACS) after staining with H2-Dd FITC antibody. 8-week male and female E3-βgl-Dd transgene livers were perfused and Dd positive (PC) and negative (PP) hepatocyte populations were separated. Unsorted (Uns) populations were used as a control. A. Representative FACS data from 8-week female hepatocytes incubated with IgG control antibody (left) or H2-Dd FITC antibody (right). R8 denotes the Dd negative hepatocytes and R7 denotes the Dd positive hepatocytes. B. RNA from unsorted (Uns) and sorted (PP and PC) cell populations was analyzed for expression of a known PP gene (Sds) and a known PC gene (OAT) [male, n=2, female, n=2, combined]. C. Cyp2a4 and D. Cyp2a5 mRNA levels in male (n=2) and female (n=2) sorted hepatocytes. Female Cyp2a4 mRNA is periportally enriched whereas male and female Cyp2a5 mRNA is pericentrally enriched. Statistics: Statistical comparisons within genes were completed using Ordinary One-way ANOVA. *=p<.05, **=p<.01.
Figure 5-5. Cyp2a4 periportal and Cyp2a5 pericentral expression patterns in the CCl₄ experimental mouse model. RNA was prepared from C3H whole liver tissue three days after 8-month old male mice had been injected interperitoneally with either mineral oil (MO, n=5) or carbon tetrachloride (CCl₄, n=4). The RNA was analyzed by RT-qPCR for expression of β-catenin (not zonally expressed), Sds (known PP gene), OAT (known PC gene), Cyp2a4, and Cyp2a5; SFRS4 was used to normalize. Statistical comparisons using Student’s unpaired t-test. *=p<.05, **=p<.01, ***p<.001.
Figure 5-6. JASPAR predictions of transcription factor binding sites within the mouse Cyp2a4 and Cyp2a5 promoters. Overview of predicted transcription factor binding sites specific to Cyp2a4 and Cyp2a5 promoter regions 2000 bp upstream of the mouse Cyp2a4 and Cyp2a5 genes (+1). Nucleotide differences between Cyp2a4 and Cyp2a5 are shown and gaps in sequence are indicated by (-). Specific transcription factors (TF’s) predicted to bind differentially to Cyp2a4 or Cyp2a5 upstream regions are shown. Asterisks next to transcription factors represent stronger predicted binding scores than the corresponding sequences in the other gene. Unique TF’s that are predicted to bind to the Cyp2a4 promoter are highlighted in green, unique TF’s that are predicted to bind to the Cyp2a5 promoter are highlighted in gold.
Figure 5-7. An EST annotated antisense to the Cyp2a4 gene. Screenshots from the UCSC Gene Browser for the regions surrounding the Cyp2a4 and Cyp2a5 genes are shown. RP23267120.3 (RP23) is designated as a transcript level 5; no single transcript from the Gencode database supports the Ensembl EST structure. Cyp2a5 does not contain an annotated EST.
Figure 5-8. Initial detection of the Cyp2a4 antisense transcript in male and female mouse hepatocytes. A. Primers designed to detect the Cyp2a4as EST RP23 transcript are shown (P1 and P2 arrowheads). The EST RP23 (green boxes) is shown aligned to the Cyp2a4 gene (exons, numbered black boxes; 4 of 9 exons are shown). Black arrows denote the transcriptional direction of the Cyp2a4 gene and the Cyp2a4as transcript. B. RT-PCR products from unsorted, U, and PP or PC sorted D<sup>d</sup> hepatocyte populations from male (M) and female (F) mice. Amplicons were resolved on a 1% agarose gel containing ethidium bromide (left panel) or 6% acrylamide gel, followed by ethidium bromide staining (right panel). Migration of the 100 bp ladder 100 and 200 bp fragments are shown. Female amplicons showed a prominent top (top arrow) and bottom band (lower
arrow) in all populations, whereas male amplicons showed a prominent bottom band in all populations and faint upper amplicon in the Dd negative hepatocyte population. In addition, the lower amplicons in the male samples appeared slightly larger than those in the female samples when resolved on the acrylamide gel. C. The Cyp2a4as primers (P1, P2) were not designed to distinguish antisense transcripts from the Cyp2a4 and Cyp2a5 genes if similar transcripts were present in both genes. The nucleotide differences between the Cyp2a4 and Cyp2a5 upstream regions spanning the Cyp2a4as are shown by gray lines. Primers P1 and P2 primer locations are shown.
Figure 5-9. RT-PCR analysis of Cyp2a4as and Cyp2a5as transcripts and regions of similarity between Cyp2a4as and Apob genes. A. Diagram showing polymorphic regions of Cyp2a4 and Cyp2a5 loci encoding Cyp2a4as and Cyp2a5as and location of
primers used for PCR amplification. The sequence in this region of Cyp2a4 and Cyp2a5 genes are nearly identical, with only two SNPs and a single 4 bp gap. Primers were designed to best utilize these differences. Cyp2a4as (P3, P4) and Cyp2a5as (P5, P6) primers are shown for comparison against the original Cyp2a4as primers (P1, P2). B. Melting curve analysis of RT-qPCR amplicons. Melting curves (from 65°C to 95 °C, with 0.5 °C increase every 5 sec), for two male (blue) and two female (red) mice are shown for Cyp2a4as (P1, P2; top panel), Cyp2a4as (P3, P4; middle panel), and Cyp2a5as (P5, P6; bottom panel). C. Analysis of RT-PCR amplicons with Cyp2a4as (P1, P2), Cyp2a4as (P3,P4), Cyp2a5as (P5, P6) and SRFS4 (control) primer pairs. Adult liver RNA from male (M) or female (F) C57BL/6 mice was used for RT-PCR and the amplicons were resolved using 1.2% agarose (top panel), 2.0% agarose (middle panel) or 6% acrylamide gels, as labeled. Migration of the 100 bp and 200 bp fragments of the 100 bp ladder are labeled. Samples from two male and two female mice are shown in the agarose gels, samples from a single male or female mouse are shown in the acrylamide gel. Amplicons of an unexpected size are most evident in the 2.0% agarose and 6% acrylamide gels. Bands from all samples were cloned and sequenced, except the top band from Cyp2a4as (P1&P2) females. Bands matching Cyp2a4as, Cyp2a5as, and ApoB are marked (*). D. Alignment of Cyp2a4as (P3, P4) and Cyp2a5as (P5, P6) primers with the section exon 26 of the Apob gene that Cyp2a4as primers misamplified. E. Annealing temperature affects PCR amplification with Cyp2a4as (P3, P4) primers. cDNA from a single male or female liver was amplified with Cyp2a4as (P3, P4) primers with annealing temperatures that ranged from 60°C to 70°C. As a control, the same cDNA was amplified with Cyp2a5as primers at 60°C. The slower migrating, Apob amplicon is most prominent at 60°C and is
gradually lost with increasing annealing temperature and is no longer evident at 70°C. In contrast, the faster migrating Cyp2a4as amplicon increases in intensity at higher temperatures in the female cDNA sample.
Table 5.1. ApoB1 versus Cyp2a4as and Cyp2a5as average RNA abundance in C3H 8-week male and female liver.

<table>
<thead>
<tr>
<th>Gene Ratio</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApoB1/Cyp2a4as</td>
<td>694000</td>
<td>1905</td>
</tr>
<tr>
<td>ApoB1/Cyp2a5as</td>
<td>6200</td>
<td>1068</td>
</tr>
</tbody>
</table>
Figure 5-10. 5’ RACE for Cyp2a4as and Cyp2a5as transcripts. A. Primer strategy for Cyp2a4as and Cyp2a5as initial 5’ RACE. Shared primer (P7) was used for cDNA synthesis from C57BL/6 female whole liver RNA. Primers P8 and P9 were used for a
first and nested amplification, respectively for Cyp2a4as while primers P6 and p10 were
used for a first and nested amplification, respectively for Cyp2a5as. The exons that match
the identified Cyp2a4as EST are shown in green; sequences from Cyp2a5as, based on
sequence from a RT-PCR clone, are shown in gold. Nucleotide differences between
Cyp2a4as and Cyp2a5as transcripts are shown by gray lines and were utilized for
Cyp2a4as-specific and Cyp2a5as-specific primer design. B. 1% agarose gel of nested
PCR products shows Cyp2a4as (P9) and Cyp2a5as (P10) specific amplicons. All bands
were cloned and sequenced; clones matching Cyp2a4as are marked accordingly (*).
Clones that did not match either gene are left unmarked. C. Cyp2a4as and Cyp2a5as
second 5’ RACE. The first Cyp2a4as and Cyp2a5as 5’ RACE reactions identified
additional upstream 5’ exons, which have been added to Cyp2a4as (green) and Cyp2a5as
(gold). Cyp2a4as (P9) or Cyp2a5as (P6) specific primers were used for cDNA synthesis
from C57BL/6 female whole liver RNA. Cyp2a4as (P11) or Cyp2a5as (P10) primers
were used for PCR, followed by Cyp2a4as (P12, which spans two antisense exons) or
Cyp2a5as (P13 or P14, which spans two antisense exons) for the nested PCR. Asterisk in
front of P11 denotes that no nucleotide differences are present in primer design.
Cyp2a5as specific primer (P14) was designed to span two antisense exons, however, the
farthest upstream Cyp2a5as exon (striped box) had not been identified from previous
sequencing. Nucleotide differences between Cyp2a4as and Cyp2a5as transcripts are
marked by gray lines. D. 1% agarose gel of nested PCR products shows Cyp2a4as and
Cyp2a5as (P13 or P14) PCR amplicons that were cloned and sequenced. One sequenced
cloned matched Cyp2a4as (*), whereas the other sequenced clones did not match
Cyp2a4as or Cyp2a5as (unmarked). Some amplicons were not cloned and sequenced because they were too faint for gel extraction. Diagrams are not to scale.
Figure 5-11. 3’ RACE identifies the 3’ ends of Cyp2a4as and Cyp2a5as transcripts.

A. Primer strategy for 3’ RACE reactions, using tagged oligo dT for the RT reaction from C57BL/6 female whole liver RNA and gene-specific primers for the first Cyp2a4as (P16, P18, P19) and Cyp2a5as (P17, P5, P20) nested Cyp2a4as (P19) and Cyp2a5as (P20) PCR reactions. Nucleotide differences between Cyp2a4as and Cyp2a5as transcripts are marked.
by gray lines. **B.** 1% agarose gel of first PCR products shows multiple amplicons in the Cyp2a4as and Cyp2a5as 3’ RACE reactions. The 100 bp ladder is designated. **C.** 1% agarose gel of second PCR products shows single amplicons in both 3’ RACE reactions. Aliquots of P18 and P5 reactions were used in nested PCR reactions for Cyp2a4as (P19) and Cyp2a5as (P20), respectively. Migration of the 100 bp ladder is designated. Amplicons were isolated, cloned and sequenced (*); both transcripts end at the same location, suggesting the same conserved cleavage polyadenylation signals are being used. Diagrams are not to scale.
Figure 5-12. RT-PCR of full length Cyp2a4as and Cyp2a5as transcripts. A.
Cyp2a4as (P8) and Cyp2a5as (P6) specific primers and a common primer (P15) were designed to amplify potential full length antisense transcripts from C57BL/6 male (M) and female (F) nuclear RNA. Nucleotide differences between Cyp2a4as and Cyp2a5as are marked by gray lines B. RT-PCR reactions from different annealing temperatures (60, 65, 70°C) were resolved using a 1% gel agarose gel. Amplicons isolated, cloned and sequenced are marked for Cyp2a4as (bracket, *) and Cyp2a5as (*). The Cyp2a5as amplicons marked with # matched the Cyp2a5as sequence. Multiple, different sized amplicons were cloned from the broad band marked *, all matching the Cyp2a4as sequence and identifying multiple alternatively processed RNAs. The 100 bp ladder is designated. Diagrams are not to scale.
Figure 5-13. Diagram of the Cyp2a4as spliced isoforms and the single Cyp2a5as spliced isoform identified by cloning and sequencing RT-PCR products. Sequencing analysis from nuclear RT-PCR and 5’/3’ RACE identified five Cyp2a4as transcript variants in female liver and a single Cyp2a5as transcript variant in male and female liver. Cyp2a4 (bottom) and Cyp2a5 (top) exons 1-5 (of a total of 9 exons) are labeled and arrows depict transcriptional directionality. Cyp2a4as and Cyp2a5as transcripts have transcriptional start sites within Cyp2a4 or Cyp2a5 Exon 4. Four Cyp2a4as isoforms contain six exons (white boxes unlabeled), and one isoform contains seven exons; introns of the antisense transcripts are denoted by dotted lines. The Cyp2a5as transcript contains seven exons. The highlighted gray sections indicate use of the same splice junctions. Vertical dotted lines indicate differences between Cyp2a4as and Cyp2a4 exons and among the antisense exons. Cyp2a4as and Cyp2a5as contain the same 3’ ends, using the same cleavage-polyadenylation sites. Diagram is drawn to scale.
Table 5.2. Coding Potential Assessment Tool (CPAT) evaluation of the Cyp2a4as and Cyp2a5as transcripts.

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>RNA size</th>
<th>ORF size (n.t.)</th>
<th>Coding Probability</th>
<th>Coding Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyp2a4as TV1</td>
<td>1098</td>
<td>174</td>
<td>0.25</td>
<td>no</td>
</tr>
<tr>
<td>Cyp2a4as TV2</td>
<td>1157</td>
<td>261</td>
<td>0.35</td>
<td>no</td>
</tr>
<tr>
<td>Cyp2a4as TV3</td>
<td>1172</td>
<td>150</td>
<td>0.30</td>
<td>no</td>
</tr>
<tr>
<td>Cyp2a4as TV4</td>
<td>1242</td>
<td>150</td>
<td>0.30</td>
<td>no</td>
</tr>
<tr>
<td>Cyp2a4as TV5</td>
<td>1333</td>
<td>150</td>
<td>0.30</td>
<td>no</td>
</tr>
<tr>
<td>Cyp2a5as</td>
<td>1235</td>
<td>324</td>
<td>0.24</td>
<td>no</td>
</tr>
<tr>
<td>H19</td>
<td>2288</td>
<td>444</td>
<td>0.30</td>
<td>no</td>
</tr>
<tr>
<td>Cyp2a4</td>
<td>1715</td>
<td>1485</td>
<td>0.99</td>
<td>yes</td>
</tr>
<tr>
<td>Cyp2a5</td>
<td>1746</td>
<td>1485</td>
<td>0.99</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 5.2. Coding Potential Assessment Tool (CPAT) evaluation of the Cyp2a4as and Cyp2a5as transcripts. The CPAT algorithm takes into consideration the maximum length of open reading frame (ORF) in nucleotides, ORF coverage of transcript length, Fickett score, and hexamer usage bias to determine whether RNA is coding or non-coding (Wang et al., 2013b). The Fickett score compares composition of bases between a given transcript against mRNA of 130 different species to come up with coding probability. Hexamer usage bias compares six nucleotide pieces within a given transcript and compares it to a large training set of coding and non-coding regions to determine coding probability. H19, a known long non-coding RNA, and Cyp2a4 and Cyp2a5 mRNAs, were used as controls within the algorithm. Each Cyp2a4as alternatively spliced RNAs and Cyp2a5as RNA was then entered and the coding label was determined based on the factors described above.
Figure 5-14. Cyp2a4as and Cyp2a5as expression patterns mimic those of their overlapping mRNA counterparts. RNA was prepared from whole liver tissues or separated hepatocyte populations and analyzed by RT-qPCR; SFRS4 was used to normalize. Cyp2a4as RNA expression was analyzed using P3 and P4 primers and 70°C annealing temperature. Cyp2a5as RNA levels were analyzed using P5 and P6 primers and 60°C annealing temperature. A. Cyp2a4 and Cyp2a4as and Cyp2a5 and Cyp2a5as RNA levels in C57BL/6 male livers (n=8) and female livers (n=7). B. C3H/HeJ Male (M, blue) and female (F, red) Cyp2a4as and Cyp2a5as RNA levels from embryonic day 17.5 (M and F, n=3), postnatal days p1 (M and F n=3), p7 (M and F n=3), p14 (M n=6, F
n=3), p21 (M n=5, F n=3), p28 (M n=5, F n=3) and p56 (M and F n=4) livers. Hash marks indicate longer than 1 week period between timepoints. C. Male (n=2) and female (n=2) Cyp2a4 as and Cyp2a5 as zonal gene expression patterns in Dd separated hepatocyte populations. Cyp2a4 as is enriched in periportal hepatocytes (PP) in female mice, whereas Cyp2a5 as is enriched in pericentral (PC) hepatocytes. Statistics: A. Student’s unpaired T-test, B. Ordinary Two-way ANOVA, C. Ordinary One-way ANOVA within each sex. *=p<.05, **=p<.001.
Table 5.3. Ratio among Cyp2a4, Cyp2a5, Cyp2a4as, and Cyp2a5as average RNA levels in C57BL/6 male and female liver.

<table>
<thead>
<tr>
<th>Gene Ratio</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyp2a4/2a5</td>
<td>0.001</td>
<td>0.907</td>
</tr>
<tr>
<td>Cyp2a4/2a4as</td>
<td>116</td>
<td>503</td>
</tr>
<tr>
<td>Cyp2a5/2a5as</td>
<td>829</td>
<td>624</td>
</tr>
<tr>
<td>Cyp2a4as/2a5as</td>
<td>0.007</td>
<td>1.13</td>
</tr>
</tbody>
</table>

Table 5.4. Ratio of Cyp2a4, Cyp2a5, Cyp2a4as, and Cyp2a5as average RNA levels between C57BL/6 female and male liver.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Female/Male Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyp2a4</td>
<td>928</td>
</tr>
<tr>
<td>Cyp2a5</td>
<td>2.33</td>
</tr>
<tr>
<td>Cyp2a4as</td>
<td>500</td>
</tr>
<tr>
<td>Cyp2a5as</td>
<td>3.09</td>
</tr>
</tbody>
</table>
Figure 5-15. JASPAR predictions of transcription factor binding sites within the potential promoters of mouse Cyp2a4as and Cyp2a5as. Overview of predicted transcription factor binding sites specific to Cyp2a4as and Cyp2a5as promoter regions 2000 bp upstream of 5’ RACE sequenced ends (+1). Nucleotide differences between Cyp2a4as and Cyp2a5as are shown and gaps in sequence are indicated by (-). Larger gaps in sequence are denoted as breaks within the Cyp2a4as sequence. Specific transcription factors (TF’s) predicted to bind differentially to Cyp2a4as or Cyp2a5as upstream regions are shown. Asterisks next to transcription factors represent stronger predicted binding scores than the corresponding sequences in the other gene. Unique TF’s that are predicted to bind to the Cyp2a4as promoter are highlighted in green, unique TF’s that are predicted to bind to the Cyp2a5as promoter are highlighted in gold.
Figure 5-16. Cyp2a4as and Cyp2a5as are nuclear-enriched transcripts. C57BL/6 male (n=2) and female (n=2) liver isolates were separated into cytoplasmic and nuclear fractions. RNA was isolated and analyzed by RT-qPCR. Spliced mRNAs for ribosomal protein L30 (L30), Cyp2a4, and Cyp2a5 were used as positive controls for cytoplasmically-enriched transcripts. Unspliced L30 (L30-U) and U6 small nuclear RNA (U6) were used as positive controls for nuclear-enriched transcripts. Transcript abundance was first calculated between cytoplasmic and nuclear fractions (C/N), and then normalized to this ratio for L30. Male and female nuclear to cytoplasmic ratios were combined since no statistically significant C/N ratio differences were found between them. L30-U, U6, Cyp2a4as, and Cyp2a5as C/N ratios were significantly different from L30, Cyp2a4, and Cyp2a5, indicating enrichment in nuclear fractions. Statistical testing was completed using Ordinary One-way ANOVA ***p<.001.
Figure 5-17. Cyp2a4as and Cyp2a5as expression patterns are similar to Cyp2a4 and Cyp2a5 in DEN-induced tumors. A. Comparison of Cyp2a4 and Cyp2a5 mRNA levels in liver tumors induced by DEN in 34-week old C57BL/6 male mice. PBS injected control mice (n=8), Non-tumor from DEN-injected mice (NT, n=9), Tumor from DEN-injected mice (T, n=9). Expression of Cyp2a4, but not Cyp2a5, is higher in tumors compared to normal tissue. B. Comparison of Cyp2a4as and Cyp2a5as RNA levels in DEN-induced male tumors. Statistics: One-way ANOVA was used for statistical comparisons between groups. ***p<.001.
Figure 5-18. Additional Cyp2 family antisense and intergenic RNA transcripts are not sex-biased and are expressed at very low levels in mouse liver. Transcript abundance of Cyp2a4as and Cyp2a5as is compared to the non-coding RNA H19 and to the newly detected Cyp2s1as and intergenic RNA RP24 transcripts in C57BL/6 male (n=5) and female (n=5) whole liver RNA. H19, Cyp2s1as, and RP24 RNA are not sex-biased. Statistical testing was completed using Student’s t-test between sexes for each RNA. *=p<.05, ***p<.001.
Figure 5-19. Cyp2a4 and Cyp2a5 DNAseI hypersensitivity regions in females overlap with the Cyp2a4as promoter. A. Comparison of DNAseI Hypersensitivity (DHS)
regions in 8-week old CD-1 mice (data gathered from (Ling, Sugathan, Mazor, Fraenkel, & Waxman, 2010). DHS regions are marked specifically for female (pink) or male (blue) and response to continuous GH in 7-week males is indicated above (strong, weak, unresponsive). Exons for the \textit{Cyp2a4} and \textit{Cyp2a5} genes are marked by numbered black boxes, whereas Cyp2a4as and Cyp2a5as transcripts are marked below gene regions by gray boxes. Spliced products are indicated by dotted lines for Cyp2a4as and Cyp2a5as transcripts. Gene regions are numbered according to distance from TSS. Diagrams are not to scale. \textbf{B.} A model for Cyp2a4as regulation of Cyp2a4 in hepatocyte nuclei. Transcriptional direction of Cyp2a4as is shown by the black arrow below Cyp2a4 exon 4. At p21, female-biased Cyp2a4as are proposed to be activated by a female-specific TF (dark pink circle) and a sex-independent TF (teal circle) common to both males and females. Between p28 and p56, continuous growth hormone may contribute to activation of female-biased TF’s which may positively regulate \textit{Cyp2a4as} gene expression. Cyp2a4as transcripts (green lines) may act to regulate \textit{Cyp2a4} transcription through complementary binding of sequences within DHS open regions near \textit{Cyp2a4}, and binding specific TF’s (gray circles) that lead to chromatin remodeling and enhanced transcription of \textit{Cyp2a4}. The transcription core machinery is also depicted (yellow circle). In males at p28, growth hormone pulsing is proposed to activate male-biased TF’s (red circles) that block basal \textit{Cyp2a4as} gene expression. By p56, it is unknown whether chromatin is still accessible to the male biased TF’s to block \textit{Cyp2a4as} gene expression or whether GH pulsing affects chromatin accessibility. \textbf{C.} A model for \textit{Cyp2a5as} regulation of \textit{Cyp2a5} in hepatocyte nuclei. Transcriptional direction of Cyp2a5as is shown by the black arrow below \textit{Cyp2a5} exon 4. At p21, a sex-independent transcription factor (teal circle), is
proposed to activate $Cyp2a5as$ gene expression and the production of $Cyp2a5as$ transcripts (gold lines) in males and females. The $Cyp2a5as$ promoter is proposed to be accessible at p21 through p56 because we were able to measure $Cyp2a5as$ at these timepoints in both males and females. Female-biased TF’s are proposed to enhance $Cyp2a5as$ gene expression in females. Females contain two additional DHS regions upstream of $Cyp2a5$ compared to males. These sites are proposed to enhance transcription of $Cyp2a5$ in females. In contrast, males lack DHS regions and female biased TF’s that may act to positively regulate $Cyp2a5as$ gene expression.
CHAPTER VI

SUMMARY AND FUTURE DIRECTIONS

Data in this dissertation focused on two major themes, gene duplication and gene regulation, centered around the Zhx gene family and their gene targets. In chapter three, I investigated the evolutionary relationship of Zhx genes. From my results, I hypothesize that an ancestral Zhx gene arose early in chordate evolution and that this primordial gene duplicated to give rise to a Zhx3-like (Zhx3L) and a Zhx1-like (Zhx1L) gene. A second duplication of the ancestral Zhx1L gene gave rise to Zhx1L and Zhx2L genes. Based on the current sea lamprey assembly, Zhx1L and Zhx2L fused and created a Zhx pseudogene (ZhxP). Chapters four and five focused on two pairs of duplicated genes that are differentially regulated by Zhx2 and at the sex-biased level, with the goal of understanding this differential control. In chapter four, I focused on the male-biased regulation of Slp by Stat5b and Zhx2. Using available large datasets, I found that Zhx2 controls fewer sex-biased genes compared to the known regulators Stat5b and HNF4α. In chapter five, I identified two long non-coding RNAs that may contribute to differential Cyp2a4 and Cyp2a5 expression.

Gene duplications create potential for generating genes with novel function and can occur through several known mechanisms, including whole genome duplication (WGD) or segmental duplication (SD) events. Evidence for two WGD in chordate evolution comes from gene number comparisons between early chordates and vertebrates; two WGD represent the simplest explanation for the increase in gene family members from lancelet to human (Putnam et al., 2008). Following WGD, some genes
may adopt a new function or partition old functions, but the majority of duplicated genes are lost as a result of selective pressures in an organism’s environment. Large-scale gene loss can be seen by comparing the number of genes across the chordate lineage (Huang et al., 2014). In contrast to WGD, SDs are duplications of smaller DNA sections (from a few base pairs up to many megabases) that can result in tandem, separated, or interspersed duplicated genes along a chromosome. While WGD can be the result of genomic doubling, gametic non-reduction (i.e. non-dysjunction), or polyspermy, SDs are generally the result of unequal crossing over or transposition events.

Previous analyses indicated that C4 and Slp are duplicated genes based on their homology, which extends into flanking regions, yet they have very different expression patterns, with Slp being strongly male-biased and C4 being expressed equally in both males and females. Slp expression is higher in Zhx2-deficient mice than normal mice. Male-biased Slp expression is mediated by an endogenous proviral element that is absent upstream of C4 (Robins, 2004). It is not yet clear how Zhx2 contributes to Slp regulation. Duplication to generate C4 and Slp, which likely was a result of nonhomologous recombination, is estimated to have occurred roughly two to three million years ago (Aida et al., 1994). Swapping of regulatory regions revealed that some mouse strains contain Slp under control of C4 regulatory elements leading to constitutive Slp expression. Similar to C4 and Slp, the Cyp2 subfamilies likely arose from duplications of ancestral Cyp2a, Cyp2g, and Cyp2b genes.

Within the Cyp2a subfamily, Mus spretus only contains the Cyp2a5 gene whereas most domestic mouse strains have both Cyp2a4 and Cyp2a5 genes, suggesting that Cyp2a5 gave rise to Cyp2a4 through a duplication event within domesticated mice. The
high level of sequence homology between Cyp2a5 and Cyp2a4 is consistent with a recent duplication event. Few studies to date have evaluated differences between Cyp2a4 and Cyp2a5 regulation and have largely focused on Cyp2a5 expression. My work has documented multiple ways in which these genes are differentially controlled, including different responses to Zhx2 loss.

The Zhx3L gene in sea lamprey lies on a separate scaffold compared to ZhxP, which suggests that a whole genome duplication of a primordial Zhx gene gave rise to Zhx3L and Zhx1L/2L, and that a common ancestor of Zhx1L/2L underwent a SD event to give rise to Zhx1L and Zhx2L. Our current data shows that Zhx3L but not ZhxP is expressed in sea lamprey, although further analysis may find that ZhxP may be expressed at low levels or in tissues that we have not analyzed. Zhx3L RNA levels in sea squirts and lancet remain to be determined. To date, it is not known what contributes to Zhx gene regulation in any organism, although all Zhx genes are ubiquitously expressed, albeit at different levels in different organs, in mouse and human. One question is what evolutionary opportunities are created by Zhx gene duplication in sea lamprey? For instance, if Zhx3L is the only Zhx gene expressed in sea lamprey, lancelet, and sea squirt, does this suggest that ZhxP confers no evolutionary advantage since it does not appear to be expressed in sea lamprey? Further examination of the sea lamprey Zhx gene structure is needed, apart from gene synten analysis, for identification of flanking regions that may contribute to potential differential ZhxP and Zhx3L expression. Identifying open chromatin regions by DNase I hypersensitivity mapping or ATAC-seq may provide clues about regulatory regions present within sea lamprey, including scaffolds which house ZhxP and Zhx3L.
Another observation stemming from the work presented in this dissertation is that Zhx genes appear earlier in evolution than at least some of the genes they regulate (AFP, Slp, Cyp2a4, and Cyp2a5). The primordial Zhx gene appears within the early chordate lineage, while Zhx2 appeared later, following a WGD and then a SD. However, the Zhx2 target genes Slp and Cyp2a4 arose much later and are found only in domesticated mice. It is not known when the first common ancestor of Cyp2 genes appeared during chordate evolution. Additionally, while Zhx2 regulates some but not all Cyp genes that have been examined to date, the full extent of mouse Cyp gene regulation (105 Cyp genes) by Zhx2 or other Zhx family members has not been investigated. Future studies evaluating Cyp51 regulation by Zhx1, Zhx2, or Zhx3 may help elucidate whether regulators appear earlier in evolution compared to genes that they regulate; if Cyp51 is regulated Zhx3, and because Cyp51 is proposed to be the primordial Cyp gene (Nelson et al., 2013), this data would suggest that regulators may be co-opted to regulate genes during chordate evolution. Alternatively, it would be interesting to perform analyses within Zhx1 and Zhx3 knockout mice to identify genes that are dysregulated and trace origins of these genes throughout chordate evolution by genome assembly analyses. If certain genes are dysregulated by individual Zhx1, Zhx2, or Zhx3 loss, this would suggest partitioning of regulator function in mice. Future studies will need to identify lamprey Zhx3L gene targets, potentially by silencing of lamprey Zhx3L in early chordates by morpholino antisense nucleotides (McCauley and Bronner-Fraser, 2006) followed by either RT-qPCR or RNA-sequencing analysis.

It would also be interesting to correlate single nucleotide polymorphisms (SNPs) within transcription factor sequences (in this case, Zhx genes) with the appearance of new
gene target sequences in different species; this type of analysis can be applied to a large number of transcriptional regulators. For example, do Zhx SNPs coincide with the appearance of new Zhx targets during chordate evolution, and are certain Zhx SNPs retained in some common chordate lineages? Lastly, although we have only begun to scratch the surface in our understanding of the complexity of gene activation under different environmental conditions, identifying when certain transcription factors appeared during evolution might provide insight about the origin of new gene expression regulatory networks. Adding to this point, Stat5b, HNF4α, and Bcl6 are annotated within the most recent sea lamprey genome assembly. Sea lamprey Stat5b and HNF4α gene targets remain to be discovered and these studies may have implications into the origin of sex-biased gene regulation in liver.
### APPENDIX TABLE 1.

<table>
<thead>
<tr>
<th>Species</th>
<th>Zhx Protein</th>
<th>NCBI Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alligator</td>
<td>Zhx1</td>
<td>XP_014459348.1</td>
</tr>
<tr>
<td></td>
<td>Zhx2</td>
<td>XP_006269908.1</td>
</tr>
<tr>
<td></td>
<td>Zhx3</td>
<td>XP_014465170.1</td>
</tr>
<tr>
<td>Anole</td>
<td>Zhx1</td>
<td>XP_003219449.1</td>
</tr>
<tr>
<td></td>
<td>Zhx2</td>
<td>XP_003228184.1</td>
</tr>
<tr>
<td></td>
<td>Zhx3</td>
<td>XP_008115797.1</td>
</tr>
<tr>
<td>Chicken</td>
<td>Zhx1</td>
<td>XP_025003919.1</td>
</tr>
<tr>
<td></td>
<td>Zhx2</td>
<td>XP_015138646.1</td>
</tr>
<tr>
<td></td>
<td>Zhx3</td>
<td>XP_015152236.1</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>Zhx1</td>
<td>NP_001074953.1</td>
</tr>
<tr>
<td></td>
<td>Zhx2</td>
<td>XP_519932.2</td>
</tr>
<tr>
<td></td>
<td>Zhx3</td>
<td>XP_024207782.1</td>
</tr>
<tr>
<td>Ciona Intestinalis</td>
<td>Zhx</td>
<td>NP_001071852.1</td>
</tr>
<tr>
<td>Western Clawed Frog</td>
<td>Zhx1</td>
<td>XP_004915112.1</td>
</tr>
<tr>
<td></td>
<td>Zhx2</td>
<td>XP_012819625.1</td>
</tr>
<tr>
<td></td>
<td>Zhx3</td>
<td>NP_001006839.1</td>
</tr>
<tr>
<td>Coelacanth</td>
<td>Zhx1</td>
<td>XP_006007084.1</td>
</tr>
<tr>
<td></td>
<td>Zhx2</td>
<td>XP_006007079.1</td>
</tr>
<tr>
<td></td>
<td>Zhx3</td>
<td>XP_005986804.1</td>
</tr>
<tr>
<td>Cow</td>
<td>Zhx1</td>
<td>NP_001103565.1</td>
</tr>
<tr>
<td></td>
<td>Zhx2</td>
<td>NP_001193335.2</td>
</tr>
<tr>
<td></td>
<td>Zhx3</td>
<td>XP_002692510.1</td>
</tr>
<tr>
<td>Dog</td>
<td>Zhx1</td>
<td>XP_022282389.1</td>
</tr>
<tr>
<td></td>
<td>Zhx2</td>
<td>XP_539154.2</td>
</tr>
<tr>
<td>Animal</td>
<td>Zhx1</td>
<td>Zhx2</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Elephant Shark</td>
<td>XP_007898748.1</td>
<td>XP_007898736.1</td>
</tr>
<tr>
<td>Guinea Pig</td>
<td>XP_023419828.1</td>
<td>XP_003467414.1</td>
</tr>
<tr>
<td>Horse</td>
<td>XP_023504683.1</td>
<td>NP_001244085.1</td>
</tr>
<tr>
<td>Human</td>
<td>NP_065885.2</td>
<td>NP_009153.3</td>
</tr>
<tr>
<td>Lancelet</td>
<td>NP_055850.1</td>
<td>XP_002597641.1</td>
</tr>
<tr>
<td>Minke Whale</td>
<td>XP_007188914.1</td>
<td>XP_007188903.1</td>
</tr>
<tr>
<td>Mouse</td>
<td>NP_898997.2</td>
<td>NP_033598.2</td>
</tr>
<tr>
<td>Opossum</td>
<td>XP_007488328.1</td>
<td>XP_001381184.1</td>
</tr>
<tr>
<td>Pig</td>
<td>XP_001928668.1</td>
<td>XP_020944632.1</td>
</tr>
<tr>
<td>Rabbit</td>
<td>XP_017197002.1</td>
<td></td>
</tr>
</tbody>
</table>

221
<table>
<thead>
<tr>
<th>Animal</th>
<th>Zhx1</th>
<th>Zhx2</th>
<th>Zhx3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rat</td>
<td>XP_008254180.1</td>
<td>NP_001040562.1</td>
<td>NP_001257985.1</td>
</tr>
<tr>
<td>Saker Falcon</td>
<td>XP_005438485.1</td>
<td>XP_005438481.1</td>
<td>XP_005439100.1</td>
</tr>
<tr>
<td>Spotted Gar</td>
<td>XP_010707584.1</td>
<td>XP_010707580.1</td>
<td>XP_010720623.1</td>
</tr>
<tr>
<td>Starling</td>
<td>XP_014725630.1</td>
<td>XP_014725614.1</td>
<td>XP_014739503.1</td>
</tr>
<tr>
<td>Turkey</td>
<td>XP_006636015.1</td>
<td>XP_006636139.1</td>
<td>XP_015220317.1</td>
</tr>
</tbody>
</table>
REFERENCES


Holloway, M.G., Laz, E.V., and Waxman, D.J. (2006). Codependence of growth hormone-responsive, sexually dimorphic hepatic gene expression on signal...
transducer and activator of transcription 5b and hepatic nuclear factor 4alpha. Mol Endocrinol 20, 647-660.
cDNA clones and differential regulation in the liver. Nucleic Acids Res 12, 6073-6090.


C4-Slp gene in mouse liver follows activation of STAT5. Proc Natl Acad Sci U S A 95, 8750-8755.


VITA

March 22, 2019

Alexandra Nichole Nail

CITIZENSHIP
Born in Duluth, MN

EDUCATION
University of Kentucky
Ph.D., Microbiology
August 2013-March 2019
Lexington, KY
U.S.A.

Morehead State University
M.S., Biology
August 2011-May 2013
Morehead, KY
U.S.A.

Morehead State University
B.S., Biology
August 2006-May 2011
Morehead, KY
U.S.A.

WORK EXPERIENCE
Graduate Research Assistant
Brett T. Spear, Ph.D.
University of Kentucky
January 2015-present
(859) 257-5167

Graduate Teaching Assistant
Geoff Gearner, Ph.D.
Morehead State University
2011-2013
(606) 783-5002

COMMITTEE MEMBERSHIPS
Biomedical Graduate Student Organization Vice President
2016-2018
Microbiology, Immunology, and Molecular Genetics Elected Student Representative
2016-2017

HONORS AND AWARDS
University of Kentucky 3-minute thesis competition semi-finalist
2017
ASBMB Graduate Student Travel Award
2017
ASBMB Graduate Student Travel Award
2016
United WAY Back on Track Scholarship Program Recipient
2014-2015
Outstanding Graduate Student College of Science and Technology
2012-2013

PROFESSIONAL SOCIETY MEMBERSHIPS
ASBMB student member
2016-present
Endocrine Society student member
2018-present
SCIENTIFIC PRESENTATIONS

Oral presentations

Microbiology, Immunology, and Molecular Genetics Seminar
Evolution of the Zhx transcription factor family and analysis of Zhx2 targets 2019
Cyp2a4 and Cyp2a5 in mouse liver
Hepatic zonal control of sex-dependent cytochrome p450 enzymes. 2017
Zhx2 regulates sex-biased liver genes. 2016
Heme oxygenase-1: Tipping the balance between healthy and unhealthy obesity. 2014

Regional poster presentations
Alexandra Nail, K. Schroder, K. Bernardo, J.J. Smith, M.L. Peterson, Spear, B.T. 2018
Characterization and evolution of zhx genes in chordates.
Microbiology, Immunology, and Molecular Genetics Annual Retreat.
Lexington, KY. Regional poster presentation.

Microbiology, Immunology, and Molecular Genetics Annual Retreat.
Lexington, KY. Regional poster presentation.

Zhx2: a potential role for sexual dimorphism in liver cancer.
Markey Cancer Center Research Day.
Lexington, KY. Regional poster presentation.

National poster presentations
Alexandra Nail, J. Jiang, M.L. Peterson, Spear, B.T. 2017
E3-βglobin-Dd mice: A new model to study hepatic zonal regulation of sex-biased cytochrome p450 mRNA expression.
American Society for Biochemistry and Molecular Biology (ASBMB) Annual Meeting
Chicago, IL. National poster presentation. (ASBMB Travel Award)

American Society for Biochemistry and Molecular Biology (ASBMB) Annual Meeting
San Diego, CA. National poster presentation. (ASBMB Travel Award)

TEACHING EXPERIENCE
University of Kentucky Critical Readings (1st Year Ph.D. course) August-December 2016
Morehead State University Graduate Teaching Assistant August 2011-May 2013

BIBLIOGRAPHY

Peer-reviewed publications
**Manuscripts in preparation**


**ASBMB Today journal news articles**


