Title

The Role of Nuclear Receptors in Tissue-Specific Gene Expression: The Impact of Genetic Variation on DNA Binding

Permalink

https://escholarship.org/uc/item/2cw7q5hj

Author

Deans, Jonathan R

Publication Date

2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

The Role of Nuclear Receptors in Tissue-Specific Gene Expression: The Impact of Genetic Variation on DNA Binding

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics, and Bioinformatics

by

Jonathan Robert Deans

December 2017

Dissertation Committee:
Dr. Frances Sladek, Chairperson
Dr. Tao Jiang
Dr. Thomas Girke

The Dissertation of Jonathan Robert Deans is approved:

____

____

____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my advisor, Dr. Frances Sladek, who has continually been a great mentor, educator, and scholar. She has given me direction, support, and inspiration throughout the development of this dissertation. I would also like to thank my committee members for their wonderful comments and suggestions throughout the years.

I would like to thank Dr. Nina Titova for her contributions to the protein binding microarray experiments. Without all her hard work on preparing and applying the nuclear extracts I don’t know if any of this would have been possible.

I would like to thank my amazingly intelligent, confident, supportive, and wonderfully loving parents, Linda and Robert, for all their guidance. I only hope that one I have a body of work that can come anywhere close to what they have each accomplished in their own careers.

Finally, I would like to thank my best friend and wife, Katy, for making me laugh every day. Her support, encouragement, and unwavering love have, without a doubt, made these past 5 years the best years of my life. Here’s to our next chapter.

In loving memory of my grandmother, Leonora Walling (1927-2015)

ABSTRACT OF THE DISSERTATION

The Role of Nuclear Receptors in Tissue-Specific Gene Expression: The Impact of Genetic Variation on DNA Binding

by

Jonathan Robert Deans

Doctorate of Philosophy, Graduate Program in Genetics, Genomics, and Bioinformatics
University of California, Riverside, December 2017
Dr. Frances Sladek, Chairperson

Nuclear receptors (NRs) are ligand-sensitive transcription factors that regulate a wide array of biological processes including development, metabolism, and circadian rhythms. All NRs share a common protein structure, including highly conserved DNA binding domains and a highly variable N-terminal A/B domain, and are very popular drug targets. To better understand the role of alternative A/B domains between NR isoforms and the impact of genetic variation on gene expression in the liver, we employed two experimental approaches. The NR hepatocyte nuclear factor 4α (HNF4α), a master regulator of liver-specific gene expression, is regulated by two promoters (P1 and P2) in the liver resulting in proteins with different A/B domains. P1-HNF4α is expressed in fetal and normal adult liver while P2-HNF4α is expressed only in the fetal liver and in liver cancer. We compared wildtype mice, which express only HNF4α1 (P1) in the adult liver, to exon-swap mice that express only HNF4α7 (P2) for global changes in gene expression (RNA-seq), chromatin binding (ChIP-seq), and unique protein interactions (RIME). The results show that P1- and P2-HNF4α isoforms differentially regulate hundreds of transcripts in the adult liver, including the NR CAR (Nr1i3), and may be recruited differentially to non-HNF4α binding sites by unique protein interactions. They also exhibit altered metabolic pathways, especially cytochrome P450 (Cyp) genes. All told, the results show that changes in just 16-30 amino acids in the AF-1 region of an NR can have profound effects on gene expression. Utilizing protein binding microarrays (PBM), we can measure the DNA binding affinity of a given NR against both alleles of 125,000 genetic variants in a single experiment to probe for affinity altering SNPs (aaSNPs). By mining SNPs from ChIP-seq peaks and eQTLs from the GTEx project, we have identified thousands of aaSNPs, hundreds of which show significant correlation to changes in gene expression within their regulatory network. Analysis of aaSNPs from GWAS studies associated with Alzheimer's disease identified a large number of genetic variants that can alter the DNA binding affinity of PPARɣ in the *APOE* locus. Additionally, we show the power of the PBMs to validate many aaSNPs derived from *in vivo* analysis and suggest a role for the PBM technology in characterizing how genetic diversity may play a role in personalized medicine.

Table of Contents

List of Figures

Chapter 1

Introduction

The Human Genome and Genetic Variation

The human genome is comprised of a system of coding and noncoding genetic features that define gene expression patterns across cell types and individuals. Over the last two decades there have been many advances in our understanding of the relationship between genetic variation in phenotypic variation and human disease. Much of this knowledge has come from genome wide association studies (GWAS) that have identified thousands of variants affecting disease and phenotype. The majority of genetic variants identified by GWAS studies are noncoding and may impact their associated phenotypes by altering the regulation of gene expression (Maurano et al., 2012; Ward and Kellis, 2012). It has been shown that noncoding genetic variants can impact tissue-specific phenotypes and play a role in disease susceptibility (Albert and Kruglyak, 2015). However, the molecular mechanisms by which genetic variation can influence gene expression are still poorly characterized. One of the most commonly used methods to characterize these variants is with expression quantitative trait loci (eQTL) that identify loci where genotypes are significantly correlated with patterns of gene expression within a population of individuals (Schadt et al., 2008). Different genotypes between individuals are distinguished by a variety of alterations in the genome. The most common genetic variations are single nucleotide polymorphisms (SNPs), sometimes referred to as single nucleotide variants (SNV) when they are less commonly found in a population. Currently there are well over 10 million common SNPs in the human genome.

The Genotype-Tissue Expression project (GTEx) is a consortium funded by the National Institutes of Health Common Fund (https://www.gtexportal.org/). The goal of the project is to generate eQTL datasets that allow one to study of the relationship between genetic variation and gene expression in human tissues. Currently, the GTEx datasets are comprised of 10,361 RNA-seq samples from 635 donors across 53 tissue types. Nonetheless, these datasets still lack the power to identify potential mechanisms by which these variants impact gene expression. As personalized medicine, and medical genetics, are increasingly used to explore the role of rare and common genetic variants, data such as that in GTEx will be ever more important for the interpretation of the mechanisms by which genetic variants can impact human disease.

Gene expression levels between individuals and cell types are regulated by transcription factors (TF) through sequence-specific interactions with genomic DNA. While chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) allows a researcher to extract and sequence DNA bound by a specific TF to interpret genome-wide occupancy, it can often be challenging to identify the true binding site within a single peak. Additionally, attempting to identify genetic variants that can disrupt or alter DNA binding affinity of a TF via ChIP-seq experiments would require a very large number of samples and would be very time-consuming. Protein binding microarrays (PBM) are high-throughput DNA binding assays. By utilizing high density, custom-designed microarrays extended on the slide to double-stranded oligonucleotides, one can test for TF binding directly to the DNA on the slide with a fluorophore-conjugated antibody. PBMs provide an alternative to ChIP-based assays of TF binding and can be custom designed to cover a broad range of k-mers or small (<30nt) DNA binding sites from the genome. The power of this technology comes from the large number of test sequences that can be spotted in a single slide (80,000-125,000) with replicated sequences throughout the design. By designing two test sequences, each with one allele of a genetic variant (plus flanking region), we can measure in vitro DNA binding affinity of any TF to both alleles and statistically identify variants that can potentially impact DNA binding in vivo.

Another level of genetic diversity with impact on tissue-specific gene expression between cell types is the use of alternative promoters. Alternative promoters are quite common in the human genome and have been verified for approximately 7,000 human genes, and expressed sequence tags (EST) and cap analysis gene expression (CAGE) mappings to the genome suggest there may even be more (Singer et al., 2008). These promoters can function in many ways to produce a wide array of transcripts from just a single gene locus. The primary role of alternate promoters is thought to be the control of gene expression under different cellular conditions, including tissue specific gene-expression (Davuluri et al., 2008). However, less frequently discussed is the notion that the different proteins resulting from the alternative promoter usage also have important physiological functions.

Nuclear Receptors

Nuclear receptors (NR) are ligand-sensitive transcription factors that regulate a wide array of biological processes including development, metabolism, and circadian rhythms. It is not surprising then that they also play a role in many diseases including obesity, diabetes, cancer, atherosclerosis, and inflammation. There are a total of 48 nuclear receptors encoded in the human genome. With the exception of NR0B1 (DAX1) and NR0B2 (SHP), which lack a DNA-binding domain, all NRs share a common protein structure: a highly variable N-terminal domain with transactivation function (A/B domain; activation function 1, AF-1), a highly conserved DNA-binding domain (C domain), a variable "hinge" region (D domain), a large highly conserved ligand-binding domain (E domain; LBD; activation function 2, AF-2) which can also play a role in dimerization of NRs, and in some cases a highly variable C-terminal tail (F domain). NRs may interact with many variations of a consensus DNA response element (AGGCTA or AGAACA) throughout the genome as monomers, homodimers, or heterodimers in a wide array of conformations, including direct or inverted repeats with anywhere from 0 to 6 nucleotides (nt) as spacers between each element. It is from these elements that NRs recruit other transcriptional co-regulators, which interact with the general transcription machinery to either repress or activate their target genes.

The A/B domain of NRs is the most variable domain within the superfamily, with no conservation in either length or structure. The steroid receptors have the longest A/B domains, as large as 602 amino acids for mineralocorticoid receptor (Lavery and

McEwan, 2005), while non-steroid receptors tend to have much shorter domains, even as small 24 amino acids in the case of the vitamin D receptor (Campbell et al., 2010). N-terminal domains of NRs are thought to be highly flexible and unstructured in the absence of binding partners (Chandra et al., 2008; Wärnmark et al., 2001). This flexibility has led to difficulties in deciphering the structure of A/B domains, which remain poorly understood.

Hepatocyte nuclear factor 4 alpha (HNF4α) is a liver-enriched transcription factor and a member of the NR superfamily (Sladek et al., 1990). HNF4α is expressed in the liver, kidney, colon, pancreas, stomach, and intestine. It is highly expressed in the liver where it is best known as a master regulator of liver-specific gene expression (Bolotin et al., 2010; Odom et al., 2004) and is essential for adult and fetal liver function. HNF4α knockout mice are embryonic lethal and adult liver HNF4α knockouts die within six weeks with a fatty-liver phenotype (Chen et al., 1994; Hayhurst et al., 2001). Results from transcriptional regulatory networks derived from ChIP promoter microarrays for HNF factors indicate that dysregulation of HNF4α may contribute to the development of type 2 diabetes (Odom et al., 2004), consistent with inherited mutations in the HNF4A gene in maturity onset diabetes of the young 1 (MODY1) (Yamagata et al., 1996).

The human HNF4A and mouse Hnf4a genes are highly conserved. Both are regulated by alternative promoters, the proximal P1 and distal P2 promoters. In the adult liver the P1 promoter is the only active promoter, while during fetal development both P1 and P2 promoters are active (Torres-Padilla et al., 2001). P1-HNF4α is expressed in the liver, small intestine, colon, and kidney while P2-HNF4α is expressed in the fetal liver, pancreas, stomach, small intestine, and colon. In liver, colon, and stomach cancers altered expression patterns of P1- and P2-HNF4α are typically found, suggesting that altered promoter usage may be important in cancer development (Tanaka et al., 2006). While it has been shown that P1-HNF4α acts as a tumor suppressor in the liver (Hatziapostolou et al., 2011; Walesky and Apte, 2015), the specific roles of the isoforms remain unclear.

The primary isoforms derived from the P1 promoter are HNF4α1/α2 while the primary isoforms derived from the P2 promoter are HNF4α7/α8. These isoforms share ≥90% homology with each other and have identical DNA-binding and ligand-binding domains. The only difference between P1 and P2 isoforms are alternative first exons that result in an altered A/B domain and the loss of the AF-1 domain for the P2 isoforms (Briançon and Weiss, 2006; Torres-Padilla et al., 2002). It should be noted that phosphorylation of HNF4α by Src kinase preferentially targets P1-HNF4α leading to protein degradation, without affecting P2-HNF4α (Chellappa et al., 2012). P2-HNF4α is not normally expressed in the adult liver, thus, to study the role of P1- and P2-HNF4α in the mouse liver we used genetically engineered (exon swap) mice that express exclusively the P1- or the P2-HNF4α isoforms (Briançon and Weiss, 2006).

Cytochrome P450s

Since the NR superfamily is a family of ligand-sensitive transcription factors, ligand synthesis and degradation play a crucial role in tissue-specific hormonal signaling and gene expression. Ligands must be synthesized and delivered throughout the body and the degradation of ligands helps to limit both the duration and the intensity of the NR-ligand response. The enzymes that regulate these processes are the cytochrome P450s (Cyp) enzymes that play a key role in the oxidative metabolism of cholesterol, steroids, bile acids, and fatty acids (Furge and Guengerich, 2006; Nebert and Russell, 2002). It has been shown that these enzymes are regulated by NRs pregnane X receptor (PXR, Nr1i2) and constitutive androstane receptor (CAR, Nr1i3) (di Masi et al., 2009; Tolson and Wang, 2010; Willson and Kliewer, 2002). It has also been shown that HNF4α is a critical factor for in vivo transcriptional activation of one of the most abundant drug metabolism enzymes, CYP3A4, and plays a critical role in liver-specific Cyp gene expression (Hwang-Verslues and Sladek, 2010; Tirona et al., 2003).

Circadian Rhythms

Circadian rhythms have been identified in essentially all living organisms -- including animals, plants, fungi and bacteria -- as a mechanism to optimize energy acquisition and storage. In higher animals such as rodents and primates, the suprachiasmatic nucleus (SCN) functions to signal peripheral clocks to the light cycle. This includes functions related to energy homeostasis and involves tissues that regulate glucose and lipid metabolism, such as the liver.

Nuclear receptors are a unique family of ligand-sensitive TFs that lie at the intersection of metabolic and circadian pathways. Core circadian clock components, which include NRs, work in an incredibly well coordinated transcriptional feedback loop that regulates mRNA expression, protein stability, chromatin states, and metabolite production and utilization (Eckel-Mahan and Sassone-Corsi, 2013). The circadian feedback loop is a highly conserved process controlling oscillating gene expression profiles every 24 hours. The core components CLOCK and BMAL1 heterodimerize and bind E-box motifs in the promoters of target genes, and drive the expression of period (Per1/2/3) and cryptochrome (Cry1/2) genes, which in turn work together to inhibit CLOCK:BMAL1 mediated expression.

Metabolic homeostasis can be defined as the balance of energy intake being equal to energy expenditure. In order for energy homeostasis to occur, rhythms in energy intake must coincide with rhythms in gene expression for metabolic homeostasis. Tissues throughout the body control their metabolic demands by using signaling molecules such as insulin, glucagon, leptin and ghrelin. The fact that many of these hormones oscillate throughout the day, or are dependent on energy intake, suggests a high level of interaction between metabolic and circadian processes.

More than half of all NRs display rhythmic patterns of expression in multiple metabolic tissues including FXR, LXR, HNF4α, PPARα, PPARɣ to name a few (Yang et al., 2006). Some NRs, such as REV-ERBα, RORα, and PPARα, have been shown to directly regulate the expression of BMAL1, while RORα and PPARα have been shown to interact directly with PER2 to modulate its activity (Schmutz et al., 2010). These findings suggest that fatty acids, sterols, and other hormones may be able to communicate information about nutrient and energy status to the clock via their cognate NRs.

In Chapter 2, we investigate the roles of HNF4α isoforms as a result of alternative promoter usage. However, HNF4A is not the only NR gene with alternative promoters: PPARG (PPARɣ), NR3C1 (glucocorticoid receptor, GR), NR1H4 (FXR), NR1I2 (PXR) and VDR (vitamin D receptor) all have alternative promoters that produce distinct isoforms with functional differences (Crofts et al., 1998; Huber et al., 2002; Kurose et al., 2005; Lee and Ge, 2014; Russcher et al., 2007). In the case of FXR, alternative promoter usage limits expression of FXRα1 to the liver and FXRα2 in the kidney and intestine (Huber et al., 2002; Zhang et al., 2003). PPARɣ alternative promoter usage also elicits a tissue-specific isoform response in adipose versus other tissues. During adipogenesis of 3T3-L1 cells PPARɣ1 is expressed early on, followed by PPARɣ2, while both are expressed at similar levels later in differentiation (Cho et al., 2009). However, during adipogenesis of brown pre-adipocytes, PPARɣ1 is expressed early on and remains the dominant isoform through differentiation (Jitrapakdee et al., 2005). Relatively little is known about the exact functions of these alternative 5’ isoforms of NRs other than the fact that these alternative promoters provide an excellent means to control tissue-specific gene expression of these isoforms.

Similarly, the alternative promoter isoforms of HNF4α control tissue-specific expression of the NR. It has been shown that both P1-HNF4α and P2-HNF4α are expressed in fetal liver, intestine and colon, while P1-HNF4α is exclusively expressed in the kidney and adult liver and P2-HNF4α in the pancreas (Harries et al., 2008). While others have shown a potential role for P1- and P2-HNF4α in cancer (Tanaka et al., 2006; Walesky and Apte, 2015), a mechanism for how these two isoforms control transcriptional activation has yet to be established. The focus of Chapter 2 is on a series of genome-scale experiments using exon-swap mice to identify the unique functions of these two isoforms in the adult liver. ChIP-seq results show that P1- and P2-HNF4α isoforms have nearly identical DNA binding affinities, further confirmed by the use of PBMs. Transcriptomics and protein-protein interaction data suggest unique roles of HNF4α isoforms in circadian rhythms and the regulation of cytochrome P450s involved in lipid metabolism.

In Chapter 3, we utilize PBMs to investigate the impact of common genetic variation on the DNA binding of a class of transcription factors known as nuclear receptors (NR). By probing DNA binding affinity with PBMs on genomic DNA flanking genetic variants found in the promoters of disease associated genes, NR ChIP-seq peaks, and significant eQTLs identified in the liver from the GTEx project, we identify thousands of in vitro affinity altering SNPs (aaSNPs) with the capability of disrupting NR binding to genomic DNA.

With a similar approach in Chapter 4, we also expand on the power of the PBM data to analyze 100,000 genetic variants that fall within PPARγ ChIP peaks, as well as 1,000 GWAS identified variants associated with Alzheimer's, a neurodegenerative disease highly commonly treated with PPARγ agonists, and identify a genomic locus with an enrichment of PPARγ aaSNPs around genes known to play a key role in the development of the disease.

References

Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. 16, 197–212.

Bolotin, E., Liao, H., Ta, T.C., Yang, C., Hwang-Verslues, W., Evans, J.R., Jiang, T., and Sladek, F.M. (2010). Integrated approach for the identification of human hepatocyte nuclear factor 4alpha target genes using protein binding microarrays. Hepatology 51, 642–653.

Briançon, N., and Weiss, M.C. (2006). In vivo role of the HNF4α AF-1 activation domain revealed by exon swapping. EMBO J. 25, 1253–1262.

Campbell, F.C., Xu, H., El-Tanani, M., Crowe, P., and Bingham, V. (2010). The yin and yang of vitamin D receptor (VDR) signaling in neoplastic progression: operational networks and tissue-specific growth control. Biochem. Pharmacol. 79, 1–9.

Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T.P., and Rastinejad, F. (2008). Structure of the intact PPAR-gamma-RXR- nuclear receptor complex on DNA. Nature 456, 350–356.

Chellappa, K., Jankova, L., Schnabl, J.M., Pan, S., Brelivet, Y., Fung, C.L.-S., Chan, C., Dent, O.F., Clarke, S.J., Robertson, G.R., et al. (2012). Src tyrosine kinase phosphorylation of nuclear receptor HNF4α correlates with isoform-specific loss of HNF4α in human colon cancer. Proc. Natl. Acad. Sci. U. S. A. 109, 2302–2307.

Chen, W.S., Manova, K., Weinstein, D.C., Duncan, S.A., Plump, A.S., Prezioso, V.R., Bachvarova, R.F., and Darnell, J.E., Jr (1994). Disruption of the HNF-4 gene, expressed in visceral endoderm, leads to cell death in embryonic ectoderm and impaired gastrulation of mouse embryos. Genes Dev. 8, 2466–2477.

Cho, Y.-W., Hong, S., Jin, Q., Wang, L., Lee, J.-E., Gavrilova, O., and Ge, K. (2009). Histone methylation regulator PTIP is required for PPARgamma and C/EBPalpha expression and adipogenesis. Cell Metab. 10, 27–39.

Crofts, L.A., Hancock, M.S., Morrison, N.A., and Eisman, J.A. (1998). Multiple promoters direct the tissue-specific expression of novel N-terminal variant human vitamin D receptor gene transcripts. Proc. Natl. Acad. Sci. U. S. A. 95, 10529–10534.

Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T.H.-M. (2008). The functional consequences of alternative promoter use in mammalian genomes. Trends Genet. 24, 167–177.

Eckel-Mahan, K., and Sassone-Corsi, P. (2013). Metabolism and the circadian clock converge. Physiol. Rev. 93, 107–135.

Furge, L.L., and Guengerich, F.P. (2006). Cytochrome P450 enzymes in drug metabolism and chemical toxicology: An introduction. Biochem. Mol. Biol. Educ. 34, 66–74.

Harries, L.W., Locke, J.M., Shields, B., Hanley, N.A., Hanley, K.P., Steele, A., Njølstad, P.R., Ellard, S., and Hattersley, A.T. (2008). The diabetic phenotype in HNF4A mutation carriers is moderated by the expression of HNF4A isoforms from the P1 promoter during fetal development. Diabetes 57, 1745–1752.

Hatziapostolou, M., Polytarchou, C., Aggelidou, E., Drakaki, A., Poultsides, G.A., Jaeger, S.A., Ogata, H., Karin, M., Struhl, K., Hadzopoulou-Cladaras, M., et al. (2011). An HNF4α-miRNA inflammatory feedback circuit regulates hepatocellular oncogenesis. Cell 147, 1233–1247.

Hayhurst, G.P., Lee, Y.H., Lambert, G., Ward, J.M., and Gonzalez, F.J. (2001). Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. Mol. Cell. Biol. 21, 1393–1403.

Huber, R.M., Murphy, K., Miao, B., Link, J.R., Cunningham, M.R., Rupar, M.J., Gunyuzlu, P.L., Haws, T.F., Kassam, A., Powell, F., et al. (2002). Generation of multiple farnesoid-X-receptor isoforms through the use of alternative promoters. Gene 290, 35–43.

Hwang-Verslues, W.W., and Sladek, F.M. (2010). HNF4α--role in drug metabolism and potential drug target? Curr. Opin. Pharmacol. 10, 698–705.

Jitrapakdee, S., Slawik, M., Medina-Gomez, G., Campbell, M., Wallace, J.C., Sethi, J.K., O'Rahilly, S., and Vidal-Puig, A.J. (2005). The Peroxisome Proliferator-activated Receptor-γ Regulates Murine Pyruvate Carboxylase Gene Expression in Vivo and in Vitro. J. Biol. Chem. 280, 27466–27476.

Kurose, K., Koyano, S., Ikeda, S., Tohkin, M., Hasegawa, R., and Sawada, J.-I. (2005). 5' diversity of human hepatic PXR (NR1I2) transcripts and identification of the major transcription initiation site. Mol. Cell. Biochem. 273, 79–85.

Lavery, D.N., and McEwan, I.J. (2005). Structure and function of steroid receptor AF1 transactivation domains: induction of active conformations. Biochem. J 391, 449–464.

Lee, J.-E., and Ge, K. (2014). Transcriptional and epigenetic regulation of PPARγ expression during adipogenesis. Cell Biosci. 4, 29.

di Masi, A., De Marinis, E., Ascenzi, P., and Marino, M. (2009). Nuclear receptors CAR and PXR: Molecular, functional, and biomedical aspects. Mol. Aspects Med. 30, 297–343.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195.

Nebert, D.W., and Russell, D.W. (2002). Clinical importance of the cytochromes P450. Lancet 360, 1155–1162.

Odom, D.T., Zizlsperger, N., Benjamin Gordon, D., Bell, 1. George W., Rinaldi, N.J., Murray, H.L., Volkert, 1. Tom L., Schreiber, J., Alexander Rolfe, P., Gifford, D.K., et al. (2004). Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. Science 303.

Russcher, H., Dalm, V.A.S.H., de Jong, F.H., Brinkmann, A.O., Hofland, L.J., Lamberts, S.W.J., and Koper, J.W. (2007). Associations between promoter usage and alternative splicing of the glucocorticoid receptor gene. J. Mol. Endocrinol. 38, 91–98.

Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 6, e107.

Schmutz, I., Ripperger, J.A., Baeriswyl-Aebischer, S., and Albrecht, U. (2010). The mammalian clock component PERIOD2 coordinates circadian output by interaction with nuclear receptors. Genes Dev. 24, 345–357.

Singer, G.A.C., Wu, J., Yan, P., Plass, C., Huang, T.H.M., and Davuluri, R.V. (2008). Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array. BMC Genomics 9, 349.

Sladek, F.M., Zhong, W.M., Lai, E., and Darnell, J.E., Jr (1990). Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. Genes Dev. 4, 2353–2365.

Tanaka, T., Jiang, S., Hotta, H., Takano, K., Iwanari, H., Sumi, K., Daigo, K., Ohashi, R., Sugai, M., Ikegame, C., et al. (2006). Dysregulated expression of P1 and P2 promoter-driven hepatocyte nuclear factor-4alpha in the pathogenesis of human cancer. J. Pathol. 208, 662–672.

Tirona, R.G., Lee, W., Leake, B.F., Lan, L.-B., Cline, C.B., Lamba, V., Parviz, F., Duncan, S.A., Inoue, Y., Gonzalez, F.J., et al. (2003). The orphan nuclear receptor HNF4alpha determines PXR- and CAR-mediated xenobiotic induction of CYP3A4. Nat. Med. 9, 220–224.

Tolson, A.H., and Wang, H. (2010). Regulation of drug-metabolizing enzymes by xenobiotic receptors: PXR and CAR. Adv. Drug Deliv. Rev. 62, 1238–1249.

Torres-Padilla, M.E., Fougere-Deschatrette, C., and Weiss, M.C. (2001). Expression of HNF4a isoforms in mouse liver development is regulated by sequential promoter usage and constitutive 3 end splicing. Mech. Dev. 109, 183–193.

Torres-Padilla, M.E., Sladek, F.M., and Weiss, M.C. (2002). Developmentally regulated N-terminal variants of the nuclear receptor hepatocyte nuclear factor 4alpha mediate multiple interactions through coactivator and corepressor-histone deacetylase complexes. J. Biol. Chem. 277, 44677–44687.

Walesky, C., and Apte, U. (2015). Role of hepatocyte nuclear factor 4α (HNF4α) in cell proliferation and cancer. Gene Expr. 16, 101–108.

Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. 30, 1095–1106.

Wärnmark, A., Wikström, A., Wright, A.P.H., Gustafsson, J.-Å., and Härd, T. (2001). The N-terminal Regions of Estrogen Receptor α and β Are Unstructured in Vitro and Show Different TBP Binding Properties. J. Biol. Chem. 276, 45939–45944.

Willson, T.M., and Kliewer, S.A. (2002). PXR, CAR and drug metabolism. Nat. Rev. Drug Discov. 1, 259–266.

Yamagata, K., Furuta, H., Oda, N., Kaisaki, P.J., Menzel, S., Cox, N.J., Fajans, S.S., Signorini, S., Stoffel, M., and Bell, G.I. (1996). Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). Nature 384, 458–460.

Yang, X., Downes, M., Yu, R.T., Bookout, A.L., He, W., Straume, M., Mangelsdorf, D.J., and Evans, R.M. (2006). Nuclear receptor expression links the circadian clock to metabolism. Cell 126, 801–810.

Zhang, Y., Kast-Woelbern, H.R., and Edwards, P.A. (2003). Natural structural variants of the nuclear receptor farnesoid X receptor affect transcriptional activation. J. Biol. Chem. 278, 104–110.

Chapter 2

The N-terminal Domain of HNF4α Regulates Xenobiotic and Lipid Metabolism in the Liver

Contributions from others:

Dr. Linh Vuong: Assisted with ChIP-seq and RNA-seq library preparation,
Dr. Poonam Deol: Performed GTT assays and metabolite analysis,
Dr. Nina Titova: Extract preparation and application to PBMs
Jane Evans, SRA: Immunoblots

Abstract

The nuclear receptor (NR) hepatocyte nuclear factor 4α (HNF4α), a master regulator of liver-specific gene expression, is regulated by two promoters (P1 and P2) resulting in proteins with different N-terminal A/B domains. P1-HNF4α is expressed in fetal and normal adult liver while P2-HNF4α is expressed only in the fetal liver and in liver cancer. We compared wildtype mice, which express only HNF4α1 (P1) in the adult liver, to exon-swap mice that express only HNF4α7 (P2) for global changes in gene expression (RNA-seq), chromatin binding (ChIP-seq), and unique protein interactions (RIME). The results show that P1- and P2-HNF4α isoforms differentially regulate hundreds of transcripts in the adult liver, including the NR CAR (*Nr1i3*). They also exhibit altered metabolic pathways, especially cytochrome P450 (*Cyp*) genes. Protein binding microarrays (PBM), ChIP-seq and RIME show that while P1- and P2-HNF4α bind canonical HNF4α binding motifs with similar specificity, they may be recruited differentially to non-HNF4α binding sites by unique protein interactions. Wildtype and exon swap mice also show differential responses to fasting. All told, the results show that changes in just 16-30 amino acids in the AF-1 region of an NR can have profound effects on gene expression.

Introduction

Hepatocyte nuclear factor 4 alpha (HNF4α) is a liver-enriched transcription factor (TF) and a highly conserved member of the nuclear receptor (NR) superfamily (Sladek et al., 1990). HNF4α is best known as a master regulator of liver-specific gene expression and mutated in Maturity Onset Diabetes of the Young 1 (MODY1) (Fajans et al., 2001; Yamagata et al., 1996). HNF4α is essential for fetal liver function (Battle et al., 2006) while *Hnf4a* adult liver knockout (KO) mice die within six weeks with a fatty liver phenotype (Hayhurst et al., 2001).

The human *HNF4A* and mouse *Hnf4a* genes are highly conserved and regulated by proximal P1 and distal P2 promoters. P1 drives the expression of transcripts containing exon 1A while P2 transcripts contain exon 1D, resulting in a loss of the N-terminal activation function 1 (AF-1). In the adult liver P1 is presumed to be the only active promoter, while during fetal liver development both P1 and P2 are active (Briançon et al., 2004; Torres-Padilla et al., 2001). The first P2-HNF4α transcript cloned, HNF4α7, was from the embryonal carcinoma cell line F9 (Nakhei et al., 1998), suggesting that it might play a role in cancer as well as fetal development. Indeed, P1-HNF4α is down regulated in liver cancer and acts as a tumor suppressor (Hatziapostolou et al., 2011; Ning et al., 2010; Tanaka et al., 2006; Walesky and Apte, 2015), while overexpression of P2-HNF4α is linked to poor prognosis in hepatocellular carcinoma (HCC) (Cai et al., 2017).

To address the physiological role of P2-HNF4α, we employed exon swap mice, which substitute exon 1A with exon 1D in the P1 promoter and demonstrate a subtle, albeit undefined role for the AF-1 domain *in vivo* (Briançon and Weiss, 2006). We compared these α7HMZ adult mice (express only P2-HNF4α) to wildtype (WT) mice (express P1-HNF4α in adult liver) using RNA-seq, ChIP-seq, rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME), protein binding microarrays (PBMs) and metabolomics. An orchestrated, altered hepatic transcriptome in P2-HNF4α livers reveals notable differences in cytochrome P450 transcripts and subtle differences in the circadian clock. The distinct P2-HNF4α transcriptome appears to be due to altered protein-protein interactions, as well as altered chromatin binding but not differences in innate DNA binding specificity, as determined by *in vitro* DNA binding reactions to ~44,000 unique sequences. Interestingly, the P2-HNF4α metabolome is characterized by altered fatty acid metabolism and a fatty liver. Our results suggest that expression of P2-HNF4α in the liver is an evolutionarily conserved mechanism to survive extreme metabolic challenges.

Materials & Methods

Animals

WT and α7HMZ mice were maintained in isolator cages under 12-h light/dark cycles at ~21^{o}C on bedding (Andersons bed OCOB Lab ⅛ 1.25CF) from Newco (Rancho Cucamonga, CA) and fed a standard lab chow (LabDiet, #5001, St. Louis, MO). They were bred and maintained in a specific pathogen free (SPF) vivarium, and all experiments were performed in an SPF vivarium. Young adult males were used for all experiments except the newborn liver analysis, for which gender was not identified.

The transgenic mice on a mixed 129/Sv plus C57BL/6 background carrying exon 1A or exon 1D in both the P1 and P2 promoter (α1HMZ and α7HMZ, respectively) have been described previously (Briançon and Weiss, 2006). Both lines were maintained as heterozygotes; wildtype (WT) and homozygous (α7HMZ) were mated for a single generation to generate mice for the experiments. The WT and α7HMZ mice in the mixed background were used for all RNA-seq, ChIP-seq, and RIME experiments. Mice of the same genotype were housed 3-5 per cage and randomly selected to treatment groups at the beginning of the experiment. Mice were euthanized by CO_2 asphyxiation and tissues harvested at the designated experiment time points. All mice used were adult males, aged 16 to 20 weeks, unless otherwise noted. Time points were 10:30 (ZT 3.5), 13:30 (ZT 6.5) and 20:30 (ZT 13.5) (lights on at 7:00 and off at 19:00).

Care and treatment of the animals were in strict accordance with guideline from the University of California Riverside Institutional Animal Care and Use Committee (Protocol# A200140014).

Immunoblot (IB) analysis

Immunoblots (IBs) in Figure 1 were carried out as previously described (Jiang et al., 1995). Proteins from nuclear extracts (NE) and whole cell extracts (WCE) were separated by 10% SDS-PAGE and then transferred to PVDF (Immobilon). Even loading was verified by Coomassie stain of the blot. Primary antibodies (Ab) were mouse monoclonal anti-HNF4α P1/P2 (R&D Systems #PP-H1415-00) which recognizes the Cterminus of both P1- and P2-HNF4α isoforms, and mouse monoclonal anti-HNF4α P1 (R&D Systems # PP-K9218-00) which recognizes the N-terminus of P1-HNF4α. Both were used at 1:10,000 overnight. Secondary antibodies were horseradish peroxidase (HRP)-conjugated goat anti-mouse (GαM-HRP) Abs from Jackson ImmunoResearch Laboratories. The procedure for NE from COS-7 and liver are described below. WCE of liver in Figure 1 were prepared using RIPA buffer (see RIME below).

Preparation of nuclear extracts (NE) for Immunoblot (IB) and Protein Binding Microarrays (PBM) analysis

Following buffers and inhibitors were used:

TE: 10 mM Tris-HCl, 1 mM EDTA pH 8.0

2X HBS: 274 mM NaCl, 10 mM KCL, 1.4 mM Na_2HPO_4, 15 mM D-glucose, 42 mM HEPES (free acid), pH 7

1X H Buffer: 100 mM HEPES, pH 7.8, 250 mM KCl, 1.5 mM spermine, 5 mM spermidine, 10 mM EGTA, 10 mM EDTA

Buffer A: 1X H Buffer, 0.32 M Sucrose

Low salt buffer: 1X H buffer, 20% glycerol

High salt buffer: 1X H buffer, 20% glycerol, 1 M KCL (for COS-7 cells) or 0.5 M KCL (for liver).

Inhibitors and DTT: Protease inhibitors (Sigma, #8340), Phosphatase inhibitor cocktail I (Sigma, #P-2850), Phosphatase inhibitor cocktail II (Sigma, #P-5726) and 200 mM phenylmethylsulfonyl fluoride (PMSF) were added at the dilution 1:200 (for cell NE) and 1:100 (for liver NE) to each buffer solution before each use. Dithiothreitol (DTT) was added to 1 mM final.

Nuclear extracts (NE) were prepared from COS-7 cells transiently transfected via $CaPO_4$ with HNF4α expression vectors for human HNF4α2 (NM_00457) and HNF4α8 (NM_175914) as previously described (Jiang et al., 1995) with some modifications. Cells (3.5×10^6) were plated in 150-mm plates and incubated in 15 ml of DMEM supplemented with 10% BCS at 37°C for 24 to 48 hours. Twenty-five µg of plasmid DNA, HNF4α2 (NM_00457) or HNF4α8 (NM_175914) in pcDNA3.1, was mixed with 450 µl TE and 500 µL 2X HBS buffer; 50 µL 2.5 M $CaCl_2$ was added to the mixture, and 25 min later the mixture was added to the cells. After approximately 10 h of incubation, the cells were washed 1x with PBS and then shocked with 3 mL 15% glycerol in 1X HBS buffer for 3 min 15 s, and then washed 2X with PBS, followed by DMEM plus 10% BCS. Then, 24 to 32 h later, the cells were harvested and NE were prepared. Cells were washed twice with cold PBS, then once with 1 ml of 0.25X Buffer H; 0.75 ml of 0.25X Buffer H was subsequently added to each 150-mm plate and incubated on ice for 5 to 20 min. Cells were scraped and resuspended in equal volume of 2X Buffer H plus 20% glycerol. After centrifugation (10 min at 2,500 rpm) the supernatant was discarded and the nuclear pellet resuspended in an equal volume of Low Salt Buffer; 0.72X volumes of High Salt Buffer was used to resuspend the nuclei followed by nutation at 4°C for 1h 10 min. The soluble NE was separated from the chromatin pellet by centrifugation (25 min at 12,000 rpm). Samples were snap-frozen and subsequently used for IBs and PBMs.

Liver NE from WT and α7HMZ mice were prepared as previously described (Yuan et al., 2009) by motorized homogenization of frozen or fresh liver in Buffer A plus 0.3% Triton X-100, protease and phosphatase inhibitors. The homogenate was filtered using 100-µm cell strainers (Fisher #08-771-19) before passing through the dounce homogenizer (Fisher #06-435B) in Buffer A plus 0.3% Triton X-100 followed by nuclei separation via centrifugation (10 min at 3,300 rpm) and multiple washes of nuclei (1X wash in Buffer A plus 0.3% Triton X-100, followed by 2X wash in Low Salt Buffer). Each wash followed by centrifugation (10 min at 2,000 rpm). After washes, nuclear pellets were resuspended in 1X volume of Low Salt Buffer, 2X volume of High Salt Buffer was added, and extraction was performed as described above for cell NE. All incubations, separations and washes were at 4°C.

Expression profiling (RNA-seq) and Analysis

Next generation sequencing of RNA (RNA-seq) was carried out as previously described (Vuong et al., 2015). WT and α7HMZ male mice were sacrificed (n=3, aged 16-18 weeks) at each time point: 10:30, 13:30, 20:30 PM (ZT 3.5, ZT 6.5, and ZT 13.5, respectively). The three mice were harvested in succession within a 30-min time frame. Two ~25 mg pieces from each liver were immediately frozen in liquid nitrogen and stored at -4^{O}C. The miRNeasy Mini Kit (Qiagen, #74104) was used to extract and purify total RNA; 4 µg of each sample was used to prepare a poly(A)+ RNA library using TruSeq RNA Sample Prep v2 Kit (Illumina, Cat# RS-122-2001). Libraries submitted for 75-bp single-end sequencing with Illumina NextSeq 500 at the UCR IIGB Genomics Core. A total of 24 libraries (3 fed time points, 1 fasted time point, 2 genotypes each, 3 replicates) were multiplexed and sequenced in two separate runs, each of which yielded ~600 M reads, averaging ~50 M reads per sample.

Reads were aligned to the mouse reference genome, mm10, with Illumina's iGenome genes.gtf file using TopHat v2.1.1 using default parameters with the exception of allowing only 1 unique alignment for a given read. Raw read counts were calculated at the gene level for each sample using HTSeq v0.6.1. Library normalization was performed with EDASeq; within-lane normalization on GC content was performed with the LOESS method and between-lane normalization was performed with non-linear full quantile method. Normalization factors from EDASeq were used for differential expression analysis with DESeq2. Normalized read counts, FPKM (fragments per kilobase per million), and rlog (regularized log transformation) results were generated for downstream analysis. Pairwise contrasts were generated for all relevant comparisons. Sample distance matrix were generated using rlog transformed values from DESeq2.

Chromatin Immunoprecipitation Sequencing (ChIP-seq) and Analysis

A freshly minced chunk from the large lobe of the liver (approximately 100-200 mg) of a 16 to 20-week old male mouse (WT and α7HMZ, 10:30) was fixed in 1%

formaldehyde (in ChIP Buffer: 1X PBS plus 1 mM PMSF, 1 mM DTT, 2 μg/mL leupeptin, 2 μg/mL aprotinin) for 15 min at room temperature (RT). The crosslink reaction was stopped with 0.125 M glycine for 5 min at RT and centrifuged (10 min at ~2,000 rpm) at 4°C. All subsequent steps were performed at 4°C. Fixed tissue was further processed using motorized and glass dounce homogenizers (Fisher #06-435B) in cold ChIP Buffer. The homogenate was filtered using 100-μm cell strainers (Fisher #08-771-19) before passing through the dounce homogenizer. Isolated liver cells were processed as previously described (Vuong et al., 2015). Briefly, cells were swelled in 1.0 mL cold Hypotonic Buffer (10 mM HEPES-KOH pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$) plus 1 mM PMSF and 1 mM DTT) for 10 min. The nuclei were collected by centrifugation and resuspended in 1.0 mL cold Nuclei Lysis Buffer (1% SDS, 50 mM Tris-HCl pH 8.0, 10 mM EDTA) plus 1 mM PMSF, 1 mM DTT, 2 μg/mL leupeptin and 2 μg/mL aprotinin. The samples were sonicated using a Sonic Dismembrator Model 500 (Fisher Scientific) to obtain DNA fragments of about 200-500 bps, diluted 1:1 with Immunoprecipitation (IP) dilution buffer (0.01% SDS, 20 mM Tris-HCl pH 8.0, 1.1% Triton X-100, 167 mM NaCl, 1.2 mM EDTA), and pre-cleared with 20 μL of packed Protein G Agarose (Pierce) beads (1:1 slurry in IP dilution buffer) that were pre-blocked with 100 μg/μL BSA for 30 min. The IP was performed with 4.2 μg of affinity-purified anti-HNF4α (α-445) (Sladek et al., 1990) or rabbit IgG control (Santa Cruz, cat#sc-2027). Thirty to forty microliters of packed protein G beads (1:1 slurry) were added to the IP sample and incubated overnight. All IPs were washed with three sequential buffers for 5 min each at RT: TSE I (0.1%

SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 150 mM NaCl), TSE II (as TSE I but with 500 mM NaCl) and TSE III (0.25 mM LiCl, 1% NP-40, 1% Deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.0). At the final wash, the IP sample was washed twice with 1X TE for 5 min at RT. The precipitated material was then eluted with IP elution buffer (1% SDS and 0.1 M $NaHCO_3$) twice. For the first elution, the sample was eluted with 150 µL buffer, incubated at RT for 20 min, centrifuged at maximum speed for 1 min and the supernatant was transferred to a new tube. The second elution was the same as the first but with an additional 1 min boiling in a 100^{o}C heat block preceding the 20-min incubation. The material from the first and second elutions were combined, incubated at 65^{o}C for 4 to 5 h to reverse the crosslinks. DNA precipitation was performed with 1 mL of 100% ethanol overnight at -20^{o}C. Protein and RNA digestions were performed for 1 h at 55^{o}C with 11 µL of 10X Proteinase K buffer (100 mM Tris-HCl pH 8.0, 50 mM EDTA, 500 mM NaCl) and 1 µL of 19 mg/mL Proteinase K solution (in 10 mM Tris-HCl pH 8.0, 1 mM $CaCl_2$ plus 30% glycerol) and 25 min at RT with 1 µL of 10 µg/uL RNaseA in ddH_2O, respectively. DNA material was purified with GeneJET PCR Purification Kit (Thermo Scientific, #K0701). Qubit fluorometer at the University of California Riverside (UCR) Institute for Integrative Gene Biology (IIGB) Genomics Core was used to measure DNA concentration; 2-22 µg of ChIP'd material was used to generate a library using the BIOO Scientific ChIPseq DNA Library Kit. Libraries were submitted for 50-bp single end sequencing by Illumina HiSEQ 2500 at the IIGB core facility.

Reads were aligned to the mouse reference genome, version mm10, with Bowtie2 using the default parameters. Peaks were called with MACS2 using default parameters for individual samples, as well as a pooled peak dataset using the SPMR (signal per million reads) parameter. Aligned reads and MACS2 peak-sets were analyzed with DiffBind to identify common and uniquely bound regions of the genome. Livers from three mice were used for each genotype. After PCA analysis, one α7HMZ replicate was identified visually as an outlier from the other two replicates so a single WT and α7HMZ replicate were omitted from downstream analysis, leaving two replicates per genotype. DiffBind analysis was performed with default parameters using DESeq2 for analysis and library size calculated as total aligned reads. ChIP peaks were excluded from analysis if MACS2 results showed -log10(p-value) ≤ 10 to produce a filter removing peaks below six-fold enrichment over background, and unique peak IDs were manually annotated to each resulting peak. Manually curated peak lists were generated by filtering all results on peaks with "concentration" ≥ 5.5. Concentration defined by DiffBind as the "mean (log) reads across all samples" in contrast.

Support Vector Machine (SVM) Predictions and Motif Generation

The kernel-based SVM was trained as previously described using results from independent HNF4α PBM experiments (Bolotin et al., 2010). All possible 13-mers in both orientations from each uniquely bound ChIP peak were submitted to the HNF4α PBM SVM for score predictions using Kernlab package in R. Each ChIP peak was then annotated with the score and sequence of the single highest predicted motif from all possible k-mers and categorized into four different bins. All sequences within each bin were submitted to seqLogo (Bembom 2017) to generate a position weight matrices (PWM) representing the strongest HNF4α binding motif within the category.

For each SVM-score category, sequences from a 200-bp window around the peak center for each ChIP peak were submitted to MEME-ChIP (Machanick and Bailey, 2011) for *de novo* motif analysis with default parameters, with the exception of number of motifs to identify (6), max word size (24), and the transcription factor binding site (TFBS) database utilized was HOCOMOCO v10.

Protein Binding Microarrays (PBM)

Protein binding microarrays (PBMs) were carried out as previously described (Bolotin et al., 2010). A custom-designed array was ordered from Agilent (SurePrint G3 Custom GE 4x180k), which contained oligonucleotides ~60 nucleotides (nt) in length, corresponding to the following sequences: sequences within 100 bp of the center of published HNF4α ChIP-seq peaks from proliferative Caco-2 cells (Verzi et al., 2010) were taken in 30-nt windows moving 5 nt at each step; 17,250 permutations of canonical HNF4α DR1 motifs (5'- AGGTCAAAGGTCA -3'); 500 permutations of DR2 motifs with variable spacer (5'- AGGTCNNNNGGTCA -3'); 900 random control 13-mer DNA sequences. A total of ~45,000 test sequences were spotted in quadruplicate on the slide as single-stranded DNA. The DNA was extended and made double-stranded on the slide using a primer to a common linker sequence (5’-TCGACCGATACTCTAATCTCCCTAGGC-3’), dNTPs (GE Healthcare) and Thermo Sequenase (Affymetrix, Cat# 78500). Extension and all incubation procedures were performed using hybridization chamber and gaskets (Agilent #G2534A, #G2534-60003, #G2534-60011). Before binding reaction, microarrays blocked with 2% milk in PBS for 3-5 hours at room temp. Binding reactions carried out with ~0.25-4 µg of human HNF4α2 or HNF4α8 in NE from transfected COS-7 cells, or NE from WT and α7HMZ livers. NEs were diluted 1:10 in low salt PBM buffer (16 mM Hepes pH 7.8, 60 mM KCL, 8 mM EDTA, 8 mM EGTA) and processed through a 30 kDa cut-off column (Amicon, Cat# UFC503096) to a final concentration of 110 mM KCl and then applied to the arrays in PBM binding buffer (16 mM HEPES pH 7.8, 100 mM KCl, 8 mM EDTA, 8 mM EGTA, 0.1% Tween 20 plus 4-20 µg sonicated salmon sperm DNA). After 2 h of incubation, arrays were washed 3X for 3 min each with PBS plus 0.1% Tween 20. Mouse monoclonal anti-HNF4α P1/P2 (R&D Systems #PP-H1415-00) diluted 1:100 in PBS buffer plus 2% non-fat milk, 0.1% Tween 20 were applied directly to the slide and incubated for 24 h at RT, followed by a conjugated secondary Ab (GαM IgG [H+L] DyLight 550, Pierce Cat# 84540) diluted 1:50 (as described above) and then incubated for 90 min. Three washes, 3 min each, in PBS plus 0.1% Tween 20 were performed after each antibody incubation. HNF4α binding was imaged with 2-µm resolution using Agilent G2565CA Microarray Scanner at the UCLA DNA Microarray Core. Extraction and normalization of the data were as described previously (Bolotin et al., 2010). Position weight matrices (PWM) were generated using SeqLogo (Bembom 2017).

Rapid Immunoprecipitation and Mass Spectrometry of Endogenous Proteins (RIME)

Hypotonic Buffer: as in ChIP-seq

Nuclei Lysis Buffer: as in ChIP-seq

IP dilution buffer: as in ChIP-seq

Shearing Buffer D3: 0.1% SDS, 10 mM Tris-HCl pH 7.6, 1 mM EDTA in biology-grade water

Inhibitors: 2 μg/mL leupeptin, 2 μg/mL aprotinin, 1:100 protease inhibitor (Sigma #8340)

RIPA buffer: 15 mM Tris-HCl pH 8.0, 150 mM NaCl, 1% NP-40, 0.7% Deoxycholate

DNaseI Buffer: 40 mM Tris-HCl pH 8.0, 1 mM $CaCl_2$, 10 mM NaCl, 6 mM $MgCl_2$

RIME was performed as previously described (Mohammed et al., 2016) with the following modifications. WT and α7HMZ male mice (n=3, 16-18 weeks of age) were sacrificed at 10:30 (ZT 3.5). Roughly 250-mg chunks of liver were collected from the same livers used for RNA-seq samples and fixed in 5 mL formaldehyde solution (1.1% MeOH-free formaldehyde, 2 μg/ml aprotinin, 2 μg/ml leupeptin in 1X PBS) for 10 min at RT. Cross-linking was stopped with 0.125 M glycine at RT for an additional 5 to 8 min. Samples were centrifuged (4 min at 2,500 rpm) at 4^{o}C. The fixative was aspirated, the tissue immediately washed with cold PBS and centrifugation repeated. Fixed tissues were frozen and stored in liquid nitrogen.

Frozen tissues were placed in 1X PBS plus inhibitors and homogenized with a motorized homogenizer at 4^{o}C as described above. The homogenized samples were passed through nylon cell strainer, dounced eight times, and centrifuged (5 min at 2,500 rpm) at 4^{o}C. Cells were swelled in 1 mL Hypotonic Buffer for 10 min at 4^{o}C, spun again to collect the nuclei, which were resuspended with Nuclei Lysis Buffer (see ChIP-seq for buffers), nutated for 20 min at 4^{o}C and then gently resuspended in 0.5mL Shearing Buffer D3 (0.1% SDS, 10 mM Tris-HCl pH 7.6, 1 mM EDTA in biology-grade water) plus inhibitors. D3 buffer was added to fill the 1-ml sonication AFA milliTUBE (Covaris #520130). Samples were sonicated for 5.5 min, 30 s break, and again for 4 min in a Covaris S220 sonicator. Immediately after sonication samples were diluted 1:1 with IP dilution buffer plus all inhibitors. Samples were centrifuged (5 min at 11,000 rpm) and the pellet, if any, was discarded. Before IP, samples were pre-cleared for 40 min at 4°C with 10 µL of pre-washed magnetic Protein A/G beads (Pierce #0088802). One day before sonication, magnetic Protein A/G beads (20 µL per sample) were pre-washed 3X with 1 mL PBS plus 0.05% Tween 20. For each sample, 3 µg of P1/P2 Ab or mouse IgG diluted in 300-400 µL PBS plus 0.05% Tween 20 and inhibitors were added to pre-washed beads and nutated for 20 h at 4°C. On the day of sonication, unbound Abs were removed from the beads, and the sonicated, precleared samples were added to the beads and nutated overnight at 4°C. The following day, the supernatant was removed and beads were washed 3X with 1 mL ice-cold RIPA buffer, and then 1X with 400 µL DNaseI Buffer and incubated in DNaseI buffer with 8 µL DNaseI enzyme (4 µg/µL, Sigma #D5319) for 20 min at 30^{o}C. Afterwards, samples washed 3x with 0.5 mL RIPA buffer at room temperature and then 2X with 1 mL ice-cold RIPA buffer. Parallel IP samples (± DNA digestion) were examined for the presence of DNA by Qubit fluorometer to confirm high efficiency digestion. IP'd material was washed 2x with 1 mL 50 mM NH_4CO_3. At the last wash, the suspension was transferred to a new non-stick tube. The wash buffer was removed and the IP beads were immediately frozen and later subjected to mass spectrometry as described below.

Multidimensional protein identification technology (MudPIT) analysis was performed by the Proteomics Core Facility in the IIGB at the University of California, Riverside. Sample preparation following IP analysed by 2D nano-liquid chromatography tandem MS (2D nano-LC/MS/MS). Briefly, following IP, beads were washed in trypsin buffer [50 mM ammonium bicarbonate, 10% (vol/vol) acetonitrile] and digested overnight at 37 °C (1µg trypsin in 100 µL buffer) and washed one time [10 min 100 µL 50% (vol/vol) acetonitrile, 5% (vol/vol) acetic acid]. Digest supernatant and post-digest wash supernatant were combined. Tryptic peptides were pelleted by SpeedVac concentrator and redissolved in 20 µL 0.1% formic acid. Peptides were separated by a 2D nano-Acquity ultra-performance LC system (2D nano-UPLC) (Waters) and analyzed with an Orbitrap Fusion MS system (Thermo Fisher). High-pH reversed-phase LC with 20 mM ammonium formate pH 10 (solvent A) and 100% (vol/vol) acetonitrile (solvent

B) was used to fractionate the tryptic peptides into five fractions on an XBridge BEH130 C18 trap column [5µm particle, 300-µm internal diameter (i.d.), 5-cm long; Waters #186003682]. Five fractions and a flush fraction were collected at (1) 13%, (2) 18%, (3) 21.5%, (4), 27%, (5) 50%, and a final flush of 60% solvent B. Each of these fractions was first concentrated and then separated with a conventional reverse phase gradient in acidic condition. A Symmetry C18 column (5-µm particle, 180-µm i.d., 20-mm long, Waters #186003514) was used to concentrate and desalt the peptides of each fraction. The samples were further separated on a BEH130 C18 column (1.7-µm particle,75-µm i.d., 20-cm long, Waters #186003544). The mobile phase A and B solvents for separation gradient were 0.2% formic acid in water and 0.2% formic acid in acetonitrile, respectively. The mobile phase nano-flow rate was 0.3 µL/min with the following 1-hour gradient: 0–1 min, 3%B; 1–30 min, 50% B; 30–31 min, 85%B; 31–35 min, 85% B; 35–36 min, 1% B; and 36–60 min, 1% B.

The MS analysis part of MudPIT was carried out with Orbitrap Fusion MS system (Thermo Fisher). A data-dependent acquisition (DDA) survey method using HCD (high-energy collision dissociation) fragmentation technique was employed in a positive ion mode. The instrument parameter included ESI spray voltage at 2300 V, ion transfer tube temperature 275°C, and 0 sweep gas. MS1 scan was carried out with Orbitrap mass analyzer with its resolution set at 120,000 and normal mass range from 300 to 2000 m/z. S-Lens RF level was 60%. AGC target was set 200,000. 50 msec was set for maximal injection time. Microscan was set for 1. For MS2 scan, top-speed scanning method was used with time window of 4 seconds. Monoisotopic selection was allowed, peptide ions with charge state from 2 to 5 and other undetermined charge states were all selected for MS2 fragmentation by HCD. Dynamic exclusion was activated after three MS2 spectra were acquired for each m/z within 1 min window. Dynamic exclusion duration was for 5 min, and exclusion mass window was +/- 30 ppm. Ion selection for MS2 acquisition was arranged from most intense peak to least intense peak with minimal intensity of NL 100,000. For HCD fragmentation, quadrupole isolation window was fixed at 2 m/z, collision energy was set at 30%, ion trap was chosen as mass detector to collect all MS2 fragments for each individual peptide ions. Ion trap scan rate was set at rapid with scan range set at normal. First fragment mass was set at 120 m/z. AGC target was set at 10,000 with maximal injection time of 0.1 second. All raw MS1 and MS2 spectra were processed with Proteome Discoverer 2.1 (Thermo Scientific) to generate mgf files, which were then submitted to Mascot searching engine to match against NCBI non redundant (nr) human protein database for protein identification. Only proteins with 1% FDR cut-off ($q \leq 0.01$) were considered for subsequent analysis.

Area under the curve, as reported by Proteome Discoverer, were averaged together for WT and α7HMZ samples (n=3). IgG samples (n=3) from both WT and α7HMZ were averaged together to create a background sample. Areas were converted to log2 scale and the fold-change above IgG background was calculated for the WT and α7HMZ samples. Proteins with less than 10-fold change above background were omitted.

Similarly, a 10-fold difference between WT and α7HMZ samples was used to identify unique protein interactions.

Dexamethasone treatment and Glucose Tolerance Test (GTT)

WT and α7HMZ mice (16-24 weeks old, backcrossed into C57bl/6N) were injected with either 4 mg/kg body weight dexamethasone (Dexamethasone 21-phosphate di-sodium salt (Sigma #D1159-100 MG) dissolved in vehicle (0.9% sterile saline (NaCl)) or vehicle alone (0.9% saline) for 7 days at the same time (11 AM) each day. On the eighth day, mice were fasted for five hours before undergoing glucose tolerance test (GTT). GTT was performed as previously described (Deol et al., 2015).

External expression datasets

The following mouse differential expression analyses were used to generate scatterplots for Figure 2.1. The HNF4α mouse liver knockout (KO) data was generated by (Walesky et al., 2013) using Affymetrix Mouse430_2.0 genechips. Data were summarized to the gene level taking the largest fold change value and associated p-value for a single gene if more than one transcript was reported. The adult vs fetal differential expression analysis was generated by the ENCODE project (ENCSR000BYS, ENCSR000BZI). Gene quantifications files were downloaded and RSEM expected_count values treated as raw read counts for differential expression analysis with DESeq2, with log2 Fold Change (log2FC) and adjusted P-values (padj.) used for plotting and highlighting values in scatterplots. The C57BL/6 vs Hepa1-6 differential expression analysis was generated with DESeq2 (Rudolph et al., 2016). Similarly, log2FC and padj were used for plotting and highlighting values in scatter plots, which were graphed using ggplot2 library in R.

Oxylipin analysis

For the analysis of non-esterified oxylipins, plasma aliquots (250 μL) or liver tissue homogenates (100 mg) were extracted and analyzed according to previously described protocols (Matyash et al., 2008; Yang et al., 2009). Briefly, samples were extracted by solid phase extraction and analyzed by ultrahigh performance liquid chromatography tandem mass spectrometry (UPLC-MS/MS) (Agilent 1200SL-AB Sciex 4000 QTrap). Analyst software v.1.4.2 was used to quantify peaks according to corresponding standard curves with their corresponding internal standards. Oxylipin concentrations are presented as pmol/gm in tissue. Data are presented as mean +/- standard error of mean (SEM). Student's T-test was used to determine statistical significance ($p < 0.05$) using GraphPad Prism v6.

Graphical and statistical analysis

Differential gene expression (DEG) was measured using raw read counts with DESeq2. Statistical significance was measured as adjusted p-value (padj.) ≤ 0.01, unless otherwise noted. Legends denote any thresholds using log2 fold change (log2FC) cutoffs.

ChIP-seq peaks were called with MACS2 and then filtered on -log10(P-value) ≥ 10, i.e., p-value ≤ 1e-10, to approach six-fold enrichment above control. Differentially bound peaks were identified using DiffBind with MACS2 output. RIME samples were analyzed with Proteome Discoverer 2.0: areas reported were converted to a log2 scale, thus fold changes were calculated on the log2 scale. Methods to filter RIME data discussed above. All heatmaps generated with pheatmap package in R. Heatmap data were row-normalized before plotting with the exception of NR heatmap in Figure 2.4. All barplots represent mean ± SEM. Transcription Factor (TF) rankings for Cleveland plots were ordered at the 13:30 (peak HNF4α expression) then manually curated with the aid of PANTHER; all TF genes with FPKM > 50 plotted using the 10:30 FPKM values. All Venn diagrams were generated with VennDiagram package in R. Unique and common RIME results were submitted to DAVID for ontology analysis. Individual statistical significance tests performed as two-way Student's T-test unless otherwise noted.

Results

Impact of the P1- and P2- HNF4α N-terminal domain (NTD) on Liver Gene Expression

When the HNF4α exon-swap mice were created, the expression of several liver genes were noted as being AF-1 dependent, including *Apoa4*, *Apoc3*, and *Nr1i3* (CAR) (Briançon and Weiss, 2006). Subtle differences in metabolism were also noted under unstressed conditions but no global analysis of gene expression was performed. We therefore decided to explore in greater depth the impact of the P1-HNF4α isoform

(HNF4α1) expressed in WT livers compared to the P2-HNF4α isoform (HNF4α7) expressed in the exon swap α7HMZ mice (Fig 2.1A, *top*). To first verify that the livers express the correct isoforms, immunoblot (IB) assays were performed on adult liver nuclear extracts (NE) using antibodies that are either specific to P1-HNF4α (P1) or that recognize both P1- and P2-HNF4α (P1/P2) (Fig 2.1B). The P1 antibody detected P1-HNF4α protein in WT adult (aged 16-18 weeks) and fetal livers (E18) but not in the α7HMZ livers, as anticipated. In contrast, the P1/P2 antibody detected a faster migrating protein in the α7HMZ adult and fetal livers, consistent with the reduced size of the HNF4α7 isoform, and the PCR genotype of the mice (not shown). There were also faster migrating bands in the WT IB detected by the P1/P2 antibodies which are most likely breakdown products as they do not migrate exactly as the bands in the α7HMZ samples and they vary from preparation to preparation. The P2-HNF4α protein was not detected at E18 in the WT fetal liver, even though transcripts have been detected by PCR (Torres-Padilla et al., 2001).

HNF4α isoforms preferentially regulate genes in fetal liver and liver cancer

To investigate whether there are global differences in gene expression in the WT and α7HMZ livers, an RNA-seq study (n=3, aged 16-18 weeks) was performed on adult livers. The aligned reads confirmed expression of the correct first exons within the WT and α7HMZ livers (Fig 2.1C). A comparison of normalized FPKMs across the duplicate samples revealed a significant (padj. ≤ 0.01) difference in 1600 genes between WT and α7HMZ livers, 831 up-regulated and 792 down-regulated in α7HMZ livers (Fig 2.1D), confirming an important role for the HNF4a NTD in the adult liver. The most down-regulated genes in α7HMZ are *Scnn1a, Cyp2c50, Rdh16f2,* and *Ces2e* and the most up-regulated are *Rad51b, Cyp2b13, Pcp4l1,* and *Cyp2b9* (Fig. 2.1E). The differentially expressed genes (DEG) were next compared to murine HNF4α liver knockout (KO) microarray expression data (Walesky et al., 2013) to identify potential HNF4α target genes (Fig. 2.1F). A small number of the significantly (padj. ≤ 0.01) α7HMZ-expressed genes were up-regulated upon P1-HNF4α KO (red spots), whereas many of the significantly WT-expressed genes were down-regulated upon P1-HNF4α KO (blue spots).

Since it has been shown that HNF4α is a tumor suppressor in mouse liver (Hatziapostolou et al., 2011; Lazarevich et al., 2004; Ning et al., 2010; Walesky et al., 2013) the WT and α7HMZ DEGs were plotted against RNA-seq data from a murine liver cancer model (Fig 2.1G). Fold changes from a differential expression analysis of C57BL/6 against murine hepatoma cell line Hepa1-6 were compared against WT vs α7HMZ expression changes. Genes that were significantly (padj. ≤ 0.01) up-regulated in the WT liver were preferentially expressed at higher levels in normal liver tissue compared to liver cancer (562 vs. 87) and genes significantly better expressed in α7HMZ mice were up-regulated in both normal and cancer cell lines equally (293 vs. 281). It is known that P2-HNF4α is not normally expressed in a healthy adult liver, but that it is expressed, at least on the RNA level, during fetal liver development (Dean et al., 2010;

Torres-Padilla et al., 2001). Thus, WT vs α7HMZ DEGs were compared with a differential expression analysis of ENCODE-generated mouse adult liver vs fetal liver (embryonic 14.5, E14.5) (Fig 2.1H). Like the cancer model comparison, most of the genes more highly expressed in the WT livers were also up-regulated in the mouse adult liver, while only a small proportion were more highly expressed in fetal livers (562 vs. 87). Genes that were up-regulated in the α7HMZ livers showed less preference for adult vs. fetal expression (335 vs. 269).

In vivo chromatin binding of HNF4α isoforms

To determine whether the altered gene expression in the WT and α7HMZ livers was due to alterations in the HNF4α isoforms binding DNA in the WT and α7HMZ livers, ChIP-seq analysis was performed with livers harvested at 10:30am. While the results revealed a large number of HNF4α binding events in the mouse liver (WT: 40,429; α7HMZ: 40,472), only a small percentage of those sites were enriched for a particular isoform (WT: 336; α7HMZ: 379) (Fig 2.2A). An analysis of the feature distribution of the ChIP-peaks showed that both WT- and α7HMZ-specific peaks were less frequently located in the promoter region, specifically ≤ 2kb from +1, than the peaks common to both genotypes (Fig 2.2B). Additionally, while both isoforms exhibited an increase in the number of unique peaks in distal intergenic regions, compared to the common peaks, the α7HMZ unique peaks were more frequently found in the intronic regions compared to the WT-unique peaks.

To determine whether any of the isoform-unique ChIP peaks were due to isoform-specific DNA binding specificity, we employed a support vector machine (SVM) trained on experimentally verified HNF4α DNA binding sites (Bolotin et al., 2010). For each uniquely bound peak, all possible 13-mers within a 200-bp window around the peak center, from both strands, were submitted to an SVM scoring algorithm. Each peak was categorized into one of four categories (>2, 2 to 1.75, 1.75 to 1.5 and <1.5) based on the the single highest-scoring SVM motif within the peak. Each bin of peaks was then submitted for *de novo* motif calling via MEME-ChIP. Both the SVM and MEME analysis highlight subtle changes in the canonical HNF4α motif between WT and α7HMZ, such as an adenine spacer between the two half sites of the direct repeat which creates a core CAAAG motif characteristic of HNF4α sites (Fang et al., 2012) (Fig 2.2C/D). The top MEME results from WT-specific peaks show HNF4α-centric CAAAG motifs in all the SVM categories except for the one with the weakest HNF4a motifs (<1.5), while α7HMZ-specific peaks had the HNF4α-centric core motif only in the top SVM category. Binding motifs for the STAT family of TFs appeared twice in the WT-specific peaks but were not found in the α7HMZ-specific peaks. In contrast, the GC-rich TF binding sites for Sp1/KLF-like factors were found in both WT- and α7HMZ-unique peaks although the Sp1-like motif was found in the second SVM category of α7HMZ-specific peaks (2 to 1.75) and in the last SVM category (<1.5) of WT-specific peaks. These results suggest that *in vivo* there may be subtle differences in the DNA binding specificity of the HNF4α isoforms, despite containing identical DNA binding domains, and that other TFs such as

STAT and SP1/KLF may play a role in the differential chromatin binding of, and hence gene regulation by, the HNF4α isoforms.

DNA binding affinity of HNF4α isoforms in vitro

To further investigate the DNA binding specificities of HNF4α isoforms we designed protein binding microarrays (PBMs) to examine variations on NR consensus sequences (a direct repeat with a spacing of 1, DR1, AGGTCANAGGTCA, and a DR2, AGGTCANNAGGTCA) as well as genomic sequence mined from HNF4α ChIP-seq peaks from human colon adenocarcinoma cell line, Caco-2 (Fig 2.3A). The PBMs are a high throughput DNA binding assay in which one can examine 45,000 test sequences in quadruplicate for a total of 180,000 DNA binding reactions. Application of human HNF4α2 (P1-HNF4α) and human HNF4α8 (P2-HNF4α) ectopically expressed in COS-7 cells to the PBM revealed that the two isoforms have nearly identical DNA binding affinity across all 45,000 test sequences (Fig 2.3B, *left*). PBM analysis of nuclear extracts from WT and α7HMZ mouse livers revealed two distinct groups of DNA sequences (Fig 2.3B, *right*). Motif analysis of those two groups with MEME, shown in green and red, highlights a preference for WT HNF4α (P1-HNF4α) to interact with GC-rich sequences recognized by Sp1 (Fig 2.3C, *right,* red spots). These spots were identified in the COS-7 scatterplot and highlighted in the same fashion to reveal a similar preference for P1-HNF4α (HNF4α2) to interact with GC-rich binding sites.

HNF4α is the most highly expressed NR in the liver

NRs are known to play an important role in regulating and/or responding to circadian rhythms, especially RORα/β/γ, Rev-erbα/β, GR, and CAR (Schmutz et al., 2010; Tahara and Shibata, 2016; Yang et al., 2006; Zhang et al., 2009; Zhao et al., 2014). Since HNF4α is an important regulator of liver-specific gene expression, RNA-seq in WT and a7HMZ livers was performed at three different time points (10:30am, 1:30pm, 8:30pm) to determine whether the HNF4α isoforms play a role in regulating or responding to circadian rhythms (Fig 2.4A). Rapid immunoprecipitation and mass spectrometry of endogenous proteins (RIME) was also performed from the same liver samples at the 10:30am time point. The expression analysis revealed that HNF4α is the most highly expressed NR in the liver, followed closely by RXRα, LXRα (*Nr1h3*), and PPARα (Fig 2.4B). While many NRs display circadian changes in expression across all three time points, CAR (*Nr1i3*) is the only NR completely dysregulated between WT and α7HMZ livers. When comparing significant (padj. ≤ 0.01, absolute log2FC ≥ 1) DEGs between time points, WT livers showed an increased number of changes compared to α7HMZ at every comparison (Fig 2.4C, *top*). Furthermore, there was a larger number of genes down-regulated versus up-regulated in α7HMZ compared to WT livers at each time point (Fig 2.4C, *bottom*). A volcano plot of the DEGs between WT and α7HMZ livers at 10:30am highlights this effect, with more significantly WT-expressed genes (*Rdh16f2*, *Cyp2c50*, *Ces2e*, *Nr1i3*) with positive fold changes in the right half of the plot and fewer α7HMZ-expressed genes (*Rad51b*, *Cyp2b13*) with negative fold changes in the left half (Fig 2.4D).

After finding HNF4α to be the most highly expressed NR in the liver, based on transcript levels, we examined where it ranked among all the TFs in the liver. Genes were sorted by WT FPKM values at the 1:30pm time point and then manually curated to select all TFs with FPKM ≥ 50; liver-enriched transcription factors (LETF) *Hnf1a* and *Hnf1b*, which had FPKM <50, were added in manually for comparison. A Cleveland dot plot was generated to show both WT and α7HMZ FPKM values from the 10:30am time point, ranked by WT FPKM (Fig 2.4E). *Hnf4a* was the fourth most highly expressed TF after *Atf5*, *Srebf1* and *Btf3*, with higher FPKM values than subunits of RNA polymerase (e.g., *Polr2m*, *Polr2b*), *Stat3* and *Stat6*. Tellingly, the other LETFs (*Cebpa*, *Onecut2*, *Foxa1*, *Hnf1a* and *HNF1b*) all had much lower transcript levels than *Hnf4a* (<100 vs. >1000 FPKM). It is also of interest to note that there were several TFs that showed statistically significant differences in gene expression between WT and α7HMZ (padj. ≤ 0.01), including *Hnf4a* (denoted by an asterisk in Fig 2.4E), which could contribute to the differential gene expression observed between WT and α7HMZ livers.

HNF4α isoforms result in altered metabolic profiles

Gene expression analysis showed very strong repression of the NR CAR (*Nr1i3*) at all three time points. CAR is a well-known regulator of cytochrome P450 (*Cyp*) genes, as well as many other Phase I and II enzymes involved in the detoxification of xenobiotics and metabolism of drugs (di Masi et al., 2009; Tolson and Wang, 2010; Willson and Kliewer, 2002). Three families of Phase I and II enzymes showed dysregulation of several members -- Cyp enzymes, glutathione S-transferases (GST), and UDP glucuronosyltransferases (UGT) (Fig 2.5A). For all three families of enzymes there are multiple members up- or down-regulated in the α7HMZ livers. Generally, the circadian pattern of expression across the three time points is unchanged for these enzymes, but the amplitude, or absolute level of expression is dysregulated. For example, *Cyp2c55* maintains a steady decreasing level of expression throughout the day in the WT liver, and a similar decreasing trend in the α7HMZ liver, but with lower values at each time point. Similarly, *Ugt1a9* maintains a cyclical pattern by showing decreased levels of expression at 1:30pm and increased levels of expression at 8:30pm in WT liver, and the pattern is repeated for the α7HMZ liver but with much lower expression levels at each time point.

While CAR (*Nr1i3*) is down-regulated in α7HMZ mice, PXR (*Nr1i2*), which is known to co-regulate many Phase I and II genes with CAR (di Masi et al., 2009; Tolson and Wang, 2010; Willson and Kliewer, 2002), is relatively unchanged between WT and α7HMZ (Fig 2.5B). Changes in gene expression in the steroid metabolism pathway were also observed in α7HMZ livers with an increase in *Cyp17a1* and a decrease in *Srd5a1* (Fig 2.5B, *top*). To determine whether there are functional differences in steroid metabolism in the exon swap mice, mice were injected with dexamethasone, a steroid agonist of GR, daily for 7 days, and then fasted for 5 hours before undergoing a glucose tolerance test (GTT). The results show that dexamethasone improves the glucose tolerance of WT mice, as evidenced by a smaller area under the curve in the GTT (Fig 2.5C, *left*). In contrast, in α7HMZ mice the GTT profile was essentially identical for the dexamethasone- and saline-injected mice (Fig 2.5C, *right*), suggesting that the steroid was more rapidly metabolized in α7HMZ livers, consistent with increased expression of *Cyp17a1*, which metabolizes dexamethasone, and decreased expression of *Srd5a*, which is a 5α-reductase of steroids that can convert cortisol and corticosteroids into 5α-dihydrocortisol and 5α-dihydrocorticosteroids, respectively.

Potential changes in fatty acid metabolism pathways were also observed in α7HMZ livers with decreases in *Cyp2b10* and *Ephx2* (Fig 2.5B, *bottom*). It is known that both *Cyp2b10* and *Ephx2* play a role in fatty acid metabolism, specifically in converting arachidonic acid into oxylipins (Wagner et al., 2011). Oxylipin analysis of WT and α7HMZ livers showed that all four DiHETrE oxylipin substrates of arachidonic acid are found at significantly lower levels in α7HMZ livers (Fig 2.5D), confirming a functional effect of the lower levels of *Cyp2b10* and *Ephx2* transcripts. Interestingly, 14,15-DiHETrE is a potent activator of PPARs (Fang et al., 2006; Ng et al., 2007).

HNF4α isoforms have unique interactomes

In order to determine whether the differences in chromatin binding between the two HNF4α isoforms are driving the changes in gene expression we cross-referenced the ChIP-seq and RNA-seq datasets at the 10:30am time point. For each uniquely bound peak, the single closest transcription start site (TSS) was annotated in a 100 kb window (±50kb). To identify WT-specific correlations, we cross-referenced the TSS from annotated unique WT peaks with all genes significantly expressed in WT livers (padj. $\leq$ 0.01 and log2FC $\geq$ 1) and found that 14.15% (45 out of 318) of these genes had a unique peak nearby (Fig 2.6A, *left*). Similarly, to identify α7HMZ-specific correlations, the TSS from annotated unique α7HMZ peaks was cross-referenced with all genes significantly expressed in α7HMZ livers (padj. $\leq$ 0.01 and log2FC $\geq$ 1): 11.73% (23 out of 196) of these genes had a unique peak nearby (Fig 2.6A, *left*). WT-specific genes matching these criteria included *Nr1i3*, *Cyp2c50*, *Cyp2c54*, *Rarres1*, *Fmn1*, *Cdhr5*, and *Camk1d* (Fig 2.6B, *left*), and α7HMZ-specific genes matching this criteria included *Cyp2b9*, *Fgfr1*, *Wnk4*, *Cyp4a14*, *Ppl*, *Vnn1*, *Acot1*, and *Cyp17a1* (Fig 2.6B, *right*). Many of the most dysregulated genes between these genotypes contained differentially bound peaks at +1, or within the first 2 kb downstream of +1, typically falling in intronic regions (Fig 2.6C).

Since the majority of dysregulated genes had no nearby HNF4α isoform-specific ChIP-peaks, we compared the protein-protein interactions between these isoforms in order to help explain the transcriptional differences. Rapid immunoprecipitation and mass spectrometry of endogenous proteins (RIME) was performed on liver tissue from WT and α7HMZ mice. Area under the curve (AUC) for each identified peptide aligning to a protein was totaled and these values were used to measure for enrichment above background samples (IgG controls). To be included for analysis, proteins needed to show $\geq$10 fold increase above background, and then to be considered uniquely interacting proteins needed to show ≥10 fold change between WT and α7HMZ samples. The RIME results show several LETFs interacting with both isoforms (CEBPA, CEBPB, HNF1A, HNF1B, and ONECUT2) as well as several NRs (ESRRA, NR0B2, NR5A2, THRB) (Fig 2.6D). GR and HNF4α regulate many of the same genes in the liver and was also identified in our RIME results to interact with both isoforms, but it was above background at only 3.4-fold change. Clock-related proteins BHLHE40 and CRY1 were also identified in the common interacting proteins, along with other TFs (KLF13, PROX1, RELA, RBBP4, and RBBP7). Several NRs were shown to interact specifically with P1-HNF4α (NR1H2, NR2C1, NR2C2, and PPARA) while only one NR (PXR, *Nr1i2*) interacted specifically with P2-HNF4α. Core circadian components CLOCK and BMAL (ARNTL) were shown to interact uniquely with P2-HNF4α while clock-related transcription factor NFIL3 interacted uniquely with P1-HNF4α. In total, more than 130 transcription factors (TF) came down with either isoform, of which nearly half interacted uniquely with HNF4α in either WT or α7HMZ livers (Fig 2.6E, *left*). Interestingly, many more RNA processing related proteins were identified in the α7HMZ-only samples compared to both common and WT-only samples (Fig 2.6E, *left*). Normalizing these counts to the total number of proteins identified in the sample revealed that the proportion of uniquely bound TFs between the isoforms was nearly equal although there was a slight increase in the proportion of RNA processing proteins among the α7HMZ-only proteins (Fig 2.6E, *right*).

HNF4α isoforms impact the circadian response

Interactions with circadian-related proteins in the RIME data suggest that HNF4α may potentially play a role in regulating or responding to circadian stimuli in the adult liver. Differentially expressed genes (DEG) were identified between any two time points (10:30am, 1:30pm, and 8:30pm) for both the WT and α7HMZ livers (padj. ≤ 0.01 & log2FC ≥ 2). There were 53 genes that showed extreme fluctuations of expression throughout the day, including commonly known circadian genes (*Cry1*, *Rorc*, *Dbp*, *Bhlhe41*, *Usp2*, *Per2*, *Per3*, *Arntl*, and *Nr1d1*) as well as many metabolism-related genes (*Fmo3*, *Lpl*, *Car3*, *Corin*, *Npas2*, *Hmgcs1*, *Mme*, *Slc45a3*, *Hsd3b4*, *Hsd3b5*, *Slc10a2*). Visualizing the normalized expression in a heatmap shows that while many of these genes do show wide ranges of expression across the three time points, nearly all of them show the same pattern of expression between the WT and α7HMZ livers (Fig 2.7A). Most of the differences between the genotypes appear to be either an increase or decrease of expression at a specific time point and not a change in the pattern of expression throughout the day. *Aqp8* is a good example of this: transcript levels go from high levels at 10:30am to moderate levels at 8:30pm in WT livers, and have the same trend in expression in α7HMZ livers except that they start at much lower levels.

In order to determine whether HNF4α isoforms play a role in responding to or regulating the circadian clock, we looked for differential expression of core circadian components and a few metabolism-related clock controlled genes across all three time points. Significant differences in expression between WT and α7HMZ livers were found in BMAL (*Arntl*) at both 10:30am and 8:30pm *Clock* and *Cry1* at 8:30pm; *Per3* and *Ppara* at 1:30pm; and *Rorc* at 10:30am (Fig 2.7B). The fact that other core components of the clock machinery did not show significant differences between the two genotypes (*Per1*, *Per2*, *Rora*), (Fig. 2.7B) also suggests that the effect of the HNF4α isoforms is a specific one. A heatmap of the core circadian components shows that apart from a few small perturbations, the core components maintain cyclic expression throughout the day in both genotypes (Fig 2.7C).

A sample distance matrix heatmap, generated as part of quality control, confirmed a subtle yet real effect of the HNF4α isoforms on the global clock (Fig. 2.7D). In the WT mouse the three samples for each time point are much more similar to each other than they are to any other time point. In contrast, in the α7HMZ samples, the only strong self-identity was seen in the 1:30pm samples (Fig 2.7D). The 10:30am and 8:30pm time points share nearly as much similarity across all three time points than they do to themselves. Since a principal component analysis (PCA) (Fig 2.S1) showed a good separation and categorization of each sample group, the differences revealed in the sample distance matrix are not likely to be due to more quality of data.

HNF4α isoforms exhibit distinct fasting response

To further examine the role of the HNF4α isoforms in the adult liver, we examined how each isoform responded to the stress of fasting. Mice from both WT and α7HMZ backgrounds were fasted for 12 hours prior to harvesting liver samples at

10:30am for RNA-seq. Categorizing significant DEGs (padj. $\underline{<}$0.01 and log2FC ≥ 1) between the fed and fasted states revealed 679 and 549 dysregulated genes for WT and α7HMZ, respectively (Fig. 2.8A). Looking more closely at the dysregulation between WT and α7HMZ in the fasted state, we found 341 genes up-regulated and 330 down-regulated in α7HMZ livers compared to WT (Fig 2.8A). These numbers were much higher than those in the fed state, especially in terms of the up-regulated genes: 176 up and 289 down in α7HMZ livers compared to WT.

In order to determine how many of the genes up- and down-regulated in both livers are uniquely regulated, we compared the fasted vs. fed datasets for each genotype. We found that of all the up-regulated genes in fasted WT livers, 56.77% (197 of 347) were uniquely regulated and of all the up-regulated genes in fasted α7HMZ livers, 50.33% (152 of 302) were uniquely regulated (Fig 2.8B, *left*). Similarly, up-regulated genes in WT livers showed 62.65% (208 of 332) uniquely down-regulated and α7HMZ livers showed 49.79% (123 of 247) uniquely regulated (Fig 2.8B, *right*). Finally, the sample distance matrix shows that the fasted livers are distinct even when compared to any time point from their own genotype, although in this case the three α7HMZ samples were more tightly correlated than the three WT samples (Fig. 2.8C). Taken together, the results from the fasting RNA-seq show that the HNF4α isoform-specific mice respond in a differential fashion to fasting on the level of gene expression, consistent with the original report of the α7HMZ mice showing a fattier liver upon fasting than the WT mice (Briançon and Weiss, 2006).

Discussion

Alternative promoter usage has been observed for many genes in the human genome, and it is estimated that over 50% of genes in the human and mouse genomes contain at least one alternative promoter (Baek et al., 2007; Pal et al., 2011). For most of these genes the physiological relevance of the transcript variants is unknown. The mouse (and human) *Hnf4a* gene has two highly conserved promoters, the proximal P1 and distal P2, that drive expression of alternative first-exons which result in alternative N-terminal domains (NTD). Both promoters are expressed in fetal livers while only the P1 promoter is expressed in adult livers. HNF4α knockout mice are embryonic lethal and adult liver HNF4α knockout mice die within six weeks with a fatty-liver phenotype. While P1-HNF4α has been shown to be a tumor suppressor in the adult liver and P2-HNF4α has been shown to up-regulated in some liver cancers, very little is known about the role that P2-HNF4α plays in the liver. Regulating expression of this isoform via an alternative promoter allows for tissue- and developmental-specific expression of a slightly modified protein with altered function.

Dysregulation of genes in the liver by P2-HNF4α

Since the only active promoter in the adult liver is the P1 promoter, we utilized previously characterized exon-swap mice to express P2-HNF4α in the adult liver. Here, we show that a 16 amino acid difference in the NTD of HNF4α is enough alter the expression of nearly 1,600 genes (Fig. 2.1D). Comparison with P1-HNF4α knockout data suggests that while P2-HNF4α may be sufficient for developing and maintaining a functioning adult liver, much of the dysregulation between WT and α7HMZ adult livers may actually be due to a loss of P1-HNF4α. Our results also show that genes more highly expressed in WT livers are preferentially expressed in normal liver tissue compared to hepatoma cell lines, while α7HMZ livers seem to be more permissive to cancer-specific gene expression with nearly an equal number of these genes up-regulated in either normal liver tissue or hepatoma cell lines (Fig. 2.1). These transcriptomic results are completely consistent with what is already known about the role of P1- and P2-HNF4α in liver cancer: P1-HNF4α is a tumor-suppressor and P2-HNF4α is up-regulated in many liver HCC cell lines.

Our results also show that in adult α7HMZ livers there is a large number of genes that normally exhibit fetal liver-specific expression, suggesting that P2-HNF4α plays an important role in the developing liver even though it is expressed at a lower level than P1-HNF4α (Briançon et al., 2004). Furthermore, it has been shown that P2-HNF4α is the primary HNF4α isoform expressed in the pancreas (Harries et al., 2008) and that, while fetal livers show dramatic gene expression changes when compared to adult livers, they also exhibit a gene expression profile similar to that of the pancreas (Lee et al., 2012), confirming a role for P2-HNF4α in the fetal liver.

One of the most notable findings from our study is a near complete repression of NR CAR (*Nr1i3*) expression in the α7HMZ livers while the expression of PXR (*Nr1i2*), which often partners with CAR, was not affected (Fig. 2.5B). It has been previously suggested that CAR accelerates the maturation of human hepatic-like cells derived from hESCs but that the NR PXR plays no role (Chen et al., 2013). While those findings are consistent with CAR (but not PXR) being up-regulated by P1- but not P2-HNF4α, since the α7HMZ mice are viable and develop apparently normal livers, it also suggests that CAR is not absolutely required for liver development *in vivo*, at least in the artificial environment of a research vivarium.

Nonetheless, there was a whole-scale dysregulation of genes encoding Phase I and II drug and xenobiotic metabolizing enzymes in α7HMZ livers (Fig. 2.5A), presumably due to the downregulation of CAR, which is well known to regulate those genes. Therefore, under more realistic conditions of exposure to environmental toxicants, the isoform of HNF4α that is expressed in the liver is likely to be very important. Finally, the finding that an HNF4α isoform that is specifically expressed in the fetal liver alters the expression of many of the genes involved in Phase I/II metabolism is consistent with the well-established difference in drug and xenobiotic metabolism between infants and adults (Cui et al., 2012; Hart et al., 2009).

HNF4α isoforms exhibit similar DNA binding specificity but altered partners in chromatin binding

Our results from ChIP-seq and PBM experiments show that both *in vitro* and *in vivo* the HNF4α isoforms exhibit nearly identical DNA binding affinity. Considering that the P1- and P2-HNF4α isoforms only differ by 16 amino acids in the N-terminal domain

(NTD) and share 100% identical DNA binding domains, this result does not come as a surprise. ChIP-seq followed by HNF4α SVM motif search revealed a small percentage of peaks were uniquely bound and contained minor variations to canonical HNF4α binding motif. Categorization of uniquely bound peaks by SVM results allows us to identify peaks with potentially direct- and indirect-binding of HNF4α isoforms based on the strength of the HNF4α binding motif identified. Thus, TF DNA binding motifs found in the "strong" SVM categories (≥2, 2-1.75 and 1.75-1.5) are highlighting direct HNF4α binding sites with other TFs interacting with complexes or co-binding with HNF4α and motifs found in the "weak" SVM categories (1.5-1.25) are highlighting indirect HNF4α binding sites with other TFs tethering HNF4α to these binding sites. This interpretation helps explain the Sp1-like binding motifs that were differentially bound in the PBM results. While these sequences showed a preference to the WT samples, we did not find binding motif results for Sp1-like motifs in the strongest SVM categories, but rather in the weakest. This would suggest that in our WT-specific ChIP-peaks with Sp1-like motifs we may be seeing the indirect binding of HNF4α to the DNA via tethering interaction with Sp1, or another Klf protein. Conversely, α7HMZ samples identified an Sp1-like motif in a strong SVM category where we are seeing direct HNF4α binding to a strong HNF4α motif, and Sp1/Klf is binding nearby.

Differential regulation of Cyp genes and gender differences in a7HMZ mice

Perhaps the most notably dysregulated gene between the two livers is the NR CAR (*Nr1i3*) which has been shown to play a crucial role in coordinating responses to exogenous and endogenous chemicals by the regulation of its target genes. Our results have shown that the loss of CAR leads to dysregulation of many Phase I & II enzymes, including cytochrome P450s (Cyp), and that these changes seem to have an impact on not only fatty acid metabolism but also on response to glucocorticoid signaling.

The regulation of steroid hormone metabolism and signaling from the liver also plays a considerable role in the maintenance of sexually dimorphic gene expression. One TF that plays a crucial role in the regulation of sexual dimorphism in the adult liver is STAT5b which responds to a variety of growth factor signals, one of which is growth hormone (GH). The cyclical release of GH into the bloodstream by the pituitary gland is stimulated by exercise, nutrition, sleep and inhibited by free fatty acids and glucose (Hartman et al., 1993). The clearance and metabolism of xenobiotics differs between males and females due largely in part to differences in expression of Cyp genes and different patterns of secretion of hormones like GH (Dhir et al., 2006; Mugford and Kedderis, 1998; Wolbold et al., 2003). STAT5b is a major mediator of sex-dependent gene expression in the liver and impacts the expression of many genes involved in the metabolism and clearance of xenobiotics. The dysregulation of many of these sex-dependent enzymes in the α7HMZ mice was perplexing at first until ChIP-seq motif analysis showed that peaks bound uniquely by P1-HNF4α were often accompanied by

STAT-like motifs. While we did not see any evidence for unique protein-protein interactions between STAT5b and either HNF4α isoform, these data would suggest a potential regulatory difference between P1- and P2-HNF4α with regards to STAT5b and sex-dependent gene expression in the liver.

Altered protein-protein interactions among the HNF4a isoforms

Mass spectrometry results identified many TFs in both WT and α7HMZ datasets of interacting proteins. Gene ontology analysis revealed slight deviations in the number of DNA repair and RNA processing related proteins, with α7HMZ having a much higher representation of RNA processing related proteins. Recent studies increasingly link RNA processing proteins with DNA damage repair (Naro et al., 2015; Wickramasinghe and Venkitaraman, 2016). Taking into context that the natural environment for P2-HNF4α is in the developing fetal liver, there may be a role for this isoform in regulating DNA damage repair and RNA processing to facilitate a safer proliferative environment. We also find that two of the top three largest common HNF4α ChIP-seq peaks are found within 10kb of TSS for Holliday junction recognition protein (*Hjurp*), a known HNF4α target, which is a histone chaperone that is often significantly overexpressed in many cancers (Filipescu et al., 2017; Hu et al., 2010; Montes de Oca et al., 2015). While *Hjurp* was not dysregulated in our RNA-seq analysis, the analysis was carried out under non-stressed conditions, leaving open the possibility that under conditions of DNA damage *Hjurp* might be dysregulated by the HNF4α isoforms.

The protein interaction results also suggest that HNF4α isoforms are capable of preferential interactions with circadian rhythm components as demonstrated by the unique interactions of P2-HNF4α with both BMAL (*Arntl*) and CLOCK (*Clock*). While we did not see complete dysregulation of the circadian pathway in the α7HMZ mice, we do find many perturbations of the timing of the cyclic expression. *Arntl* and *Clock* expression look as if they are cycling slightly faster or earlier in α7HMZ mice compared to WT. We also find significant changes in *Cry1*, *Per3*, and *Rorc*, which are known to be involved the cyclic maintenance of circadian rhythms. Several clock-controlled genes (*Mdr1*, *Oat*) are dysregulated between α7HMZ and WT livers as well, perhaps because of these changes. Noting the subtle changes circadian gene expression, interaction with BMAL and CLOCK, and fewer differentially expressed genes between all time points for α7HMZ mice might explain why there is a clear distinction between the three WT time points in the sample distance matrix but this distinction is less clear for the α7HMZ mice.

Since HNF4α plays a major role in the regulation of energy homeostasis in the liver and its natural ligand, linoleic acid, is an essential fatty acid that can only be obtained through the diet, we decided to determine whether these HNF4α isoforms would respond differently to the stress of fasting. What we found was that while both WT and α7HMZ livers showed a similar number of genes being up- and down-regulated in response to fasting, there seems to be a much larger difference in the transcriptomes of these two livers. In the normal fed state (n=3, 10:30am) when comparing expression levels between α7HMZ and WT we found 166 and 291 genes up- and down-regulated, respectively. These data would suggest that the loss of the AF-1 domain in the P2-HNF4α N-terminal domain may lead to a loss of activation of many P1-specific genes but retain the capability to repress negative HNF4α targets. However, in a fasted state (n=3, 10:30am) we found 458 and 382 genes up- and down-regulated respectively in α7HMZ when compared to WT levels of expression. Many of these genes are also up- and down-regulated in the fed state, but when we examine only the genes whose expression levels were significantly dysregulated in response to fasting we find hundreds of genes differentially regulated. When considering genes that are normally up-regulated in response to fasting, the α7HMZ livers have lost the ability to properly regulate 195 of these genes and gained the ability to increase 141. Similarly, α7HMZ livers have lost the ability to repress 204 fasting response genes and gained the ability to repress 116 others. The GTT results show the physiological effects of many of these changes resulting in potentially dysregulated glucose metabolism or insulin resistance from chronic dexamethasone treatments.

The loss of the expression of NR CAR could be playing a role in these changes. CAR is known to be important for the regulation of energy homeostasis and fasting typically induces CAR expression, while CAR-deficient mice do not respond to fasting well (Maglich et al., 2004). It is known that fatty acids are natural ligands for PPARα and a fasting-induced increase in fatty acid levels would increase PPARα activity (Nakamura et al., 2004, 2014). It was also shown that induction of CAR by fasting was ablated by PPARα knockout mice. It should be noted that RIME results showed unique protein-protein interactions with PPARα for P1-HNF4α suggesting that P2-HNF4α's inability to interact with PPARα may lead to a loss of activation of CAR in both fed and fasted states.

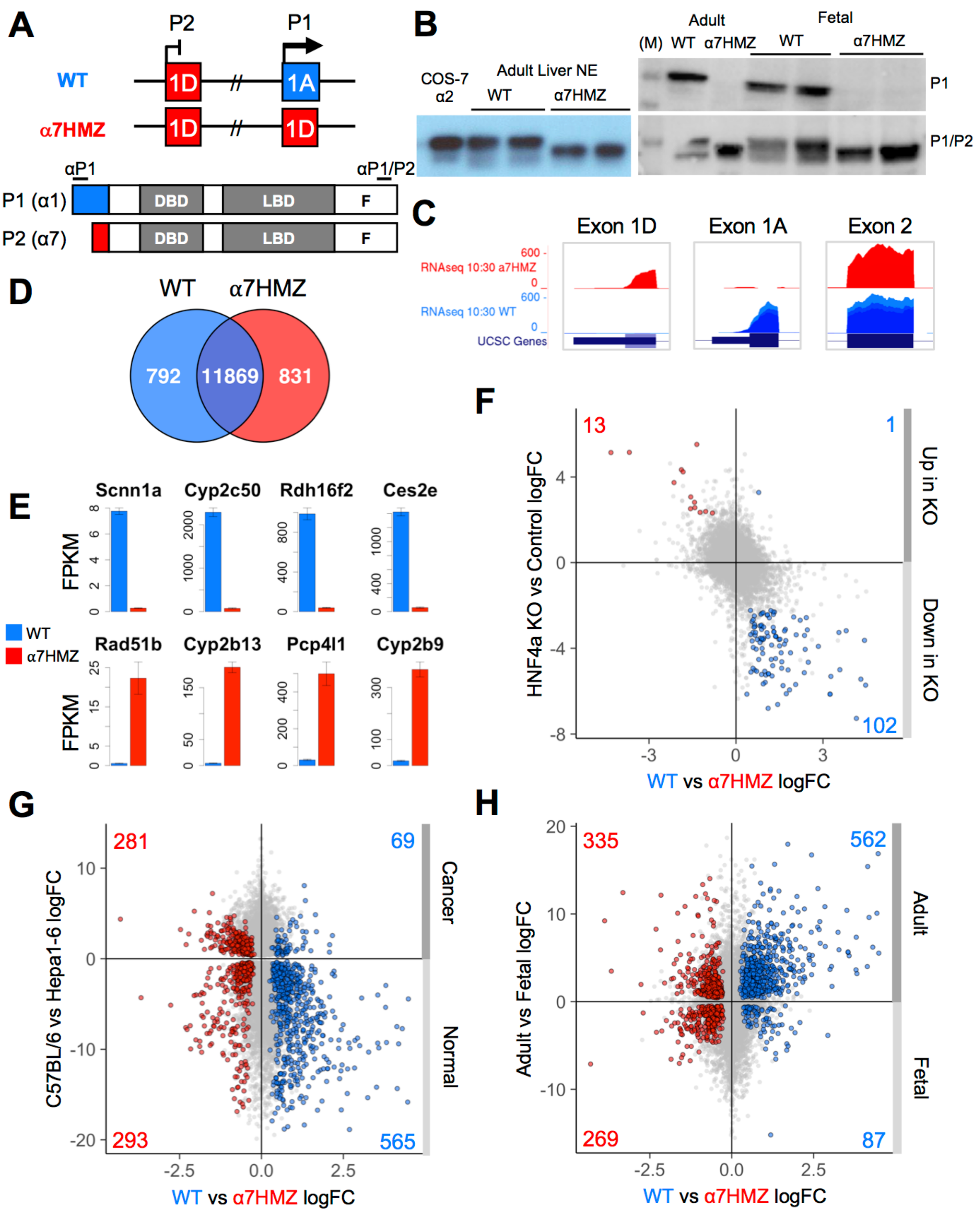
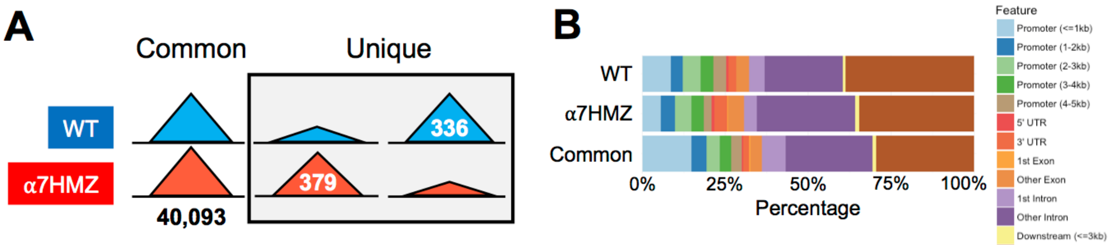

A
P2
P1
WT
1D
1A
α7HMZ
1D
1D
αP1
αP1/P2
P1 (α1)
DBD
LBD
F
P2 (α7)
DBD
LBD
F
B
COS-7
α2
Adult Liver NE
WT
α7HMZ
Adult
Fetal
(M)
WT
α7HMZ
WT
α7HMZ
P1
P1/P2
C
Exon 1D
Exon 1A
Exon 2
600
RNAseq 10:30 a7HMZ
0
600
RNAseq 10:30 WT
0
UCSC Genes
D
WT
α7HMZ
792
11869
831
E
Scnn1a
Cyp2c50
Rdh16f2
Ces2e
FPKM
WT
α7HMZ
Rad51b
Cyp2b13
Pcp4l1
Cyp2b9
FPKM
F
13
1
102
Up in KO
Down in KO
HNF4a KO vs Control logFC
WT vs α7HMZ logFC
G
281
69
293
565
Cancer
Normal
C57BL/6 vs Hepa1-6 logFC
WT vs α7HMZ logFC
H
335
562
269
87
Adult
Fetal
Adult vs Fetal logFC
WT vs α7HMZ logFC

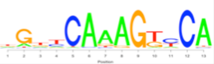
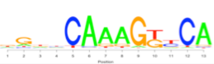
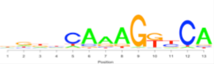
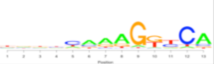
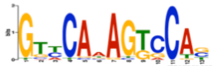
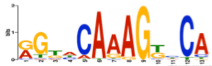
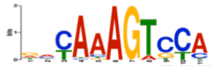
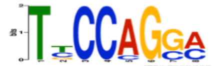
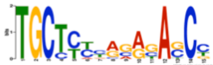
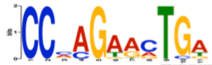
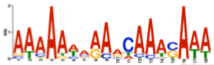
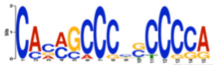
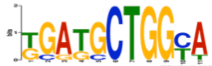
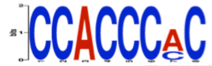
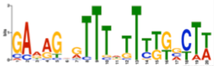
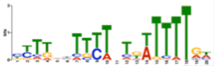
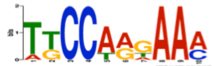
Figure 2.1

Figure 2.1 HNF4α isoforms preferentially regulate genes in fetal liver and liver cancer.

(*A*) *Top:* Diagram of *Hnf4a* P1 and P2 promoters and first exons in wildtype (WT) and α7HMZ exon-swapped mice. *Bottom:* Protein domains of P1-HNF4α (HNF4α1/2) and P2-HNF4α (HNF4α7/8) with the A/B domain color-coded to exons 1A and 1D. Other domains (DNA binding domain, DBD; ligand binding domain, LBD; and F domain) are not changed. Epitopes for P1-specific and P1/P2-common antibodies are indicated. (*B*) Immunoblots of adult liver nuclear extracts (NE) and fetal liver whole cell extracts (WCE) using P1 and P1/P2 antibodies as indicated. M, molecular weight markers, top band is 54 kD. Cos7 α2, NE of Cos7 cells transfected with human HNFα2. (*C*) UCSC Genome Browser view of RNA-seq data from WT and α7HMZ livers showing unique reads in Exons 1A and 1D in each genotype. (*D*) Venn diagram of number of genes from WT and α7HMZ adult male livers at 10:30am. Uniquely expressed genes have an adjusted P-value (padj.) ≤ 0.01. (*E*) FPKM barplots of the most up- and down-regulated genes in α7HMZ livers compared to WT. (*F*) Scatterplot of RNA-seq log2 fold-change (log2FC) values between WT and α7HMZ livers, plotted against microarray log2FC values between HNF4α knockout (KO) and control mouse liver. Colored data points with padj ≤ 0.01 in both datasets. Blue dots, up in WT versus α7HMZ. Red dots, up in α7HMZ versus WT. (*G*) As in (*F*) except that the WT and α7HMZ RNA-seq data are plotted versus RNA-seq data from a murine hepatoma cell line (Hepa1-6) and WT C57BL/6 livers. Numbers indicate number of colored dots (genes) in each quadrant. (*H*) As in (*F*) except that the RNA-seq data from WT and α7HMZ livers are plotted versus adult and E14.5 fetal mouse livers.



HNF4α SVM Score

A WT (336)	≥ 2	2 – 1.75	1.75 – 1.5	1.5 – 1.25
# Peaks	41 (12.2%)	68 (20.2%)	124 (36.9%)	97 (28.8%)
SVM Motif				
MEME Motifs #1				
#2				
#3				
#4				
Transcription Factor	NRs (1) N/A (2) GCM (3)	NRs (1) N/A (2) GLI2 (3) STAT (4)	NRs (1) FOX (2,3)	BCL6/STAT (1) SP1/2 (2) FUBP1 (3)

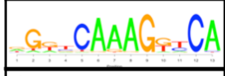
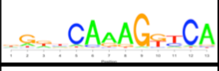
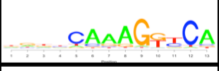
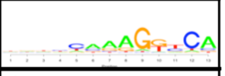
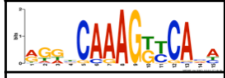
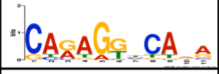
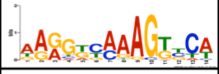
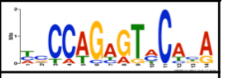
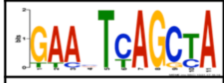
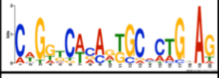
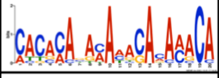
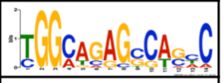
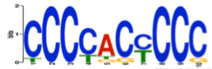
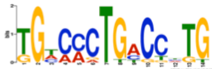
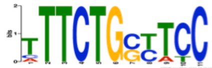
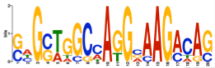
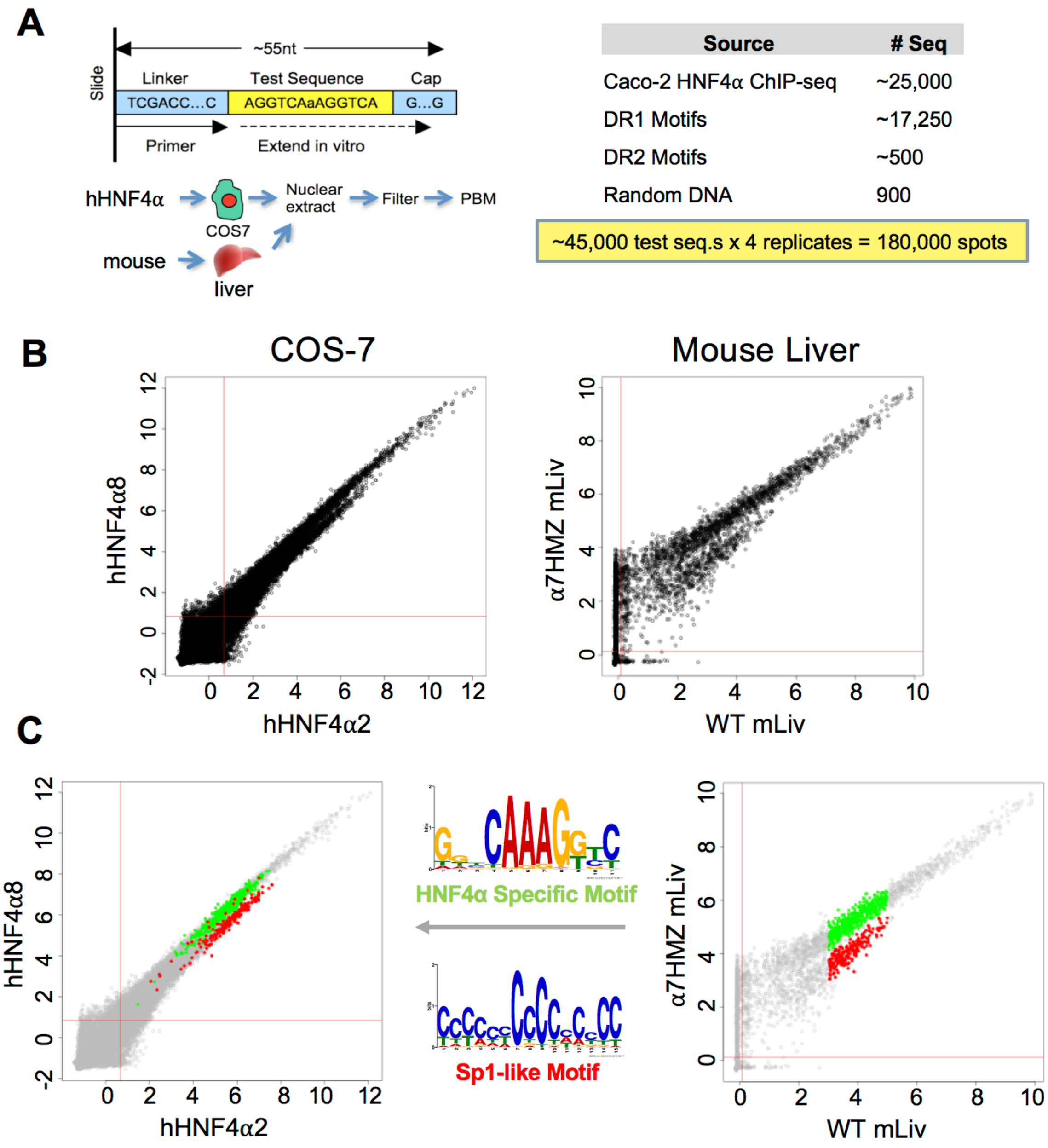
B α7HMZ (379)	≥ 2	2 – 1.75	1.75 – 1.5	1.5 – 1.25
# Peaks	59 (15.5%)	80 (21.1%)	156 (41.1%)	82 (21.6%)
SVM Motif				
MEME Motifs #1				
#2				
#3				
#4				
Transcription Factor	NRs (1) N/A (2)	COUPTF (1) ESR (2) SP1/2 (3) UBIP1 (4)	NRs (1) FOX (2) NR6A1 (3) SMAD3 (4)	FOXA (1) PAX5 (2)

Figure 2.2

Figure 2.2 HNF4α isoforms exhibit similar but not identical chromatin binding profiles *in vivo*.

(*A*) HNF4α ChIP-seq peaks were categorized as common (40,093), WT unique (336) or α7HMZ unique (379) peaks. (*B*) Feature distribution plots for common, WT, and α7HMZ unique peaks as determined by ChIPseeker. (*C*) Categorization of WT unique ChIP peaks, based on highest SVM HNF4α motif score, into four groups. The top four DNA motifs from *de novo* MEME-ChIP analysis are shown along with the transcription factor family known to bind the motifs. Numbers refer to the MEME motif. Transcription factors in bold are discussed in the text. (*D*) As in (*C*) except for α7HMZ unique ChIP peaks.



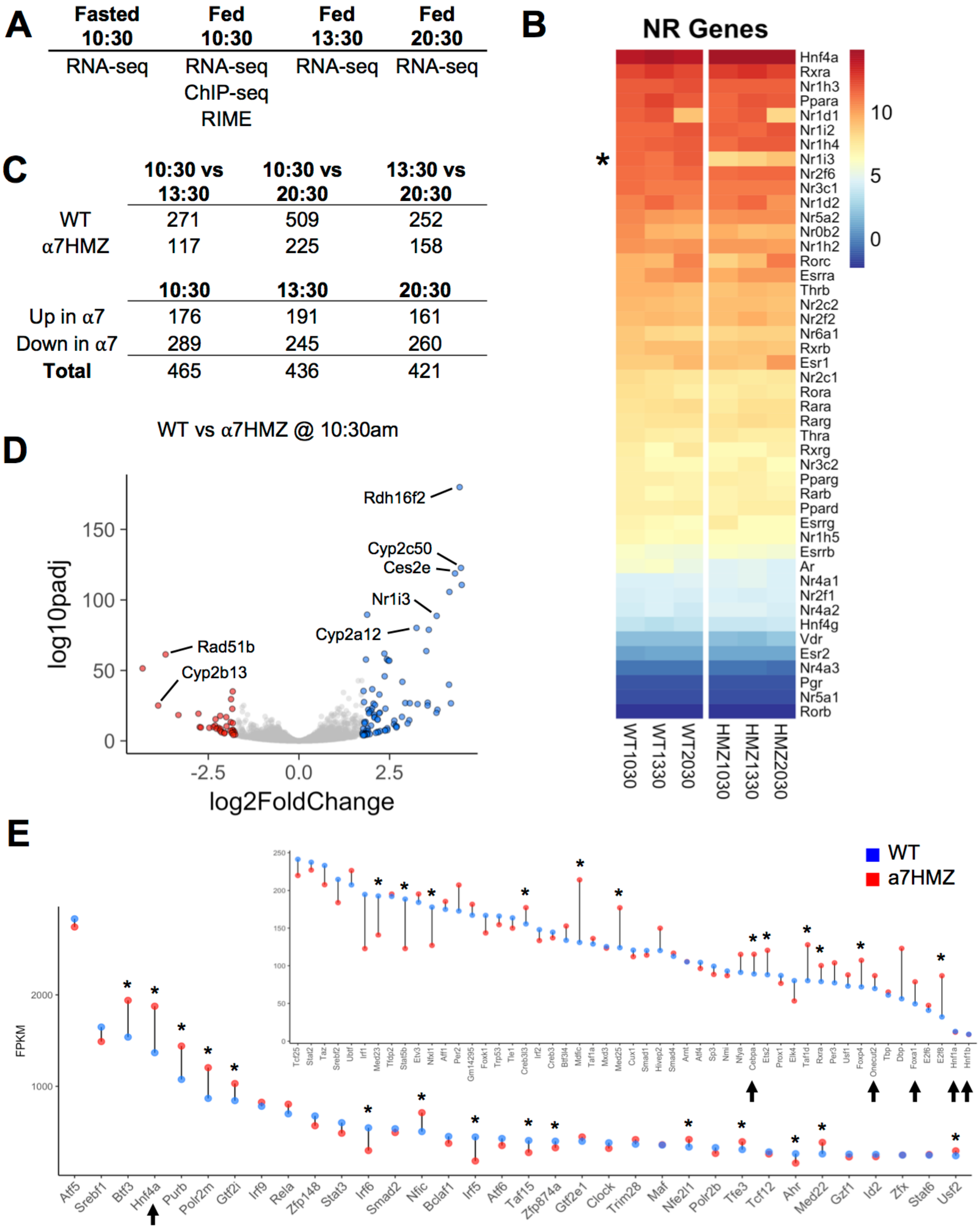
Source	# Seq
Caco-2 HNF4α ChIP-seq	~25,000
DR1 Motifs	~17,250
DR2 Motifs	~500
Random DNA	900

~45,000 test seq.s x 4 replicates = 180,000 spots

Figure 2.3

Figure 2.3 HNF4α isoforms exhibit similar DNA binding profiles *in vitro* but differ in binding Sp1-like motifs.

(*A*) Schematic of the protein binding microarray (PBM). Arrays are extended *in vitro* and nuclear extracts from COS-7 cells transfected with human HNF4α or mouse livers were applied to slides. containing ~45,000 test sequences data-mined from HNF4α ChIP-seq peak centers from a colorectal adenocarcinoma cell line (Caco-2) as well as DR1 and DR2 motif permutations and 900 random DNA test spots. All sequences were printed in quadruplicate on the slide for a total of 180,000 spots. (*B*) Scatter plots of log2 average binding intensities for each test sequence with ectopically expressed human HNF4α8 plotted against ectopically expressed human HNF4α2, in COS-7 cells and mouse liver nuclear extracts from α7HMZ mice plotted against WT mice. (*C*) *Right*, two groups, indicated in green and red, were selected from the α7HMZ vs WT mouse liver plot for *de novo* motif analysis. *Middle*, PWMs for the red (Sp1-like) and green (HNF4α) groups. *Left,* The red and green groups of sequences are highlighted in the COS-7 plots.



A

Fasted 10:30	Fed 10:30	Fed 13:30	Fed 20:30
RNA-seq	RNA-seq ChIP-seq RIME	RNA-seq	RNA-seq

C

	10:30 vs 13:30	10:30 vs 20:30	13:30 vs 20:30
WT	271	509	252
α7HMZ	117	225	158

	10:30	13:30	20:30
Up in α7	176	191	161
Down in α7	289	245	260
Total	465	436	421

Figure 2.4

Figure 2.4 HNF4α isoforms result in different transcriptional profiles: HNF4α is the most highly expressed NR in the liver.

(*A*) Overview of analyses performed on WT and α7HMZ livers with the time of day the livers were harvested. (*B*) Heatmap of regularized log-transformed read counts for all NR genes sorted from highest to lowest for the 10:30 AM fed WT mice. (*C*) *Top*, number of genes with significant gene expression changes between the indicated time points. *Bottom*, number of genes significantly up- or down-regulated in α7HMZ livers compared to WT at each time point. Significance measured by padj ≤ 0.01 and log2FC ≥ 1. (*D*) Volcano plot of gene expression between WT and α7HMZ at the 10:30 AM fed time point. Spots representing genes with log2FC ≥ 1.75 are in bold. Select genes are indicated. (*E*) Cleveland plots of WT (blue) and α7HMZ (red) FPKM values for the top 85 expressed transcription factors in WT fed livers at 10:30 AM. Arrows point to known liver-enriched transcription factors. Significant gene expression differences with padj ≤ 0.01 between WT and α7HMZ are marked with asterisks.

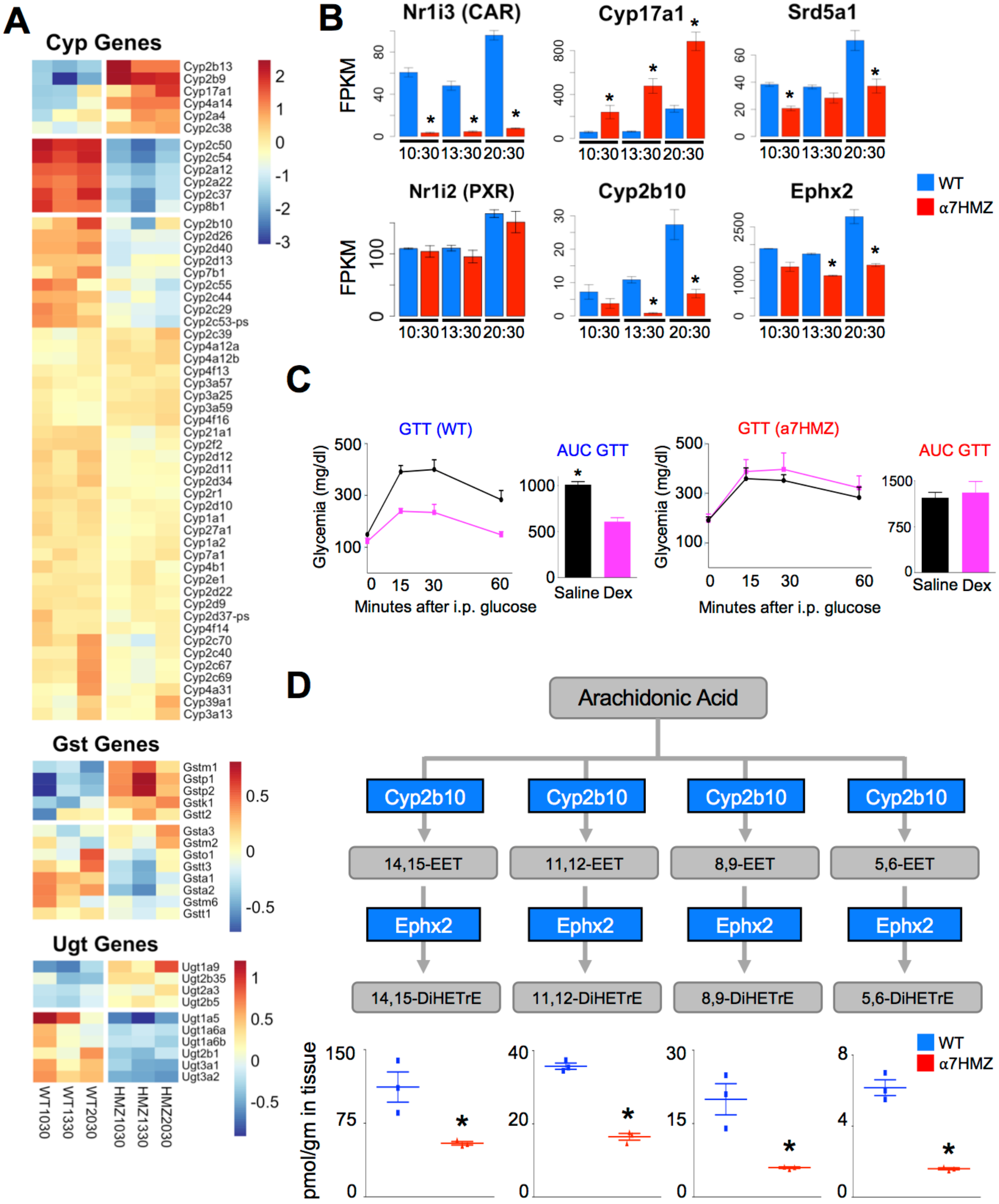
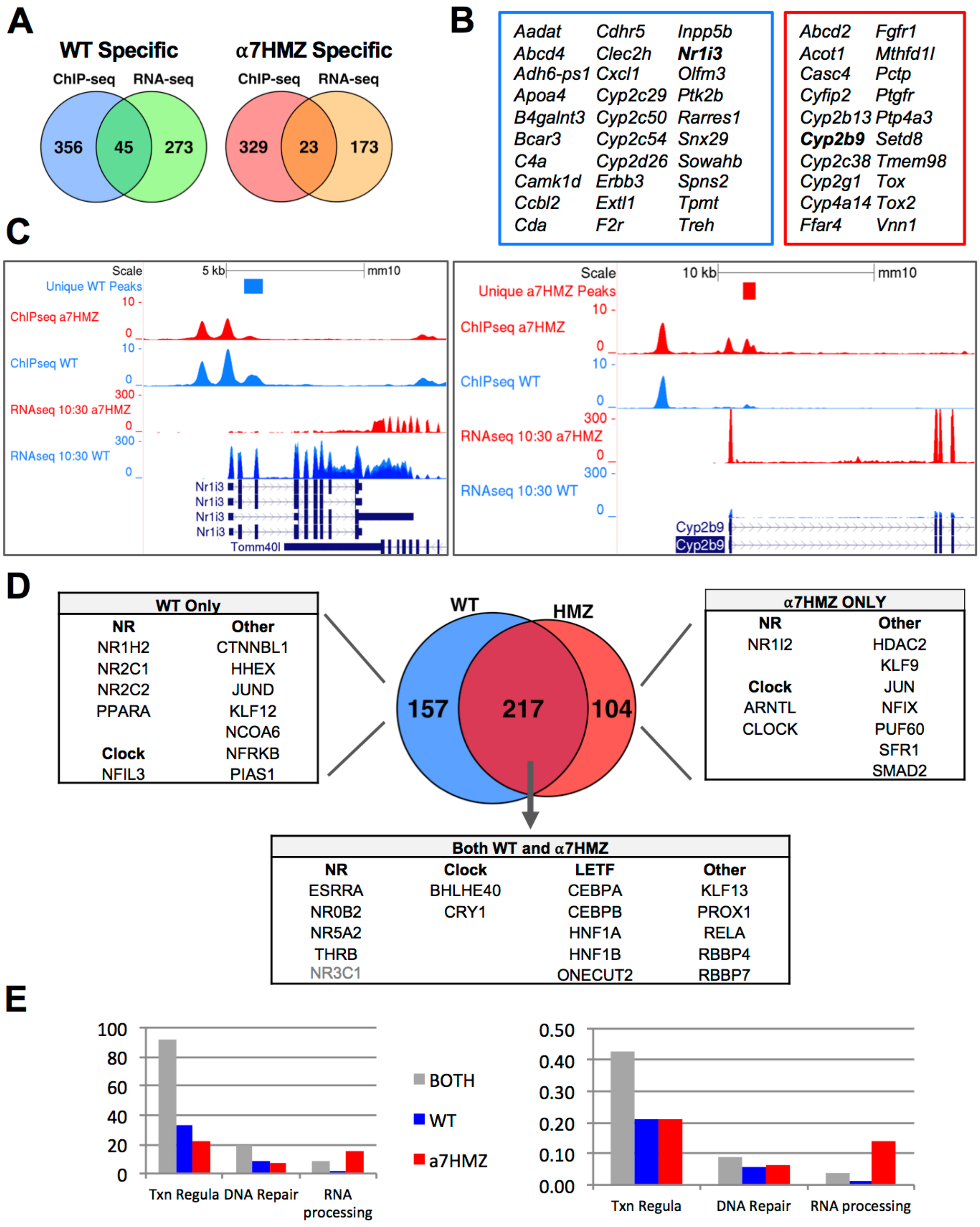

A
Cyp Genes
Cyp2b13
Cyp2b9
Cyp17a1
Cyp4a14
Cyp2a4
Cyp2c38
Cyp2c50
Cyp2c54
Cyp2a12
Cyp2a22
Cyp2c37
Cyp8b1
Cyp2b10
Cyp2d26
Cyp2d40
Cyp2d13
Cyp7b1
Cyp2c55
Cyp2c44
Cyp2c29
Cyp2c53-ps
Cyp2c39
Cyp4a12a
Cyp4a12b
Cyp4f13
Cyp3a57
Cyp3a25
Cyp3a59
Cyp4f16
Cyp21a1
Cyp2f2
Cyp2d12
Cyp2d11
Cyp2d34
Cyp2r1
Cyp2d10
Cyp1a1
Cyp27a1
Cyp1a2
Cyp7a1
Cyp4b1
Cyp2e1
Cyp2d22
Cyp2d9
Cyp2d37-ps
Cyp4f14
Cyp2c70
Cyp2c40
Cyp2c67
Cyp2c69
Cyp4a31
Cyp39a1
Cyp3a13
2
1
0
-1
-2
-3
Gst Genes
Gstm1
Gstp1
Gstp2
Gstk1
Gstt2
Gsta3
Gstm2
Gsto1
Gstt3
Gsta1
Gsta2
Gstm6
Gstt1
0.5
0
-0.5
Ugt Genes
Ugt1a9
Ugt2b35
Ugt2a3
Ugt2b5
Ugt1a5
Ugt1a6a
Ugt1a6b
Ugt2b1
Ugt3a1
Ugt3a2
1
0.5
0
-0.5
WT1030
WT1330
WT2030
HMZ1030
HMZ1330
HMZ2030
B
Nr1i3 (CAR)
Cyp17a1
Srd5a1
FPKM
80
40
0
800
400
0
60
40
20
0
10:30
13:30
20:30
*
Nr1i2 (PXR)
Cyp2b10
Ephx2
FPKM
100
0
30
20
10
0
2500
1000
0
WT
α7HMZ
C
GTT (WT)
Glycemia (mg/dl)
500
300
100
0
15
30
60
Minutes after i.p. glucose
AUC GTT
1000
500
0
Saline Dex
GTT (a7HMZ)
500
300
100
AUC GTT
1500
750
0
D
Arachidonic Acid
Cyp2b10
14,15-EET
11,12-EET
8,9-EET
5,6-EET
Ephx2
14,15-DiHETrE
11,12-DiHETrE
8,9-DiHETrE
5,6-DiHETrE
pmol/gm in tissue
150
75
0
40
20
0
30
15
0
8
4
0
WT
α7HMZ

Figure 2.5

Figure 2.5 HNF4α isoforms result in altered metabolic profiles.

(*A*) Heatmaps of regularized log-transformed read counts for Phase I and II detoxification enzymes: select cytochrome P450s (Cyp), glutathione S-transferases (Gst), and UDP glucuronosyltransferases (Ugt) across all three time points in WT and α7HMZ livers. (*B*) Bar plots of FPKM values for NRs CAR (*Nr1i3*) and PXR (*Nr1i2*) and select metabolic genes. * padj ≤ 0.01 between WT and α7HMZ. (*C*) GTT assays and area under the curve (AUC) with intraperitoneal injection (i.p.) of glucose in WT and α7HMZ male mice (n=4-7, 16-24 weeks old) after daily injections with 4 mg/kg body weight dexamethasone (pink) or saline (black) for 7 days. On eighth day mice were fasted for 5 hours prior to GTT. * Significantly greater (P-value ≤ 0.05). (*D*) *Top*, Schematic of arachidonic acid metabolic pathway, shown are enzymes down-regulated in α7HMZ livers, *Cyp2b10* and *Ephx2*. *Bottom*, Levels of DiHETrE oxylipins (dihydroxytrienoic acids) in livers of WT or α7HMZ mice (n=3, 12-13 weeks old). * Significantly different (P-value ≤ 0.05; Ttest).



WT Only	
NR	**Other**
NR1H2	CTNNBL1
NR2C1	HHEX
NR2C2	JUND
PPARA	KLF12
	NCOA6
Clock	NFRKB
NFIL3	PIAS1

α7HMZ ONLY	
NR	**Other**
NR1I2	HDAC2
	KLF9
Clock	JUN
ARNTL	NFIX
CLOCK	PUF60
	SFR1
	SMAD2

Both WT and α7HMZ			
NR	**Clock**	**LETF**	**Other**
ESRRA	BHLHE40	CEBPA	KLF13
NR0B2	CRY1	CEBPB	PROX1
NR5A2		HNF1A	RELA
THRB		HNF1B	RBBP4
NR3C1		ONECUT2	RBBP7

Figure 2.6

Figure 2.6 HNF4α isoforms have unique protein-protein interactions.

(*A*) Venn diagrams showing overlap between genes with uniquely bound ChIP-peaks within a 50-kb window and differentially expressed genes in WT and α7HMZ livers. (*B*) Select genes from the intersection of ChIP-seq and RNA-seq in the Venn diagram. (*C*) UCSC Genome Browser view of differentially expressed genes with a unique ChIP-signal near +1, *Nr1i3* (CAR) and *Cyp2b9*. ChIP-seq and uniquely bound regions are in the top two tracks and RNA-seq from 10:30 AM is in the bottom two tracks. (*D*) Venn diagram summarizing RIME results. All nuclear receptors (NR), clock related proteins, and liver enriched TFs (LETF) are shown with other selected proteins of interest. NR3C1 shown in grey because it did not pass the arbitrary cut-off but is a gene of interest. (*E*) All proteins 10-fold above background categorized into three categories; transcription factors (Txn Regula.), DNA repair, and RNA processing. *Left*, counts of total proteins in WT, α7HMZ, and common groups. *Right*, counts of proteins in each category normalized to total number of proteins in unique and common datasets.

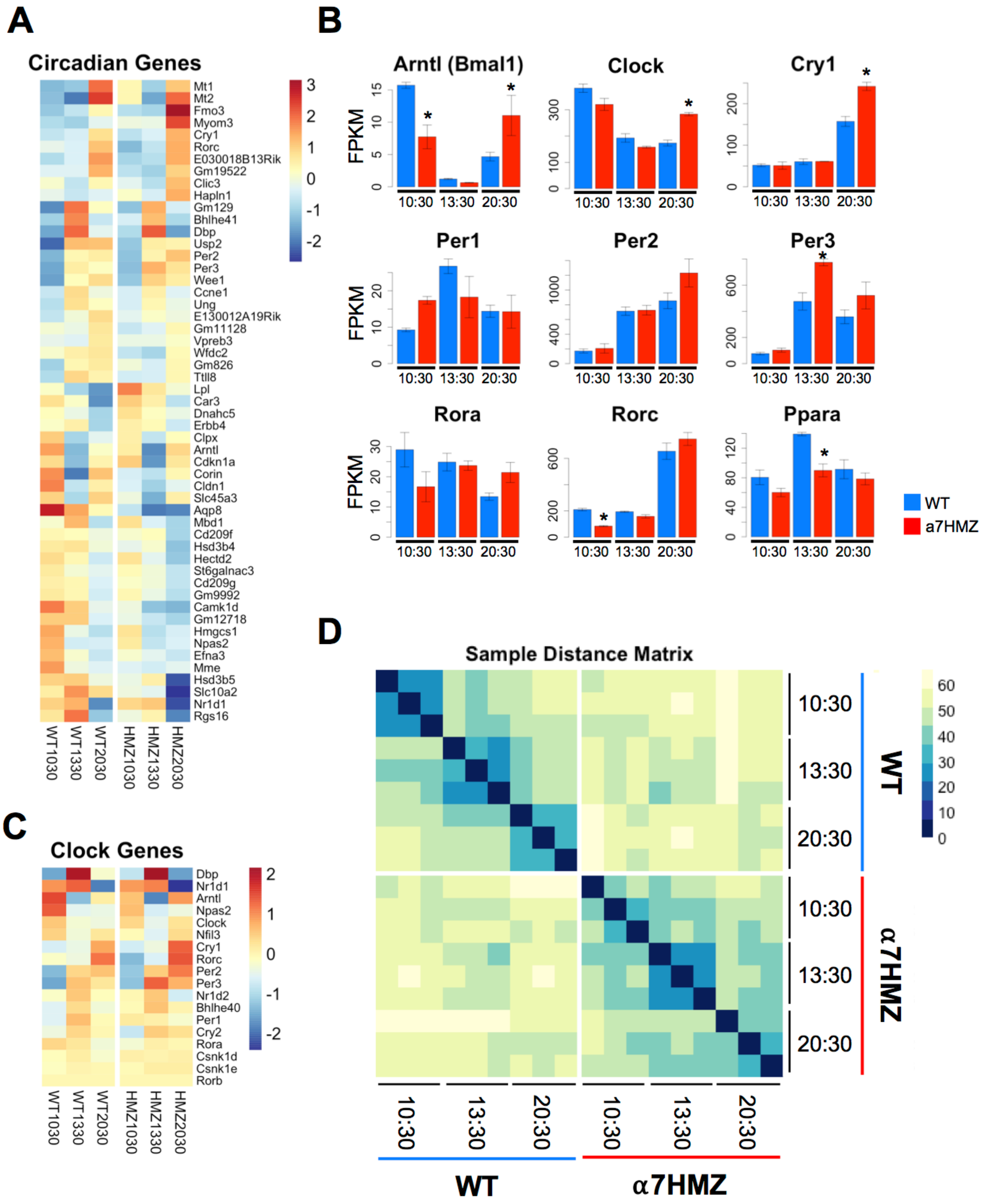
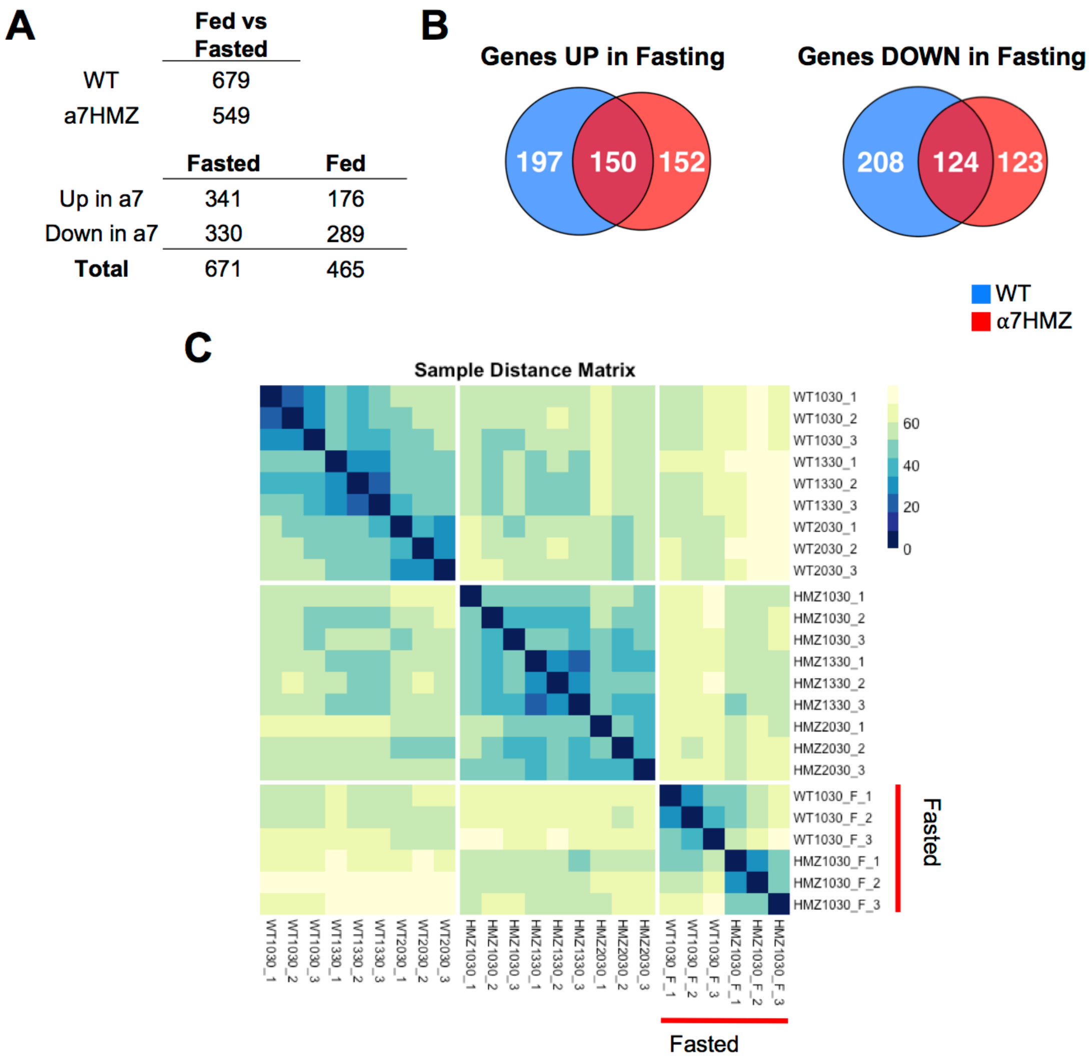

A
Circadian Genes
Mt1
Mt2
Fmo3
Myom3
Cry1
Rorc
E030018B13Rik
Gm19522
Clic3
Hapln1
Gm129
Bhlhe41
Dbp
Usp2
Per2
Per3
Wee1
Ccne1
Ung
E130012A19Rik
Gm11128
Vpreb3
Wfdc2
Gm826
Ttll8
Lpl
Car3
Dnahc5
Erbb4
Clpx
Arntl
Cdkn1a
Corin
Cldn1
Slc45a3
Aqp8
Mbd1
Cd209f
Hsd3b4
Hectd2
St6galnac3
Cd209g
Gm9992
Camk1d
Gm12718
Hmgcs1
Npas2
Efna3
Mme
Hsd3b5
Slc10a2
Nr1d1
Rgs16
3
2
1
0
-1
-2
WT1030
WT1330
WT2030
HMZ1030
HMZ1330
HMZ2030
B
Arntl (Bmal1)
Clock
Cry1
Per1
Per2
Per3
Rora
Rorc
Ppara
FPKM
*
10:30 13:30 20:30
WT
a7HMZ
C
Clock Genes
Dbp
Nr1d1
Arntl
Npas2
Clock
Nfil3
Cry1
Rorc
Per2
Per3
Nr1d2
Bhlhe40
Per1
Cry2
Rora
Csnk1d
Csnk1e
Rorb
2
1
0
-1
-2
WT1030
WT1330
WT2030
HMZ1030
HMZ1330
HMZ2030
D
Sample Distance Matrix
60
50
40
30
20
10
0
10:30
13:30
20:30
WT
α7HMZ

Figure 2.7

Figure 2.7 HNF4α isoforms impact the circadian response.

(*A*) Heatmap of regularized log-transformed read counts for all genes with significant gene expression difference (padj. ≤ 0.01 and log2FC ≥ 2) between any pair of time points for either genotype. (*B*) Barplots for select core circadian clock genes and several metabolic clock-controlled genes. Significant expression changes between genotypes at a given time point are denoted with asterisks (padj. ≤ 0.01). (*C*) Heatmap of regularized log-transformed read counts for core circadian machinery. (*D*) Distance matrix for all fed RNA-seq samples across all three time-points (N=3 per condition). Dark blue indicates smaller distance which implies high degree of similarity.



	Fed vs Fasted
WT	679
a7HMZ	549

	Fasted	Fed
Up in a7	341	176
Down in a7	330	289
Total	671	465

Figure 2.8

Figure 2.8 HNF4α isoforms impact the fasting response.

(*A*) *Top*, total number of genes dysregulated (padj. ≤ 0.01 & log2FC ≥ 1) in WT and α7HMZ livers between the fed and the fasted state. *Bottom*, total number of genes up- and down-regulated (padj. ≤ 0.01 & log2FC ≥ 1) in α7HMZ livers compared to WT in both fed and fasted mice. (*B*) Venn diagrams of total genes up or down (padj. ≤ 0.01 & log2FC ≥ 1) in WT and α7HMZ livers in fasted versus fed mice. (*C*) Distance matrix for all RNA-seq samples, including fasted (n=3 per condition). Dark blue indicates smaller distance which implies high degree of similarity.

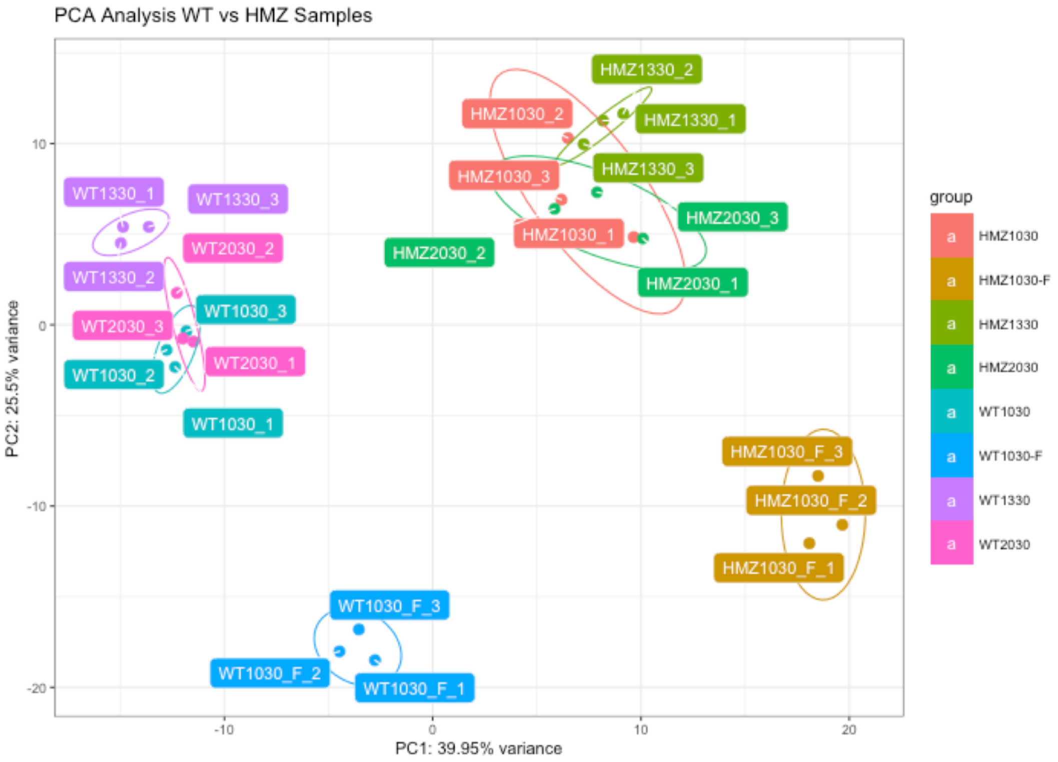

PCA Analysis WT vs HMZ Samples
HMZ1330_2
HMZ1030_2
HMZ1330_1
HMZ1030_3
HMZ1330_3
HMZ2030_3
HMZ2030_2
HMZ1030_1
HMZ2030_1
WT1330_1
WT1330_3
WT2030_2
WT1330_2
WT1030_3
WT2030_3
WT2030_1
WT1030_2
WT1030_1
HMZ1030_F_3
HMZ1030_F_2
HMZ1030_F_1
WT1030_F_3
WT1030_F_2
WT1030_F_1
10
0
-10
-20
-10
0
10
20
PC1: 39.95% variance
PC2: 25.5% variance
group
a HMZ1030
a HMZ1030-F
a HMZ1330
a HMZ2030
a WT1030
a WT1030-F
a WT1330
a WT2030

Figure 2.S1

Figure 2.S1 PCA Analysis of WT vs α7HMZ Samples

PCA analysis of rlog transformed read counts from DESeq2 for all WT and α7HMZ samples, including fed and fasted (-F).

References

Bembom O. 2017. seqLogo: Sequence logos for DNA sequence alignments. R package version 1.40.0.

Baek, D., Davis, C., Ewing, B., Gordon, D., and Green, P. (2007). Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. Genome Res. *17*, 145–155.

Battle, M.A., Konopka, G., Parviz, F., Gaggl, A.L., Yang, C., Sladek, F.M., and Duncan, S.A. (2006). Hepatocyte nuclear factor 4alpha orchestrates expression of cell adhesion proteins during the epithelial transformation of the developing liver. Proc. Natl. Acad. Sci. U. S. A. *103*, 8419–8424.

Bolotin, E., Liao, H., Ta, T.C., Yang, C., Hwang-Verslues, W., Evans, J.R., Jiang, T., and Sladek, F.M. (2010). Integrated approach for the identification of human hepatocyte nuclear factor 4alpha target genes using protein binding microarrays. Hepatology *51*, 642–653.

Briançon, N., and Weiss, M.C. (2006). In vivo role of the HNF4α AF-1 activation domain revealed by exon swapping. EMBO J. *25*, 1253–1262.

Briançon, N., Bailly, A., Clotman, F., Jacquemin, P., Lemaigre, F.P., and Weiss, M.C. (2004). Expression of the alpha7 isoform of hepatocyte nuclear factor (HNF) 4 is activated by HNF6/OC-2 and HNF1 and repressed by HNF4alpha1 in the liver. J. Biol. Chem. *279*, 33398–33408.

Cai, S.-H., Lu, S.-X., Liu, L.-L., Zhang, C.Z., and Yun, J.-P. (2017). Increased expression of hepatocyte nuclear factor 4 alpha transcribed by promoter 2 indicates a poor prognosis in hepatocellular carcinoma. Therap. Adv. Gastroenterol. *10*, 761–771.

Chen, F., Zamule, S.M., Coslo, D.M., Chen, T., and Omiecinski, C.J. (2013). The human constitutive androstane receptor promotes the differentiation and maturation of hepatic-like cells. Dev. Biol. *384*, 155–165.

Cui, J.Y., Renaud, H.J., and Klaassen, C.D. (2012). Ontogeny of novel cytochrome P450 gene isoforms during postnatal liver maturation in mice. Drug Metab. Dispos. *40*, 1226–1237.

Dean, S., Tang, J.I., Seckl, J.R., and Nyirenda, M.J. (2010). Developmental and tissue-specific regulation of hepatocyte nuclear factor 4-alpha (HNF4-alpha) isoforms in rodents. Gene Expr. *14*, 337–344.

Deol, P., Evans, J.R., Dhahbi, J., Chellappa, K., Han, D.S., Spindler, S., and Sladek, F.M. (2015). Soybean Oil Is More Obesogenic and Diabetogenic than Coconut Oil and Fructose in Mouse: Potential Role for the Liver. PLoS One *10*, e0132672.

Dhir, R.N., Dworakowski, W., Thangavel, C., and Shapiro, B.H. (2006). Sexually dimorphic regulation of hepatic isoforms of human cytochrome p450 by growth hormone. J. Pharmacol. Exp. Ther. *316*, 87–94.

Fajans, S.S., Bell, G.I., and Polonsky, K.S. (2001). Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. N. Engl. J. Med. *345*, 971–980.

Fang, B., Mane-Padros, D., Bolotin, E., Jiang, T., and Sladek, F.M. (2012). Identification of a binding motif specific to HNF4 by comparative analysis of multiple nuclear receptors. Nucleic Acids Res. *40*, 5343–5356.

Fang, X., Hu, S., Xu, B., Snyder, G.D., Harmon, S., Yao, J., Liu, Y., Sangras, B., Falck, J.R., Weintraub, N.L., et al. (2006). 14,15-Dihydroxyeicosatrienoic acid activates peroxisome proliferator-activated receptor-alpha. Am. J. Physiol. Heart Circ. Physiol. *290*, H55–H63.

Filipescu, D., Naughtin, M., Podsypanina, K., Lejour, V., Wilson, L., Gurard-Levin, Z.A., Orsi, G.A., Simeonova, I., Toufektchan, E., Attardi, L.D., et al. (2017). Essential role for centromeric factors following p53 loss and oncogenic transformation. Genes Dev. *31*, 463–480.

Harries, L.W., Locke, J.M., Shields, B., Hanley, N.A., Hanley, K.P., Steele, A., Njølstad, P.R., Ellard, S., and Hattersley, A.T. (2008). The diabetic phenotype in HNF4A mutation carriers is moderated by the expression of HNF4A isoforms from the P1 promoter during fetal development. Diabetes *57*, 1745–1752.

Hart, S.N., Cui, Y., Klaassen, C.D., and Zhong, X.-B. (2009). Three patterns of cytochrome P450 gene expression during liver maturation in mice. Drug Metab. Dispos. *37*, 116–121.

Hartman, M.L., Veldhuis, J.D., and Thorner, M.O. (1993). Normal control of growth hormone secretion. Horm. Res. *40*, 37–47.

Hatziapostolou, M., Polytarchou, C., Aggelidou, E., Drakaki, A., Poultsides, G.A., Jaeger, S.A., Ogata, H., Karin, M., Struhl, K., Hadzopoulou-Cladaras, M., et al. (2011). An HNF4α-miRNA inflammatory feedback circuit regulates hepatocellular oncogenesis. Cell *147*, 1233–1247.

Hayhurst, G.P., Lee, Y.H., Lambert, G., Ward, J.M., and Gonzalez, F.J. (2001). Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. Mol. Cell. Biol. *21*, 1393–1403.

Hu, Z., Huang, G., Sadanandam, A., Gu, S., Lenburg, M.E., Pai, M., Bayani, N., Blakely, E.A., Gray, J.W., and Mao, J.-H. (2010). The expression level of HJURP has an independent prognostic impact and predicts the sensitivity to radiotherapy in breast cancer. Breast Cancer Res. *12*, R18.

Jiang, G., Nepomuceno, L., Hopkins, K., and Sladek, F.M. (1995). Exclusive homodimerization of the orphan receptor hepatocyte nuclear factor 4 defines a new subclass of nuclear receptors. Mol. Cell. Biol. *15*, 5131–5143.

Lazarevich, N.L., Cheremnova, O.A., Varga, E.V., Ovchinnikov, D.A., Kudrjavtseva, E.I., Morozova, O.V., Fleishman, D.I., Engelhardt, N.V., and Duncan, S.A. (2004). Progression of HCC in mice is associated with a downregulation in the expression of hepatocyte nuclear factors. Hepatology *39*, 1038–1047.

Lee, J.S., Ward, W.O., Knapp, G., Ren, H., Vallanat, B., Abbott, B., Ho, K., Karp, S.J., and Corton, J.C. (2012). Transcriptional ontogeny of the developing liver. BMC Genomics *13*, 33.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696–1697.

Maglich, J.M., Watson, J., McMillen, P.J., Goodwin, B., Willson, T.M., and Moore, J.T. (2004). The nuclear receptor CAR is a regulator of thyroid hormone metabolism during caloric restriction. J. Biol. Chem. *279*, 19832–19838.

di Masi, A., De Marinis, E., Ascenzi, P., and Marino, M. (2009). Nuclear receptors CAR and PXR: Molecular, functional, and biomedical aspects. Mol. Aspects Med. *30*, 297–343.

Matyash, V., Liebisch, G., Kurzchalia, T.V., Shevchenko, A., and Schwudke, D. (2008). Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. J. Lipid Res. *49*, 1137–1146.

Mohammed, H., Taylor, C., Brown, G.D., Papachristou, E.K., Carroll, J.S., and D'Santos, C.S. (2016). Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes. Nat. Protoc. *11*, 316–326.

Montes de Oca, R., Gurard-Levin, Z.A., Berger, F., Rehman, H., Martel, E., Corpet, A., de Koning, L., Vassias, I., Wilson, L.O.W., Meseure, D., et al. (2015). The histone chaperone HJURP is a new independent prognostic marker for luminal A breast carcinoma. Mol. Oncol. *9*, 657–674.

Mugford, C.A., and Kedderis, G.L. (1998). Sex-dependent metabolism of xenobiotics. Drug Metab. Rev. *30*, 441–498.

Nakamura, M.T., Cheon, Y., Li, Y., and Nara, T.Y. (2004). Mechanisms of regulation of gene expression by fatty acids. Lipids *39*, 1077–1083.

Nakamura, M.T., Yudell, B.E., and Loor, J.J. (2014). Regulation of energy metabolism by long-chain fatty acids. Prog. Lipid Res. *53*, 124–144.

Nakhei, H., Lingott, A., Lemm, I., and Ryffel, G.U. (1998). An alternative splice variant of the tissue specific transcription factor HNF4alpha predominates in undifferentiated murine cell types. Nucleic Acids Res. *26*, 497–504.

Naro, C., Bielli, P., Pagliarini, V., and Sette, C. (2015). The interplay between DNA damage response and RNA processing: the unexpected role of splicing factors as gatekeepers of genome stability. Front. Genet. *6*, 142.

Ng, V.Y., Huang, Y., Reddy, L.M., Falck, J.R., Lin, E.T., and Kroetz, D.L. (2007). Cytochrome P450 Eicosanoids are Activators of Peroxisome Proliferator-Activated Receptor. Drug Metab. Dispos. *35*, 1126–1134.

Ning, B.-F., Ding, J., Yin, C., Zhong, W., Wu, K., Zeng, X., Yang, W., Chen, Y.-X., Zhang, J.-P., Zhang, X., et al. (2010). Hepatocyte nuclear factor 4 alpha suppresses the development of hepatocellular carcinoma. Cancer Res. *70*, 7640–7651.

Pal, S., Gupta, R., Kim, H., Wickramasinghe, P., Baubet, V., Showe, L.C., Dahmane, N., and Davuluri, R.V. (2011). Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. Genome Res. *21*, 1260–1272.

Rudolph, K.L.M., Schmitt, B.M., Villar, D., White, R.J., Marioni, J.C., Kutter, C., and Odom, D.T. (2016). Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. PLoS Genet. *12*, e1006024.

Schmutz, I., Ripperger, J.A., Baeriswyl-Aebischer, S., and Albrecht, U. (2010). The mammalian clock component PERIOD2 coordinates circadian output by interaction with nuclear receptors. Genes Dev. *24*, 345–357.

Sladek, F.M., Zhong, W.M., Lai, E., and Darnell, J.E., Jr (1990). Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. Genes Dev. *4*, 2353–2365.

Tahara, Y., and Shibata, S. (2016). Circadian rhythms of liver physiology and disease: experimental and clinical evidence. Nat. Rev. Gastroenterol. Hepatol. *13*, 217–226.

Tanaka, T., Jiang, S., Hotta, H., Takano, K., Iwanari, H., Sumi, K., Daigo, K., Ohashi, R., Sugai, M., Ikegame, C., et al. (2006). Dysregulated expression of P1 and P2 promoter-driven hepatocyte nuclear factor-4alpha in the pathogenesis of human cancer. J. Pathol. *208*, 662–672.

Tolson, A.H., and Wang, H. (2010). Regulation of drug-metabolizing enzymes by xenobiotic receptors: PXR and CAR. Adv. Drug Deliv. Rev. *62*, 1238–1249.

Torres-Padilla, M.E., Fougere-Deschatrette, C., and Weiss, M.C. (2001). Expression of HNF4a isoforms in mouse liver development is regulated by sequential promoter usage and constitutive 3 end splicing. Mech. Dev. *109*, 183–193.

Verzi, M.P., Shin, H., He, H.H., Sulahian, R., Meyer, C.A., Montgomery, R.K., Fleet, J.C., Brown, M., Liu, X.S., and Shivdasani, R.A. (2010). Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. Dev. Cell *19*, 713–726.

Vuong, L.M., Chellappa, K., Dhahbi, J.M., Deans, J.R., Fang, B., Bolotin, E., Titova, N.V., Hoverter, N.P., Spindler, S.R., Waterman, M.L., et al. (2015). Differential Effects of Hepatocyte Nuclear Factor 4α Isoforms on Tumor Growth and T-Cell Factor 4/AP-1 Interactions in Human Colorectal Cancer Cells. Mol. Cell. Biol. *35*, 3471–3490.

Wagner, K., Inceoglu, B., and Hammock, B.D. (2011). Soluble epoxide hydrolase inhibition, epoxygenated fatty acids and nociception. Prostaglandins Other Lipid Mediat. *96*, 76–83.

Walesky, C., and Apte, U. (2015). Role of hepatocyte nuclear factor 4α (HNF4α) in cell proliferation and cancer. Gene Expr. *16*, 101–108.

Walesky, C., Gunewardena, S., Terwilliger, E.F., Edwards, G., Borude, P., and Apte, U. (2013). Hepatocyte-specific deletion of hepatocyte nuclear factor-4α in adult mice results in increased hepatocyte proliferation. Am. J. Physiol. Gastrointest. Liver Physiol. *304*, G26–G37.

Wickramasinghe, V.O., and Venkitaraman, A.R. (2016). RNA Processing and Genome Stability: Cause and Consequence. Mol. Cell *61*, 496–505.

Willson, T.M., and Kliewer, S.A. (2002). PXR, CAR and drug metabolism. Nat. Rev. Drug Discov. *1*, 259–266.

Wolbold, R., Klein, K., Burk, O., Nüssler, A.K., Neuhaus, P., Eichelbaum, M., Schwab, M., and Zanger, U.M. (2003). Sex is a major determinant of CYP3A4 expression in human liver. Hepatology *38*, 978–988.

Yamagata, K., Furuta, H., Oda, N., Kaisaki, P.J., Menzel, S., Cox, N.J., Fajans, S.S., Signorini, S., Stoffel, M., and Bell, G.I. (1996). Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). Nature *384*, 458–460.

Yang, J., Schmelzer, K., Georgi, K., and Hammock, B.D. (2009). Quantitative profiling method for oxylipin metabolome by liquid chromatography electrospray ionization tandem mass spectrometry. Anal. Chem. *81*, 8085–8093.

Yang, X., Downes, M., Yu, R.T., Bookout, A.L., He, W., Straume, M., Mangelsdorf, D.J., and Evans, R.M. (2006). Nuclear receptor expression links the circadian clock to metabolism. Cell *126*, 801–810.

Yuan, X., Ta, T.C., Lin, M., Evans, J.R., Dong, Y., Bolotin, E., Sherman, M.A., Forman, B.M., and Sladek, F.M. (2009). Identification of an endogenous ligand bound to a native orphan nuclear receptor. PLoS One *4*, e5609.

Zhang, Y.-K.J., Yeager, R.L., and Klaassen, C.D. (2009). Circadian expression profiles of drug-processing genes and transcription factors in mouse liver. Drug Metab. Dispos. *37*, 106–115.

Zhao, X., Cho, H., Yu, R.T., Atkins, A.R., Downes, M., and Evans, R.M. (2014). Nuclear receptors rock around the clock. EMBO Rep. *15*, 518–528.

Chapter 3

Identification of Affinity Altering SNPs (aaSNPs) Using Protein Binding Microarrays

Contributions from others:

Dr. Nina Titova: Extract preparation and application to PBMs
Dr. Eugene Bolotin: Design of Disease SNP-PBM
Dr. Bin Fang: Design of HNF4α-ChIP-SNP PBM

Abstract

Gene expression is regulated by transcription factors (TFs) that bind specific DNA sequences in regulatory regions. There are more than 1000 TFs encoded in the human genome and each one may bind 1000s of different sequences, making it very difficult to predict which TFs regulate which genes. This problem is compounded by individual genetic variation, referred to as single nucleotide polymorphisms (SNPs) which are often located in TF binding sites in noncoding regions of the human genome. More than 150 million SNPs have been identified among human populations and more than 70,000 associations between SNPs and disease have been made. In this chapter, we developed a high throughput DNA binding assay called protein binding microarrays (PBM) to better define the DNA sequences to which TFs bind. We created several different PBM platforms with which we could perform one million DNA binding reactions in a single experiment in order to identify SNPs that alter the ability of TFs to bind DNA, referred to as affinity altering SNPs (aaSNP). This PBM technology allowed us to identify more than 9,700 aaSNPs with a high degree of confidence (padj. ≤ 0.01) for a subset of nuclear receptors, ligand-dependent TFs that play important roles in physiology and disease. Since nuclear receptors are popular drug targets and since these aaSNPs are just beginning to scratch the surface of common genetic variation, the high throughput PBM system could play an important role in personalized medicine in the future.

Introduction

Recent efforts with genome-wide association studies (GWAS) to identify genetic variations linked to common human disease and phenotypes have identified many variants in intergenic regions of the genome. Until recently, it has been very challenging to identify which genes are affected by these variants, and by what mechanism. The genotype-tissue expression project (GTEx) is a resource aimed at providing insight into human gene expression and regulation and its relationship to genetic variation (Lonsdale et al., 2013). The project collects multiple tissues from organ donors, which are also densely genotyped by whole genome sequencing or Illumina OMNI 5M Array. By analyzing RNA-seq transcriptome data for each tissue and treating expression levels of every transcript as a quantitative trait, genomic variants that are highly correlated can be identified as expression quantitative trait loci (eQTLs). These data are pre-computed by the GTEx project, comparing every variant against transcript expression levels within a ±1 megabase window.

The identification of eQTLs has now provided a very large dataset of correlations between gene expression changes and genetic variants with the added context of the tissue the data was taken from. One limitation of eQTL data, as well GWAS results, is that this type of information only has the power to identify trait loci as opposed to causal variants. This means there may be one or more variants in the immediate vicinity that could also be contributing to the traits being identified because of linkage disequilibrium (LD) and co-inherited traits. Likewise, frequently eQTLs are linked to changes in expression of more than gene (eGene) at a time. The next step in utilizing this power data lies in ascribing a specific mechanism by which any of these eQTLs might be acting to alter the gene expression of nearby genes. One possible mechanism is that these eQTLs may be disrupting the binding sites of transcription factors (TFs) and thereby reducing their effectiveness in transcriptional activation.

The nuclear receptor superfamily also plays a role in modern medicine as they are common drug targets (Kojetin and Burris, 2014; Roshan-Moniri et al., 2014; Sladek, 2003). This is because many of these proteins can have their transcriptional activity modulated by the presence or absence of their corresponding ligands. Additionally, many of these proteins have a tissue-specific expression profile and the result is that synthetic drugs that target the NR ligand binding domain should, in theory, impact only the tissues that express those NRs, thereby reducing the number of off-target effects.

Drugs and endogenous ligands will bind to NRs in the cytosol (class I) or nucleus (class II) and result in receptor localization to the nucleus. Here the NR will bind promoters and enhancers of target genes in a sequence-specific manner. Thus, human genetic variation can play a key role in the effectiveness of a NR. Should an individual contain a SNP that disrupts a binding site for an NR such as HNF4α there is an increased chance of reduced effectiveness of transcriptional activation at the locus. Similarly, it is possible there exist SNPs that convert non-functional promoter binding sites at off-target promoters and enhancers into fully functional binding sites, thus increasing transcriptional activity of genes causing unintended consequences. Here, we investigate whether human genetic variation can lead to a disruption of NR DNA binding sites and whether the identified variants could in turn be linked to changes in nearby gene expression.

By utilizing the high-throughput protein binding microarray (PBM) technology we can investigate *in vitro* the potential of 125,000 SNPs to alter the DNA binding affinity of a NR in a single experiment. The results of these PBM experiments can be cross-referenced with NR regulatory networks to identify disruptive genetic variants within regulatory regions near their target genes. The ultimate goal is to develop a knowledge base of aaSNPs that impact the regulation of NR target genes to better understand NR regulatory networks with an outlook on personalized medicine.

The GTEx project adds another exciting layer to this experiment because we now have access to hundreds of RNA-seq datasets that have paired genotyping across millions of variants. GTEx provides pre-computed eQTL analysis for 53 tissues and many of these tissues have ≥100 samples. Pairing this data with aaSNP PBM datasets will allow for a high-throughput approach to identify human variants that disrupt or enhance NR binding and, importantly, that are highly correlated with allele-specific gene expression from human tissue samples. While the PBM approach does not prove causality, it can show that variants correlated with gene expression changes can alter *in vitro* the DNA binding affinity of nuclear receptors that are known to be enriched in those tissues.

The knowledge that a variant disrupts NR binding and correlates with a change in gene expression of a nearby gene is not sufficient to prove that the change is due to the disruption of binding of the NR. In order to provide some evidence for that causality, we can use the Nuclear Receptor Signaling Atlas (NURSA) Transcriptomine (TM) database, which contains curated RNA-seq and qPCR fold-change data specific to NRs in ligand treatment or NR knockout experiments. (Becnel et al., 2017; Ochsner et al., 2012). The database is updated quarterly to identify new transcriptomics studies involving NR, NR-ligand, and coregulator-dependent gene expression. The fold change values extracted are the processed (summarized and normalized) values submitted by investigators to NCBI's GEO database and EBI ArrayExpress.

In this chapter, we use custom-designed PBMs to identify aaSNPs in NR binding sites and cross reference the associated genes to both GTEx and TM. The net result are eGenes that are targets of specific NRs and whose expression is correlated with SNPs that alter the ability of the NR to bind DNA.

Materials and Methods

Preparation of nuclear extracts for Protein Binding Microarrays

COS-7 cells were transiently transfected with expression vectors for human HNF4α2 (NM_00457), HNF4α8 (NM_175914), RXRα (NM_002957), COUPTFII (NM_021005), GR (NM_000176), TRβ1 (NM_000461), RARα (NM_000964), PPARα (NM_005036) and nuclear extracts were prepared as in Chapter 2. GR applied to Liver

GTEx PBM was extracted from H1993 lung cancer cell line. The concentrations of ligands added 2 h 30 min before harvesting was: GW7647 for PPARα (0.2µM, Cayman Chemicals), 9-cis retinoic acid for RXRα (2µM, Sigma), trans-retinoic acid for RARα (2µM, Sigma #R2500), T3 (3,3′,5-Triiodo-L-thyronine sodium salt) for TRβ1 (70nM, Sigma #T6397), Dexamethasone for GR (100nM, Sigma). The RXRa ligand, 9-cis retinoic acid (Sigma) was also simultaneously added to the PPARα cells at the same concentration as the PPAR ligand.

Protein Binding Microarrays

Protein Binding Microarrays (PBMs) were designed and processed as in Chapter 2 and Bolotin et al. (Bolotin et al., 2010). Protease inhibitor (Sigma) and ligands were added at all incubation steps starting from hybridization. 1µM of each ligand was added to all washes. After purification, NR complexes were applied to arrays and incubated for 15h at 4°C, arrays were washed 3x for 2 min 30 sec each with PBS plus 0.1% Tween 20. Mouse monoclonal anti-Flag antibody (Ab) (Sigma, #F3165) diluted 1:100 in PBS buffer plus 2% non-fat milk, 0.1% Tween 20 were applied directly to the slide and incubated for 48h at 4°C, followed by a conjugated secondary Ab (GαM IgG [H+L] DyLight 550, Pierce #84540) diluted 1:50 (as described above) and then incubated for 4h at room temperature. Three washes, 2 min 30 sec each in PBS plus 0.1% Tween 20 were performed after each antibody incubation.

The same method with the following modifications was used for GR: lung cancer H1993 cells were treated with 0.1 μM Dexamethasone for 1h before harvesting; rabbit monoclonal anti-GR Abs ((D6H2L) XP® Rabbit mAb #12041, Cell Signaling) were used at 1:60 dilution, followed by a conjugated secondary Ab (GαR IgG [H+L] DyLight 550, ThermoFisher #84541) at 1:30 dilution. For RARα, co-expression with RXRα, cells were treated with 9-cis-retinoic acid + trans-retinoic acid (1μM each), mouse monoclonal anti-RARPα Ab (Sigma, #WH0005914M1). For TRβ1, co-expression with RXRα, cells were treated with T3 (3,3′,5-Triiodo-L-thyronine sodium salt) (Sigma #T6397) at 50nM added 3 h before harvesting. Flag Ab as described above but method as in Chapter 2.

Disease SNP-PBM Design

The Disease SNP-PBM was designed by extracting 3000 genes associated with disease from the Genetic Association Database at the time (2011) and extracting all common SNPs (appear in ≥ 1% of the population or are 100% non-reference) from dbSNP v130 within a -6kb to +1kb window around transcription start site (TSS). The net result was 119,743 SNPs. Both major and minor alleles were selected as 29-mers with 14 nucleotides (nt) of genomic DNA before and after the variant to accommodate a full NR binding site of up to 15 nt. As a negative control, 400 random *in silico* 29-mer DNA sequences were included in the design. All alleles plus negative controls were replicated four times on the slide, resulting in a one-million (1M) spot design. Nuclear extracts containing receptors HNF4α2, COUPTFII, RXRα, GR, and TRb1 were applied and PBMs were processed as in Chapter 2.

HNF4α ChIP-SNP-PBM Design

The HNF4α ChIP-SNP-PBM was designed by extracting all common SNPs (≥1% population or 100% non-reference) from dbSNP v132 within a 200-bp window (±100 bp) of the peak center of ChIP-seq data from HepG2 and CaCo-2 (proliferating & differentiated) cell lines for HNF4α (Verzi et al., 2010; Wallerman et al., 2009), and HepG2 cell line for RXRα from the Myers Lab (ENCSR000BHU) (ENCODE Project Consortium, 2012). As in the Disease SNP-PBM, both major and minor alleles were spotted as 29-mers. The same set of negative controls from Disease SNP-PBM were included in this design. All alleles and controls were spotted in four replicates to yield a 1M spot design. Nuclear extracts for receptors HNF4α2, HNF4α8, and RXRα were applied as in Chapter 2.

Liver GTEx SNP-PBM

The Liver GTEx SNP-PBM was designed by taking all variant-gene associations from Liver GTEx v6 database and filtering for the top 140,000 unique eQTLs with an absolute Beta (slope) score ≥ 0.3, P-value ≤ 0.0001, and median RPKM ≥ 1. Variants were ranked in descending order by median RPKM values reported by GTEx to select the top 120,000 unique variants. Alleles were tested as ref and alt as denoted by GTEx ids. The reported chromosome position from GTEx ids was used to extract flanking 14 nt to create a 29mer, like previous designs. The same set of negative controls from previous designs were included. As in previous designs all spots replicated 4 times, resulting in 1 million spot design. Nuclear extracts for receptors HNF4α2, COUPTFII, RXRα, RARα, PPARα, GR, and TRβ1 were applied as in Chapter 2.

Graphical and statistical analysis

Due to the technical variability of the PBM slides, some quality control and normalization steps were necessary. Slides were spatial normalized using the MANOR library in R. Averages and standard deviation were calculated across replicated spots. To identify test sequences with potential outliers, a coefficient of variation (cv) adjusted for small sample sizes was calculated with non-log values as follows:

$$\frac{sd}{avg} * (1 + (\frac{1}{4n}))$$

Any set of replicated spots with cv ≥ 0.5 are selected for outlier removal with a custom Python script. The distance from lowest to second highest score, and distance from highest to second lowest score are measured and compared within each group of four replicates. Outliers were removed with the following criteria; lowest value removed if first comparison was the largest, highest value removed if second comparison was the largest. Averages, standard deviation, and cv of the remaining three replicates are returned to the dataset.

To calculate binding levels of each test sequence in the experiment the entire population of negative controls are averaged together and used as a control population. Each individual test sequence is measured against the control with a one-tailed Student's T-test. Reported P-values are corrected for multiple hypothesis testing using the

Benjamini-Hochberg (“fdr”) option with the p.adjust function from the stats library in R. These values are reported as “padj.” or adjusted P-values.

To calculate affinity altering SNPs (aaSNPs), the PBM binding score of the major and minor (ref and alt) allele averages were compared with a two-tailed Student’s T-test, using standard deviation from control population as pooled variance. P-values were corrected with Benjamini-Hochberg option of the p.adjust function as above. To measure the effect size between the two alleles a Cohen’s D value was calculated by dividing the difference of means by the standard deviation of the pooled negative controls.

All scatterplots were generated in R with ggplot2 library. All receptors applied within a single design were quantile normalized with ‘preprocessCore’ package in R, before plotting. Since aaSNPs were calculated prior to quantile normalization their significance was matched to quantile normalized values before highlighting on the scatterplots. Venn diagrams were generated with the ‘VennDiagram’ package in R.

NR regulatory network datasets from nuclear receptor signaling atlas (NURSA) Transcriptomine (TM) database (https://www.nursa.org/nursa/transcriptomine/) were downloaded in July 2017. Datasets for each NR derived from human or mouse liver were filtered on P-value ≤ 0.05.

Results

Identification of disease-associated aaSNPs

To identify genetic variants that may alter the binding of nuclear receptors (NRs) and also impact transcriptional regulation of disease-associated genes, common SNPs (≥1% in population or 100% non-reference) were extracted from the promoter regions (+6kb to -1kb) of about 3000 genes identified in the Genetic Association Database. Utilizing protein binding microarrays (PBMs) we can measure the binding affinity of a single NR to both the major and minor alleles of about 125,000 SNPs in a single experiment. Slides are designed with a 26-bp "linker" sequence that is attached to the slide which raises the test sequence away from the surface of the slide, and allows for the annealing of a primer that can be extended *in vitro* to make the test sequence double-stranded DNA. For each allele we construct a test sequence by taking the leading and trailing 14 nucleotides (nt) around the variant, and end with a gcgcg cap (Fig 3.1A). Each test sequence is replicated on the slide in quadruplicate to allow for the single removal of an outlier in case of technical variability of the slide (Fig 3.1B).

A total of five NRs were applied to the Disease SNP-PBM -- hepatocyte nuclear factor 4-alpha 2 (HNF4α2, *NR2A1*), COUP transcription factor 2 (COUPTFII, *NR2F2*), retinoic acid receptor alpha (RXRα, *NR2B1*), glucocorticoid receptor (GR, *NR3C1*), thyroid hormone receptor beta (TRβ1, *NR1A2*). Results for the total number of significantly bound alleles (Binders, padj. ≤ 0.01) and the number of affinity altering SNPs (aaSNPs, padj. ≤ 0.01 & Cohen's ≥ 2) are reported, along with the ratio of aaSNPs to total binders (Fig 3.1C). All receptors, except for TRβ1, averaged around 250 aaSNPs with a ratio less than 1% of total aaSNPs to total binders. We find nearly 1,000 aaSNPs for TRβ1with a ratio of roughly 2% of aaSNPs to total binders.

Comparison of disease aaSNPs between receptors

To further characterize this population of aaSNPs, Venn diagrams were generated comparing all receptors against HNF4α2. When looking at significant (padj. ≤ 0.01 & Cohen's ≥ 2) aaSNPs for each receptor, there were very few (a total of 34) that altered the binding affinity of both HNF4α2 and another NR (Fig 3.2A). The NRs that had the greatest number of aaSNPs in common were RXRα and its heterodimeric partners COUTPF and TRβ1: nearly half of all RXRα aaSNPs (113/145, 43.79%) were also aaSNPs for TRβ1 (Fig 3.2B). The low number of shared aaSNPs between HNF4α2 and other NRs gives more evidence that HNF4α2 binds DNA as a homodimer and cannot for heterodimers with other NRs (Bogan et al., 2000; Jiang et al., 1995).

To visualize the aaSNP data between two NRs, scatterplots were generated showing the entire population of test sequences for each receptor and highlighting significant aaSNPs. These plots help to further highlight differences in binding affinities between the receptors, and the diverging profiles of aaSNPs specific to each receptor (Fig 3.2C). While most aaSNPs specific to COUTPF showed low binding affinity for HNF4α2, as seen by the clustering of blue points from (-2,2) on the x-axis, there is a population of HNF4α2-specific aaSNPs that show very high binding affinity for

COUPTF, seen by the yellow points from (2,4) on the y-axis. These are SNPs which successfully altered HNF4α2 DNA binding on the PBM and yet resulted in very high PBM binding scores for COUPTF without altering DNA binding affinity. The distribution of PBM binding scores for COUPTF and RXRα reveal that these two NRs show much more similar DNA binding affinity across all test sequences than any other combination of NRs (Fig 3.2C, *Top Right*). There are many uniquely identified aaSNPs along the diagonal of this plot, suggesting that while statistically these aaSNPs are not identified as common aaSNPs between the two NRs, they are resulting in relatively similar PBM binding scores for the two receptors. Lowering the cut off for designation as an aaSNP to padj ≤ 0.05 increased the shared aaSNPs ratios for COUPTF and RXRα from 16.36% to 42.62% for COUPTF, and from 13.95% to 35% for RXRα for a total of 286, suggesting the initial statistical thresholds might be too strict.

Of particular note is the divergence of PBM binding scores for GR and HNF4α2, which is not surprising given that GR classically binds inverted repeat 1 (IR1; AGAACANTGTTCT) while HNF4α2 prefers DR1 response elements. Similarly, we see a broad range of PBM binding scores between HNF4α, COUPTF, and RXRα despite all three receptors having common binding motifs (DR1).

Identification of aaSNPs Derived from ChIP-seq Peaks

The Disease SNP-PBM is a generic design that can be used to test any given NR, or TF, for aaSNPs within the promoter region of disease-associated genes, but it does have its limitations. Even though we examine the promoter region of the disease-associated genes (-6000 to +1000 of the TSS), we do not know that the NRs actually bind to those promoters. Furthermore, we know that there are many NR binding sites in intergenic and intragenic regions that could be contribute to transcriptional regulation. Therefore, we designed the HNF4a CHIP-SNP PBM using ChIP-seq datasets for HNF4α from HepG2 and CaCo-2 cell lines, both of which express high levels of HNF4α. Even though we used all HNF4α ChIPseq peaks, there was room on the slide for RXRα CHIPseq data from HepG2 cells as well. All common SNPs from a 200 bp (±100 bp) window around peak center were extracted to create the HNF4α ChIP-SNP-PBM (Fig 3.3A). As in the Disease-SNP design, we selected 125,000 SNPs and spotted both alleles with 14 nt of genomic sequence flanking each variant, and included the same set of negative controls.

Three receptors were applied to this design, the two alternative promoter-driven isoforms of HNF4α (HNF4α2 and HNF4α8) and RXRα. This design resulted in a greater number of total binders as well as aaSNPs compared to the Disease SNP-PBM (total bidners for HNF4α2: 84,767 vs 28,415, respectively; aaSNPs for HNF4α2: 1,104 vs 277, respectively) (Fig 3.3B). The same was true for RXRα: (total binders for RXRα: 65,936 vs 27,511, respectively; aaSNPs for RXRα: 822 vs 258, respectively)

HNF4α8 had the largest ratio of aaSNPs identified compared to total binding alleles. While the HNF4α isoforms have identical DNA binding domains, we find that they only share 56.61% (HNF4α2) and 27.77% (HNF4α8) of their aaSNPs with each other. These differences could be due either to subtle but real differences in DNA binding specificity, and/or to technical issues of the variance of the negative controls for that slide, resulting in both fewer binders as well as more aaSNPs. It should be noted that the HNF4α ChIP-seq datasets used to make this design were from HepG2, which expresses both P1- and P2-HNF4α, but mostly P1-HNF4α, and CaCo-2 differentiated and proliferative cells.

RXRα shares fewer aaSNPs with HNF4α2 (24.13%) than with HNF4α8 (39.97%) but these represent a similar, albeit lower, percentage of the HNF4α2 and HNF4α8 aaSNPs -- 17.66% and 14.32%, respectively (Fig 3.3B). It is interesting to note that Spearman's coefficient for RXRα and HNF4α8 are worse (0.29, P-value: 2.2e-16) than between RXRα and HNF4α2 (0.70, P-value: 2.2e-16).

Scatter plots of PBM binding scores for these receptors shows that the HNF4α isoforms bind relatively similarly to most alleles (Fig 3.3C, *left*). The similar PBM binding scores between HNF4α2 and HNF4α8 taken together with the distribution of many of the shared aaSNPs (green) overlaying HNF4α8-specific aaSNPs suggests that a less strict statistical analysis would likely reveal an even greater overlap between the two isoforms.

Comparing RXRα to both HNF4α isoforms we find a broader range of PBM binding scores within each comparison. Highlighting unique (red, yellow, and blue) and common (purple) aaSNPs we can see that most of the RXRα-unique aaSNPs (red spots) result in low PBM binding scores for HNF4α isoforms as seen by the vertical clustering of red (0,2.5) in the x-axis of both the HNF4α2 and HNF4α8 plots. Many of the HNF4α2 and HNF4α8 aaSNPs are scattered vertically throughout the plot suggesting that while they significantly alter the binding of HNF4α isoforms, RXRα still binds many of these variants (Fig 3.3.C, *right*).

While both HNF4α and RXRα are known to bind DR1 motifs, these results show that there is a wide range of sequences derived from ChIP-seq datasets of both factors that bind the NRs. In comparing predicted HNF4α binding sites we find a few interesting features between the common, HNF4α-, and RXRα-specific aaSNPs. Many of the HNF4α-specific aaSNPs showed a genetic variant within the central "CAAAG" of the consensus HNF4α motif (Fig 3.4A). The aaSNPs shared between the two NRs generally resulted in genetic variants in the right half site, but still showed a preference for the "CAAAG" motif in the center of the motif (Fig 3.4B). Perhaps because of the preference for the "GGTCA" in the right half site by RXRα we also find that many of these shared aaSNPs disrupt a near perfect "AGGTCA" right half site. As expected, RXRα-specific aaSNPs have almost completely lost the central "CAAAG" that HNF4α prefers, but retain the "GGTCA" ending to the right half-site that is so prevalent in RXRα-specific motifs (Fig 3.4C). Considering the importance of the right half-site in the DR1 for RXRα-specific binding, we would have expected more aaSNPs to disrupt the last 4 nt of that motif. Nonetheless, we find most of the RXRα-specific aaSNPs in the first 2 nt of the half site "AGGTCA". Thus, genetic variants that disrupt HNF4α-specific binding do so via the central “CAAAG”, and those that disrupt RXRα-specific binding do so via the right half-site of DR1 motifs lacking a “CAAAG” motif.

Identification of aaSNPs in NR Regulatory Pathways

While the ChIP-SNP-PBM successfully identified a large number of genetic variants with the capability of altering the binding affinity of a NR, still missing was any information about the impact of the variants on gene expression. With the release of v6 of the GTEx Project we gained access to 97 RNA-seq and genotyped human liver samples. Utilizing the eQTLs calculated by the GTEx Project we identified the top 125,000 genetic variants correlated with the largest and most significant changes in gene expression of the most highly expressed genes in the liver (Slope ≥ 3, P-value ≤ 0.0001, FPKM ≥ 1) (Fig 3.5A). As in the previous designs, both alleles for each variant were spotted with the flanking 14 nt, and spotted in quadruplicate. The same set of 400 negative controls from the previous designs were included as controls for the Liver GTEx SNP-PBM design.

Since this design was not derived from any specific ChIP-seq binding profiles, we applied a wide array of NRs expressed in the liver in order to identify additional genetic variants that correlate with changes in gene expression and that alter the binding affinity for a NR. Overall, the total number of alleles with significant levels of binding (padj. ≤ 0.01) were much lower than the previous ChIP-SNP-PBM design, but in a similar range as the Disease SNP-PBM (Fig 3.5B, *top*). The number of significant aaSNPs (padj. ≤ 0.01 and Cohen’s ≥ 2) identified for each receptor was highly variable with the highest number found for HNF4α2 (1,713) and the lowest number found for GR (48). A comparison of the significant aaSNPs between HNF4α2 and the other receptors show similar results as with the Disease SNP-PBM; namely, a small number of aaSNPs that disrupt more than one NR (Fig 3.5B, *bottom*). Plotting quantile normalized PBM binding scores for reference and alternate alleles of a genetic variant against themselves reveals that for most NRs there does not appear to be any allelic bias, except for HNF4α2 and GR (Fig 3.5C): HNF4α2 has a slight preference for the reference allele (1,113 vs 600) while GR has a preference for the alternate allele (20 vs 28). At first, these results might suggest that HNF4α2 binds the most prevalent alleles, perhaps due to conservation of functional binding sites. However, the reference allele of a SNP is not always the major frequency allele most commonly found in a population. Why these two NRs would show any preference to a SNP designation given based on genotyping the reference genome is beyond us at this point, but would warrant further research.

To visualize the overlap in the aaSNP data between the receptors, scatterplots of PBM binding scores were generated as done for the previous designs (Fig 3.6). The profiles between HNF4α2 and the other NRs were strikingly similar to those from the Disease-SNP design: this is not unexpected as both PBMs contain genomic sequences with unknown TF DNA specificity from the human genome.

Highlighting significant aaSNPs in each plot shows that most of the variants, while capable of altering the DNA binding affinity of one receptor, do not alter the binding of a second NR. This can be easily seen in the COUPTF vs HNF4α2 plot where most COUPTF-specific aaSNPs (blue spots) cluster on the x-axis at (0,2), and likewise most HNF4α2-specific aaSNPs (yellow spots) cluster on the y-axis at (0,2). Again, the NR that is most divergent from HNF4α2 in this dataset is GR, which binds a different half site from HNF4α (AGAACA vs AGGTCA) and as an inverted, not a direct, repeat.

While all the genetic variants chosen for this PBM design were statistically significant eQTLs in human livers, and while the PBM results show that >1000 significantly disrupt DNA binding affinity of NRs expressed in the liver, we still needed to verify that the related eGenes can actually be regulated by the cognate receptor. To accomplish this, we examined NR regulatory datasets from the TM database to identify potential target genes for each NR. Statistically significant changes in gene expression in either knockout or ligand-treated studies were cross-referenced with the eGenes from GTEx based on the PBM results. HNF4α2 showed the largest number of target genes implicated in this aaSNP-eQTL analysis (125), while the other receptors showed 11 or less, with GR resulting in no overlap (Fig 3.7A). The top 25 aaSNP-eGene comparisons for HNF4α2 based on PBM padj are shown in Fig 3.7B. These (and 100 other aaSNP-eGenes) all have statistically significant aaSNP values from the PBM, eQTL values from GTEx and fold change in TM based on HNF4α knockout. For example, in *SCG5* the alt allele of "15_33065311_G_A_b37" decreases HNF4α2 binding (Cohen's 4.31, padj 4.57e-04) and results in a an associated decrease in expression in GTEx and is a verified positive target of HNF4α2 in TM being downregulated -5.03-fold in HNF4α knockout experiment. It is interesting to note that all the top 25 aaSNP-eGenes in this table report stronger binding of HNF4α2 to the reference allele, while GTEx and TM associations are quite mixed in their directionality.

We find two aaSNPs with ≥1 eGene associated with them, 6_32610868_C_T_b37 (*C4A* and *HLA-DQB1*) and 6_32633354_C_CT_b37 (*C4A* and *TAPBP*), and four separate aaSNPs correlated with changes in gene expression of complement gene *C4A* (Fig 3.7B).

To highlight the advantages of the PBM approach for identifying aaSNPs *in vitro* with potential *in vivo* implications, we compared the impact of reference and alternate alleles within the dataset. The eQTL beta score, or slope as it is referred to in GTEx v7, indicates whether the associated transcripts are up- or down-regulated in individuals with the alternate allele. Similarly, we can indicate whether the alternate alleles increase or decrease the DNA binding affinity of a given NR. After identifying the regulatory role of the NR on a given target gene using the TM dataset, we can identify aaSNPs with positive or negative correlations with the changes in gene expression.

We find two positive correlated eQTLs with variants 7_940267_T_C_b37 and 6_32610868_C_T_b37. The former showing increased expression of *PRKAR1B* and increased DNA binding affinity with the alternate allele (C), and a -2.08-fold change in expression in HNF4α mouse liver knockout, suggesting that *PRKAR1B* is activated by HNF4α2 (Fig 3.8A). The genetic variant 7_940267_T_C_b37, with minor allele frequency T=0.331, is associated with an increase in *PRKAR1B* gene expression in individuals with the alternate allele and the PBM results show increased DNA binding affinity with the alternate allele. This means that in an individual carrying the alternate allele, 66.9% of the population, HNF4α2 will bind better and expression of *PRKAR1B* will increase. *PRKAR1B* encodes protein kinase cAMP-dependent type I regulatory subunit beta, a regulatory subunit of cAMP protein kinase A (PKA) and is involved in the signaling of the second messenger cAMP. It has been shown that HNF4α DNA binding activity can be inhibited by PKA phosphorylation of the DBD (Viollet et al., 1997), suggesting individuals with the alternate allele could see decrease HNF4α binding activity in response to cAMP inducers.

The variant 6_32610868_C_T_b37 is associated with a decreased expression of *C4A* and decreased DNA binding affinity with the alternate allele (T), and a -3.08-fold change in expression in HNF4α2 mouse liver knockout, suggesting that *C4A* is positive target of HNF4α2 (Fig 3.8B). This variant, with minor allele frequency T=0.206, is associated with a decrease in *C4A* gene expression in individuals with the alternate allele and PBM results also show decreased DNA binding affinity for this allele. This means someone with the alternate allele, 79.4% of the population, will have reduced HNF4α2 binding and the expression of *C4A* will decrease. The *C4A* gene encodes complement C4A, the acidic form of complement factor 4 and is essential for the classical complement pathway. Proteolytic degradation cleaves C4 into three chains (alpha, beta, and gamma). The alpha chain, C4a anaphylatoxin, an antimicrobial peptide and a mediator of local inflammation. Deficiency of C4A is associated with systemic lupus erythematosus and type I diabetes mellitus.

In a third example, the correlation between the PBM and GTEx data holds even for a negative target gene of HNF4α2. *CTSS* is up-regulated 5-fold in HNF4α knockout in mouse liver data, suggesting that HNF4α2 represses the expression of *CTSS* (Fig 3.8C). The genetic variant 1_150821847_G_A_b37 is associated with an increase in *CTSS* gene expression in individuals with the alternate allele and the PBM results show disruption of DNA binding affinity with the alternate allele. Thus, in individuals carrying the alternate allele, HNF4α2 will bind less well and expression of the gene will increase. *CTSS* encodes cathepsin S, a member of the peptidase C1 family and a lysosomal cysteine proteinase that may participate in the degradation of antigenic proteins to peptides for presentation on MHC class II molecules. It is not too surprising that we find aaSNPs in genes such as *CTSS* and *C4A* that are part of the immune system, which is known to have a high degree of variability between individuals. Indeed, there are a total of four aaSNP-eQTLs in *C4A* among the top 25 most statistically significant aaSNPs for HNF4α2 (Fig. 3.7B).

Discussion

Over the past decade GWAS studies have identified thousands of noncoding genetic variants associated with various human diseases and phenotypes. An important step in elucidating the mechanism by which these variants may impact gene expression is to determine whether the variants disrupt TF recruitment to promoters and enhancers. Other techniques to measure protein-TF interactions (EMSA gel shifts, DNAse footprinting, ChIP, and Yeast two-hybrid assays) prove to be either low-throughput or too time-consuming to generate data for multiple TFs in parallel. PBMs have the advantage of simultaneous measurement of DNA binding affinity of a TF against tens of thousands (15,000-225,000) of DNA sequences in a single experiment. Additionally, once a slide is designed the experiment can be repeated for additional replicates, or compared across any TF with western blot quality antibody.

Potential *in vivo* approaches to determine DNA binding affinity of a TF involve chromatin immunoprecipitation (ChIP) to cross-link and pull down genomic DNA fragments bound a given TF. The results of these experiments will differ largely based on cell conditions and cell type. Additionally, to analyze rare genetic variants many samples may be required until a sample is processed from an individual carrying the rare minor allele. With the PBM technology, a rare minor allele with flanking DNA sequence can be printed on the slide and analyzed for DNA binding affinity with any number of TFs. This makes PBMs the ideal technique for initial identification of aaSNPs from any GWAS and eQTL datasets.

Our results show that the “broad” aaSNP approaches, such as those used in the Disease SNP-PBM and Liver GTEx PBM designs derived from genomic sequence with unknown TF binding specificity are useful in their wide application towards many TFs, but are limited in their power to identify as many aaSNPs as possible. This can be seen by the lower aaSNP/binder ratio in both broad designs compared to higher ratios seen in the HNF4α ChIP-SNP-PBM.

The Disease SNP-PBM results show that for significant aaSNPs there is very little overlap between two NRs, except for TRβ1 and RXRα which are known to heterodimerize: nearly 43% of RXRα aaSNPs also disrupt binding of TRβ1 (Fig 3.2A). A major drawback of the Disease SNP-PBM is that without an *in vivo* DNA binding assay, such as ChIP-seq, we lack any knowledge about the physiological relevance of any aaSNPs identified from the PBM. These results can be easily cross-referenced with ChIP-seq datasets, if available, to verify that a given TF does in fact bind the locus in vivo. CHIP-seq results can also inform about tissue- or condition-specific binding in the promoters of disease-associated genes.

The HNF4α ChIP-SNP-PBM attempts to improve upon the previous design by coupling the genetic variant selection with ChIP-seq datasets for HNF4α2 from HepG2 and CaCo-2 cell lines, and RXRα from HepG2. While the scope of this design allows us to identify aaSNPs from *in vivo* DNA binding data, there may be more HNF4α and RXRα loci from different tissues and conditions that are not captured from in these three datasets. As a result of this more narrow approach we identified a much higher ratio of aaSNPs to binders for both HNF4α isoforms. In contrast, the aaSNP/binder ratio was not increased in the RXRα dataset, which could be due to either noisier data from the slide or to the fact that most of the sites were from HNF4a CHIP-seq peaks, not RXRa peaks.

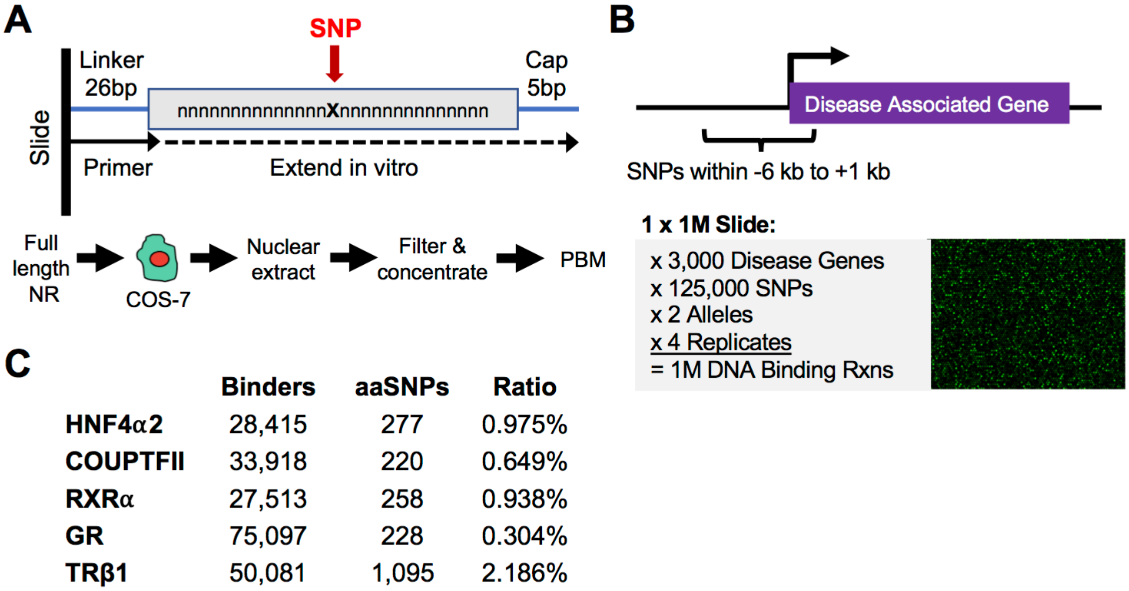
Perhaps the most interesting result from this experiment was the broad range of PBM binding scores between HNF4α isoforms and RXRα. Looking more closely at the aaSNPs shared and unique to both NRs reveals some insight into why the two receptors do not bind so similarly. Many of the RXRα-specific aaSNPs are for DNA sequences which do not contain very strong HNF4α motifs, as noted by the lack of “CAAAG”, but still contain a strong right half site “NGGTCA” that is critical for RXRα binding. Similarly, the HNF4α2-specific motifs contain the “CAAAG” preferred by HNF4α2 but have slightly degenerated right half sites like “AGTCCA” which do not contain the “TCA” in the last three positions that RXRα seems to prefer. It is interesting to note that while the scope of the project is to identify common genetic variants that can alter DNA binding affinity of NRs for potential impact on transcriptional regulation, the data can still be used to further define NR-specific motifs, as we did previously (Fang et al., 2012).

Linking aaSNPs to in vivo gene expression

The Liver GTEx PBM was successful in identifying many aaSNPs associated with changes in gene expression of target genes of HNF4α2. Many of the other NRs were less successful, in part due to the smaller TM datasets available but also potentially because of the broad approach taken in the design.

As the GTEx project continues to grow we should gain more power in identifying eQTLs in the liver, but with nearly 11 million genetic variants in total we will need to continue to be scrutinous in our selection of eQTL for testing. ChIP-seq datasets can help inform designs to select genetic variants from regions where NRs bind in normal liver tissue. Alternative approaches could include first selecting a set of target genes for each

NR and filling the design with all significant eQTLs related to those genes. If that number does not reach 125,000, then the design could be completed with the next most significant eQTLs that fall within known *in vivo* binding sites, but we are currently limited in the number of ChIP-seq datasets for TFs from normal tissues publicly available.



	Binders	aaSNPs	Ratio
HNF4α2	28,415	277	0.975%
COUPTFII	33,918	220	0.649%
RXRα	27,513	258	0.938%
GR	75,097	228	0.304%
TRβ1	50,081	1,095	2.186%

Figure 3.1

Figure 3.1 Identification of aaSNPs in Promoter Regions of Disease-Associated Genes

(*A*) Schematic of protein binding microarray (PBM) and workflow. (*B*) Disease SNP-PBM design. About 3000 disease associated genes were collected from the Genetic Association Database. All common SNPs that appear in at least 1% of the population or are 100% non-reference were selected from a -6kb to +1 kb window around TSS to a total 125,000 SNPs. Both reference (ref) and alternate (alt) alleles spotted in quadruplicate generating a total of 1 million (1M) spots. Included in the design were 400 random DNA control sequences to measure a population of non-specific binding. Diagram showing total number of SNPs, alleles, and replicates spotted on 1M design, with image of COUPTFII binding intensities as seen on microarray scanner. (*C*) Shown are the total number of significantly bound alleles (adjusted P-value; padj. ≤ 0.01) and total number of affinity altering SNPs (aaSNP; padj. ≤ 0.01 & Cohen's ≥ 2) for the five NRs applied to this design. Cohen's D statistic was calculated using the standard deviation of 400 random controls. The ratio of aaSNPs to total binders is also shown.

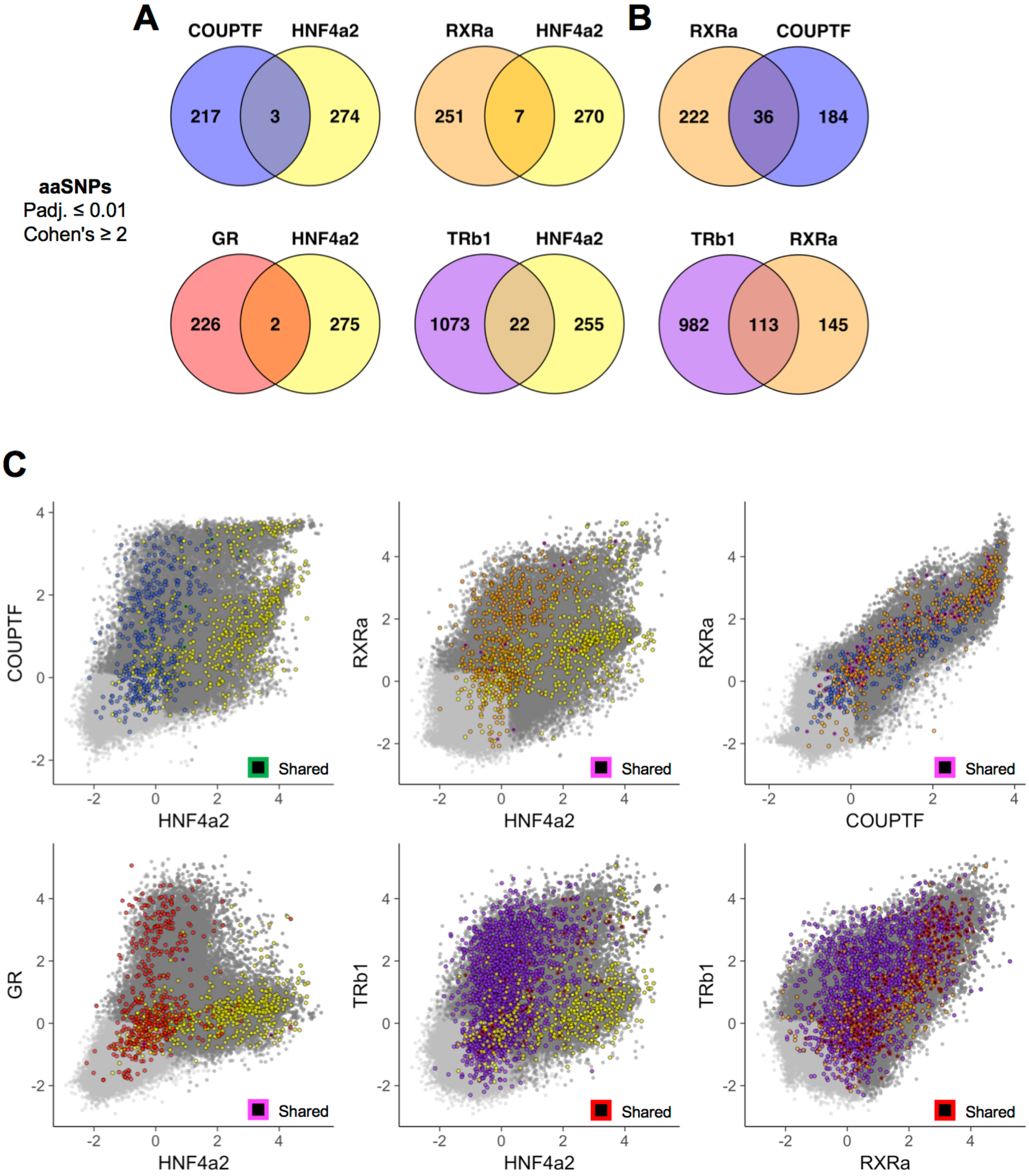
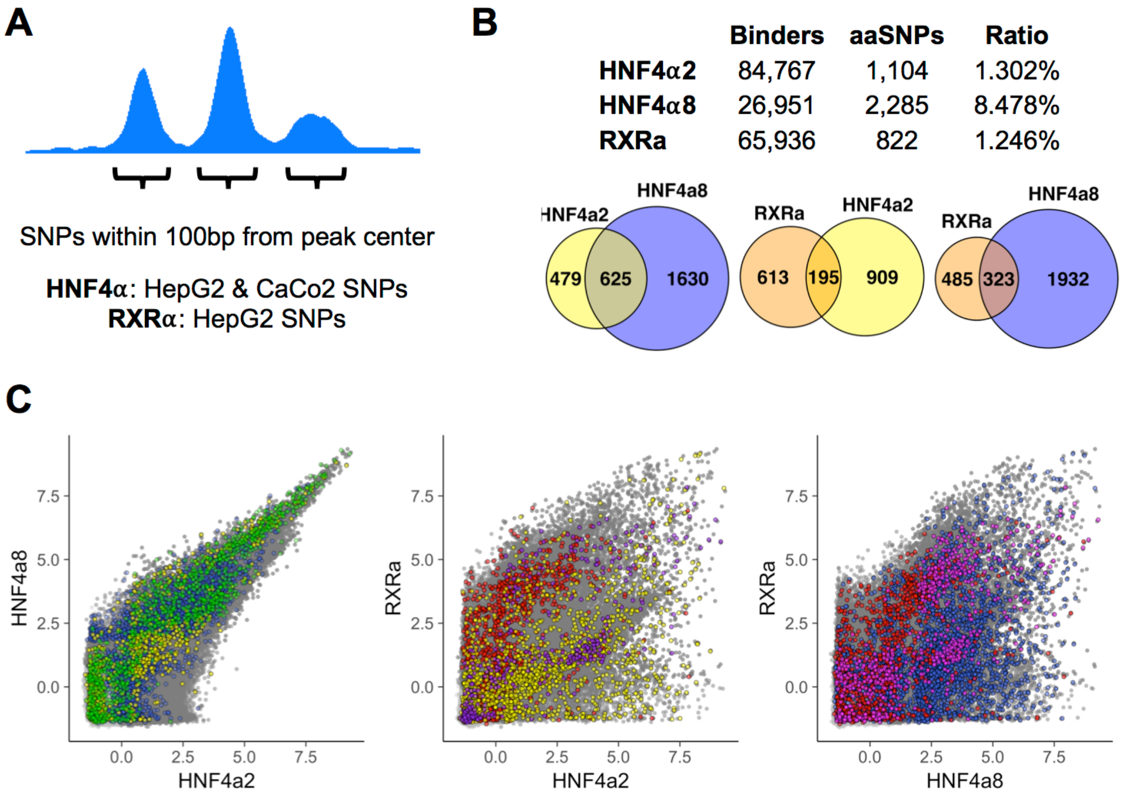

A
B
C
aaSNPs
Padj. ≤ 0.01
Cohen's ≥ 2
COUPTF
HNF4a2
217
3
274
RXRa
HNF4a2
251
7
270
RXRa
COUPTF
222
36
184
GR
HNF4a2
226
2
275
TRb1
HNF4a2
1073
22
255
TRb1
RXRa
982
113
145
COUPTF
HNF4a2
Shared
RXRa
HNF4a2
Shared
RXRa
COUPTF
Shared
GR
HNF4a2
Shared
TRb1
HNF4a2
Shared
TRb1
RXRa
Shared
-2
0
2
4

Figure 3.2

Figure 3.2 Low Occurrence of aaSNPs Disrupting Multiple Receptors in Disease-SNP PBM

(*A*) Venn diagrams comparing aaSNPs identified for each NR based on PBM binding scores. A difference of binding between reference and alternate alleles is counted as a single aaSNP (padj. ≤ 0.01 & Cohen's ≥ 2). (*B*) Scatterplots of quantile normalized PBM binding scores, each point is the average binding score of a single allele. All binders (padj. ≤ 0.01) plotted in dark grey, ref and alt alleles for aaSNPs are highlighted with same coloring from Venn diagrams. Common aaSNPs are highlighted with an alternate color for each grid.



	Binders	aaSNPs	Ratio
HNF4α2	84,767	1,104	1.302%
HNF4α8	26,951	2,285	8.478%
RXRa	65,936	822	1.246%

Figure 3.3

Figure 3.3 Identification of aaSNPs in HepG2 and CaCo-2 ChIP-seq Peaks

(*A*) HNF4α ChIP-SNP-PBM design. All common SNPs within ±100 bp of peak center in HNF4α ChIP-seq peaks from HepG2 and CaCo-2 cell lines, and RXRα ChIP-seq peaks from HepG2 cell line were spotted with reference and alternate alleles in quadruplicate. Roughly 125,000 total SNPs were selected for a 1 million spot slide. (*B*) *Top*, three NRs were applied to this design; alternative-promoter isoforms of HNF4α, HNF4α2 and HNF4α8, and RXRα. Reported are total number of significantly bound alleles (padj. ≤ 0.01), number of significant aaSNPs (padj. ≤ 0.01 & Cohen's ≥ 2), and the ratio of aaSNPs to total binders. *Bottom*, Venn diagrams of aaSNPs identified between NRs. (*C*) Scatter plots of quantile normalized comparisons. All significant binders (padj. ≤ 0.01) plotted in dark grey, reference and alternate alleles for aaSNPs are highlighted with same color code from venn diagrams. aaSNPs common to both NR are highlighted with an alternate color for each grid.

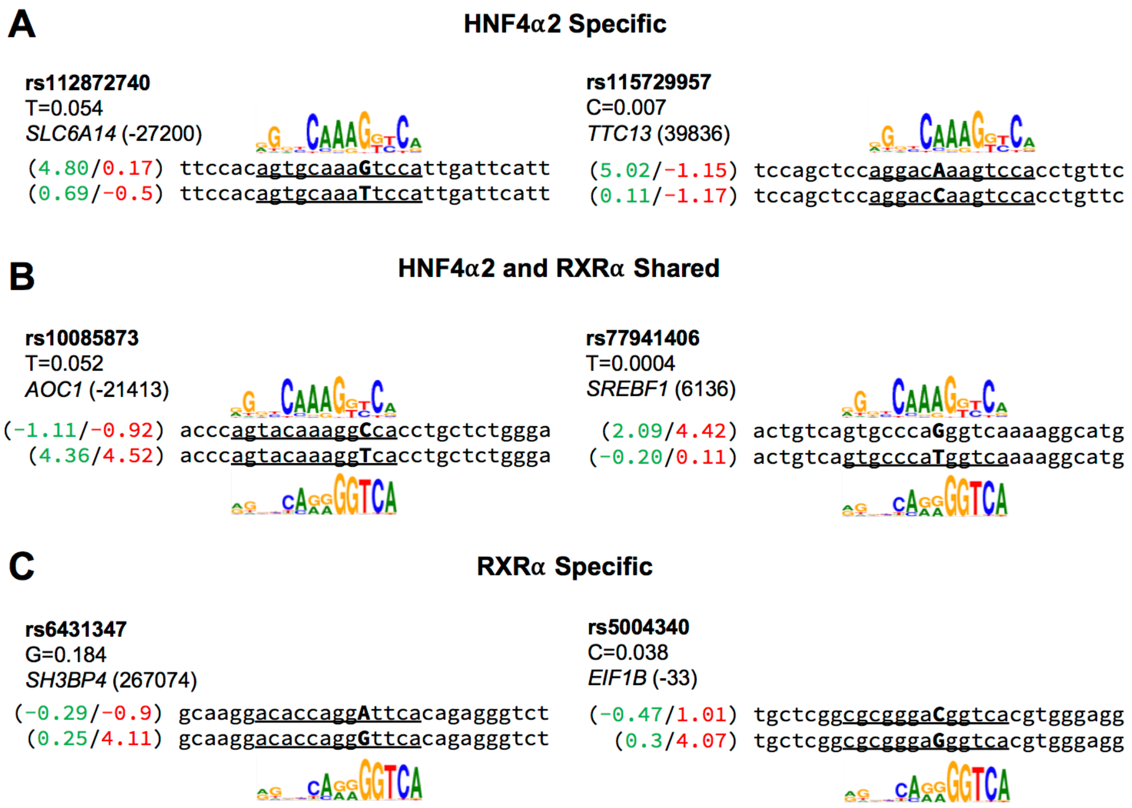
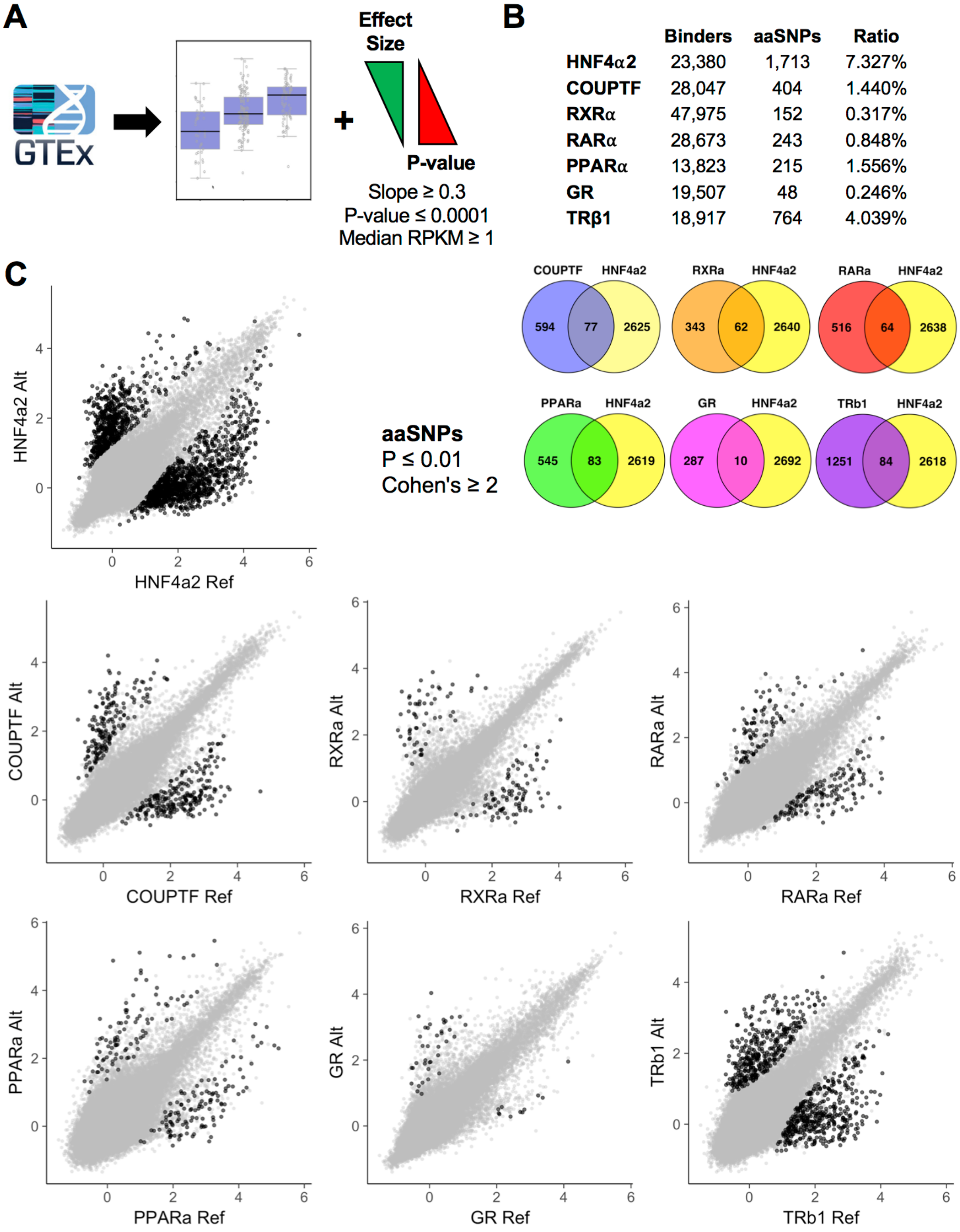

A
HNF4α2 Specific
rs112872740
T=0.054
SLC6A14 (-27200)
(4.80/0.17) ttccacagtgcaaaGtccattgattcatt
(0.69/-0.5) ttccacagtgcaaaTtccattgattcatt
rs115729957
C=0.007
TTC13 (39836)
(5.02/-1.15) tccagctccaggacAaagtccacctgttc
(0.11/-1.17) tccagctccaggacCaagtccacctgttc
B
HNF4α2 and RXRα Shared
rs10085873
T=0.052
AOC1 (-21413)
(-1.11/-0.92) acccagtacaaaggCcacctgctctggga
(4.36/4.52) acccagtacaaaggTcacctgctctggga
rs77941406
T=0.0004
SREBF1 (6136)
(2.09/4.42) actgtcagtgcccaGggtcaaaaggcatg
(-0.20/0.11) actgtcagtgcccaTggtcaaaaggcatg
C
RXRα Specific
rs6431347
G=0.184
SH3BP4 (267074)
(-0.29/-0.9) gcaaggacaccaggAttcacagagggtct
(0.25/4.11) gcaaggacaccaggGttcacagagggtct
rs5004340
C=0.038
EIF1B (-33)
(-0.47/1.01) tgctcggcgcgggaCggtcacgtgggagg
(0.3/4.07) tgctcggcgcgggaGggtcacgtgggagg

Figure 3.4

Figure 3.4 ChIP-SNP-PBM Examples

Shown are examples of aaSNPs (padj. ≤ 0.01 & Cohen's ≥ 2) from the HNF4a CHIP-SNP PBMs. (*A*) Spotted test sequences and PBM binding scores for HNF4α2-specific aaSNPs. Minor allele frequency and nearest TSS with distance shown below dbSNP rs number. HNF4α2 binding scores in green, RXRα binding scores in red. Underlined sequence represents best HNF4α motif alignment with consensus HNF4α2 motif shown below. (*B*) As in (*A*) but with aaSNPs common to HNF4α2 and RXRα with HNF4α2 motif shown above and RXRα motif shown below. (*C*) as in (*A*) but with aaSNPs unique to RXRα and RXRα motif shown below.



	Binders	aaSNPs	Ratio
HNF4α2	23,380	1,713	7.327%
COUPTF	28,047	404	1.440%
RXRα	47,975	152	0.317%
RARα	28,673	243	0.848%
PPARα	13,823	215	1.556%
GR	19,507	48	0.246%
TRβ1	18,917	764	4.039%

Figure 3.5

Figure 3.5 Identification of aaSNPs in Tissue-Specific eQTLs

(*A*) Liver GTEx SNP-PBM design. All significant SNP-gene associations were extracted from GTEx project v6 with effect size ≥ 0.3 (Slope), P-value ≤ 0.0001, and median RPKM ≥ 1 across all liver samples. (*B*) *Top*, seven nuclear receptors were applied to this design. Reported are total number of significantly bound alleles (padj. ≤ 0.01), significant aaSNPs (padj. ≤ 0.01 & Cohen's ≥ 2), and ratio of aaSNPs to total binders. *Bottom*, Venn diagrams of comparisons of aaSNPs identified for multiple NRs. (*C*) Scatterplots of quantile normalized PBM binding scores of reference alleles plotted against alternate alleles. Highlighted (black) spots are significantly different (padj. ≤ 0.01 & Cohen's ≥ 2; aaSNPs).

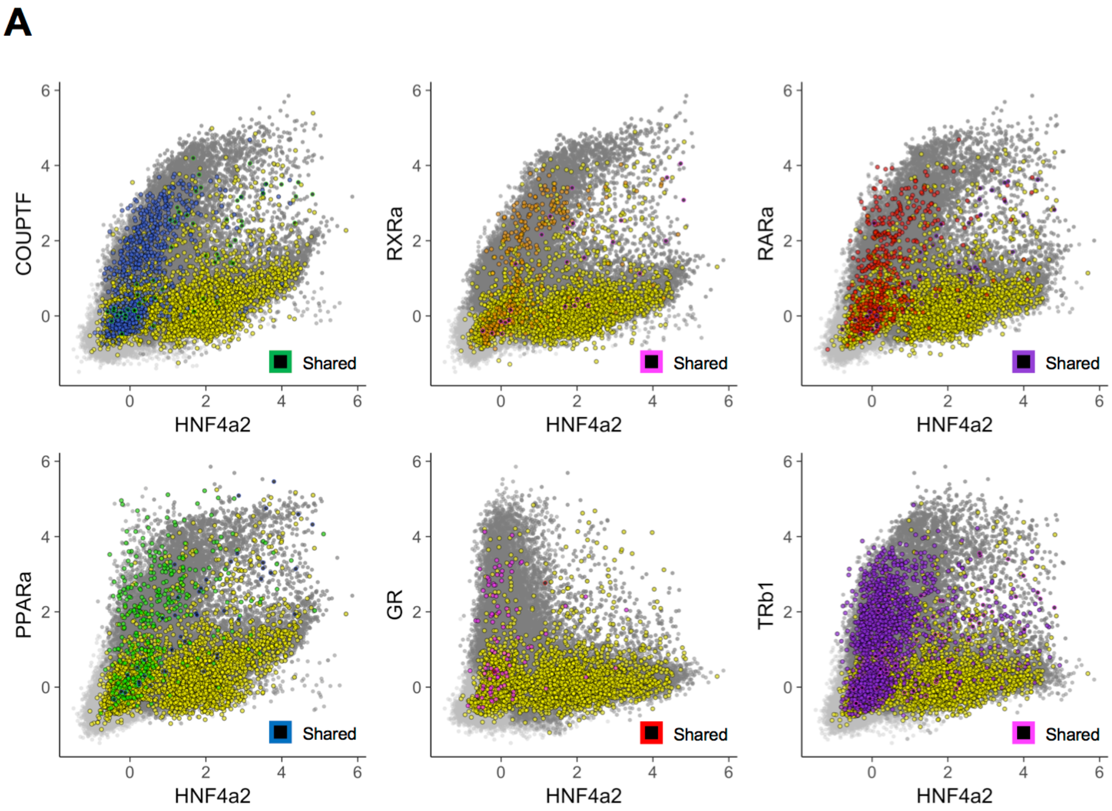
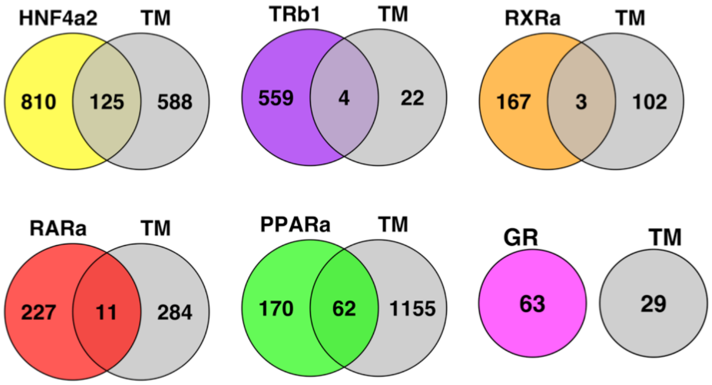


Figure 3.6

Figure 3.6 Distribution of aaSNPs between receptors

Scatterplots of quantile normalized PBM binding scores for each allele. All binders (padj. ≤ 0.01) plotted in dark grey, ref and alt alleles for aaSNPs are highlighted with same coloring from Venn diagrams in Fig 3.5. Common aaSNPs are highlighted with an alternate color for each grid.

A



B

PBM						GTEx				Transcriptomine	
ID	**Ref**	**Alt**	**Cohen's**	**Obs**	**Padj**	**Pval**	**Beta**	**RPKM**	**Sym**	**FC**	**Pval**
5_172630811_C_G_b37	1.54	0.11	4.18	1.43	2.16E-04	3.92E-05	0.45	2.36	BNIP1	2.07	1.39E-02
11_95523123_C_G_b37	1.88	0.27	4.69	1.60	3.44E-04	1.01E-05	0.49	8.31	CEP57	-2.28	3.02E-08
20_25267893_A_G_b37	1.01	-0.19	3.50	1.20	4.14E-04	1.95E-06	0.50	7.30	ABHD12	-2.22	1.16E-04
15_33065311_G_A_b37	**1.84**	**0.37**	**4.31**	**1.47**	**4.57E-04**	**3.15E-04**	**-0.31**	**2.80**	**SCG5**	**-5.03**	**2.37E-03**
1_40257157_T_C_b37	3.29	1.17	6.22	2.12	8.64E-04	1.13E-05	0.70	3.51	PPIE	2.19	1.99E-02
11_119967918_A_C_b37	0.98	-0.06	3.03	1.04	9.12E-04	4.05E-04	-0.35	4.11	C2CD2L	-2.81	1.35E-02
2_121530351_C_T_b37	1.11	0.08	3.00	1.02	1.09E-03	2.02E-04	0.44	7.62	EPB41L5	-2.13	1.15E-04
11_73698738_T_C_b37	1.96	0.29	4.87	1.66	1.60E-03	3.94E-04	0.31	1.89	UCP2	-4.35	1.05E-03
9_123421556_C_T_b37	1.68	-0.03	5.02	1.72	1.94E-03	7.11E-06	-0.58	5.97	PSMD5	-2.93	2.35E-02
9_37528290_C_A_b37	1.82	0.62	3.52	1.20	2.32E-03	4.85E-04	0.58	5.93	POLR1E	-2.23	6.20E-07
6_31117075_C_A_b37	2.82	0.14	7.83	2.68	2.58E-03	2.64E-07	-0.75	150.90	HLA-C	2.19	5.00E-10
1_150821847_G_A_b37	**1.79**	**0.16**	**4.78**	**1.63**	**2.73E-03**	**6.96E-07**	**0.34**	**5.03**	**CTSS**	**4.91**	**1.18E-02**
6_32593592_T_C_b37	2.58	0.61	5.77	1.97	3.12E-03	4.69E-04	-0.32	28.24	**C4A**	-3.09	0.00E+00
5_122144906_G_A_b37	1.75	0.51	3.63	1.24	3.52E-03	1.37E-05	0.55	9.55	PPIC	2.17	3.00E-10
6_32551595_C_T_b37	1.37	0.09	3.74	1.28	3.61E-03	4.75E-04	-0.41	28.24	**C4A**	-3.09	0.00E+00
6_32651694_C_T_b37	3.49	1.32	6.37	2.18	3.78E-03	1.13E-05	0.32	13.45	ATF6B	2.18	4.09E-02
12_31206764_T_C_b37	1.51	0.32	3.50	1.20	4.21E-03	1.84E-05	-0.62	3.35	DDX11	2.16	7.03E-06
19_15853717_C_A_b37	2.53	0.85	4.90	1.67	4.70E-03	6.34E-07	0.39	15.42	CYP4F12	-11.83	1.42E-05
21_40763986_T_G_b37	1.38	0.31	3.11	1.06	5.32E-03	2.95E-05	-0.56	1.88	SH3BGR	4.01	1.43E-02
15_45713801_C_T_b37	2.61	0.90	5.00	1.71	6.18E-03	5.91E-06	0.41	2.30	SPATA5L1	2.12	7.19E-04
6_32800224_C_A_b37	1.27	0.21	3.12	1.07	6.95E-03	3.68E-04	-0.36	10.43	PSMB9	5.68	0.00E+00
6_32610868_C_T_b37	1.03	-0.10	3.30	1.13	7.39E-03	3.93E-05	-0.63	1.24	HLA-DQB1	12.12	8.61E-04
6_32610868_C_T_b37	**1.03**	**-0.10**	**3.30**	**1.13**	**7.39E-03**	**3.31E-05**	**-0.39**	**28.24**	**C4A**	**-3.09**	**0.00E+00**
6_32633354_C_CT_b37	1.65	0.59	3.09	1.06	8.62E-03	4.70E-04	-0.38	37.73	TAPBP	3.69	0.00E+00
6_32633354_C_CT_b37	1.65	0.59	3.09	1.06	8.62E-03	2.38E-04	-0.37	28.24	**C4A**	-3.09	0.00E+00

Figure 3.7

Figure 3.7 eQTLs of Nuclear Receptor Target Genes

(*A*) Venn diagrams of distinct gene symbols from all significant aaSNPs (padj. ≤ 0.01 & Cohen's ≥ 2) for the indicated NR cross-referenced with distinct gene symbols from related Transcriptomine (TM) NR regulatory networks. (*B*) Table of the 25 most significant (based on PBM padj.) aaSNPs identified for HNF4α2 with associated TM fold-change values based on HNF4α knockout data. Green columns refer to PBM-based data; GTEx/Spot ID, ref and alt allele PBM binding scores, Cohen's D effect size of aaSNP, observed effect (ref-alt), and FDR corrected P-values (padj.). Yellow columns refer to GTEx-based data (v6); P-value and Beta (Slope) of the eQTL analysis, median RPKM, and gene symbol of associated eGene. Blue columns refer to Transctiptomine-based data; fold change of eGene in HNF4α knockout and P-value associated with effect.

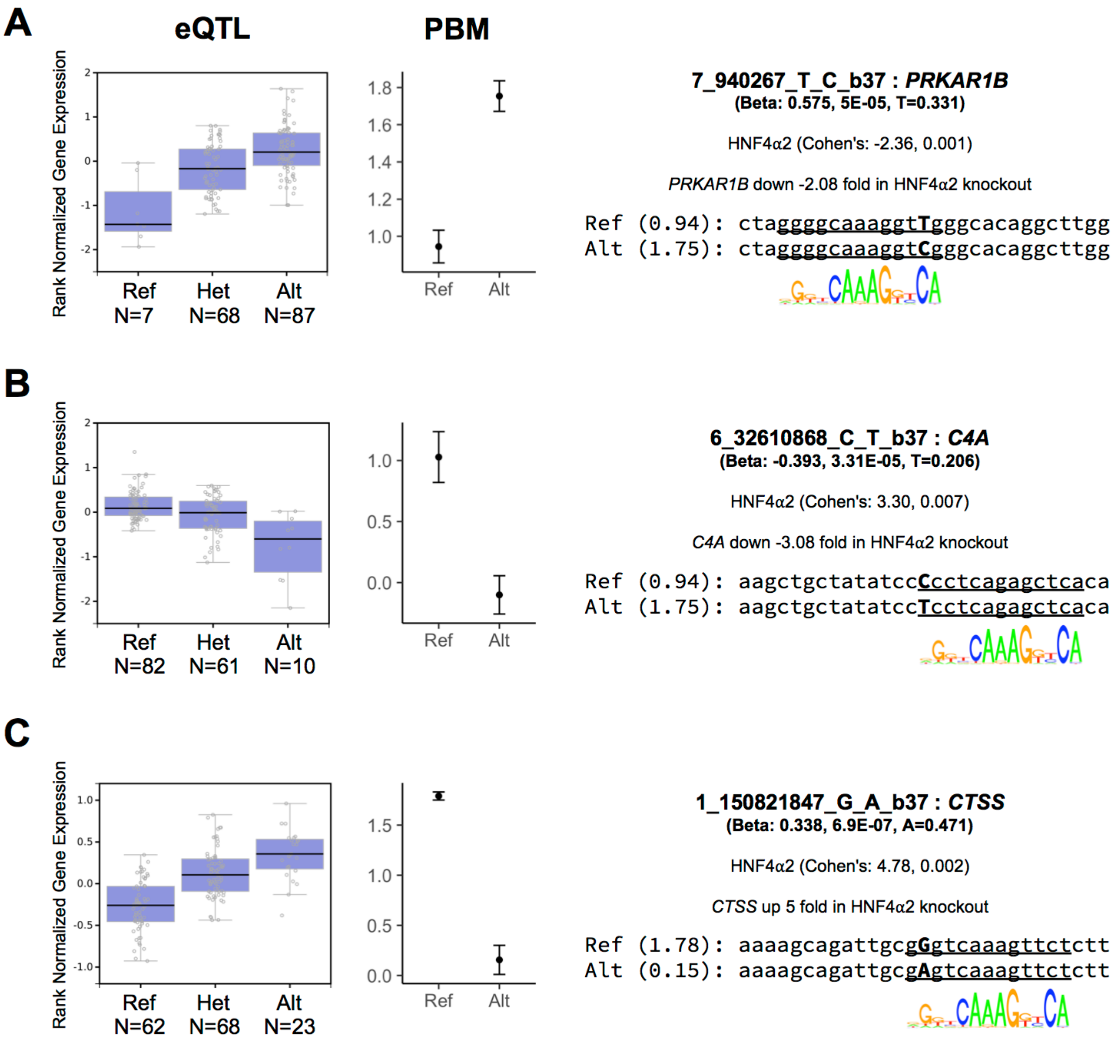

A
eQTL
PBM
Rank Normalized Gene Expression
2
1
0
-1
-2
Ref N=7
Het N=68
Alt N=87
1.8
1.6
1.4
1.2
1.0
Ref
Alt
7_940267_T_C_b37 : PRKAR1B
(Beta: 0.575, 5E-05, T=0.331)
HNF4α2 (Cohen's: -2.36, 0.001)
PRKAR1B down -2.08 fold in HNF4α2 knockout
Ref (0.94): ctaggggcaaaggtTgggcacaggcttgg
Alt (1.75): ctaggggcaaaggtCgggcacaggcttgg
B
Rank Normalized Gene Expression
2
1
0
-1
-2
Ref N=82
Het N=61
Alt N=10
1.0
0.5
0.0
Ref
Alt
6_32610868_C_T_b37 : C4A
(Beta: -0.393, 3.31E-05, T=0.206)
HNF4α2 (Cohen's: 3.30, 0.007)
C4A down -3.08 fold in HNF4α2 knockout
Ref (0.94): aagctgctatatccCcctcagagctcaca
Alt (1.75): aagctgctatatccTcctcagagctcaca
C
Rank Normalized Gene Expression
1.0
0.5
0.0
-0.5
-1.0
Ref N=62
Het N=68
Alt N=23
1.5
1.0
0.5
0.0
Ref
Alt
1_150821847_G_A_b37 : CTSS
(Beta: 0.338, 6.9E-07, A=0.471)
HNF4α2 (Cohen's: 4.78, 0.002)
CTSS up 5 fold in HNF4α2 knockout
Ref (1.78): aaaagcagattgcgGgtcaaagttctctt
Alt (0.15): aaaagcagattgcgAgtcaaagttctctt

Figure 3.8

Figure 3.8 Examples of aaSNP-eGenes for HNF4α2

(*A*) *Left*, normalized gene expression associated with genotype (reference, heterozygous, alternate) for eQTL linking 1_150821847_G_A_b37 to *CTSS*. *Middle*, average PBM binding scores represented as mean ± standard error of the mean. *Right*, Summary and motif analysis of aaSNP. Shown below eQTL id are GTEx derived stats for Beta/Slope and P-value, Cohen's effect size and padj associated with aaSNP, and Transcriptomine (TM) reported expression change of gene in HNF4α knockout. (*B*) as in (*A*) but with aaSNP-eQTL: 7_940267_T_C_b37 to *PRKAR1B*. (*C*) as in (*A*) but with aaSNP-eQTL: 6_32610868_C_T_b37 to *C4A*.

References

Becnel, L.B., Ochsner, S.A., Darlington, Y.F., McOwiti, A., Kankanamge, W.H., Dehart, M., Naumov, A., and McKenna, N.J. (2017). Discovering relationships between nuclear receptor signaling pathways, genes, and tissues in Transcriptomine. Sci. Signal. *10*.

Bogan, A.A., Dallas-Yang, Q., Ruse, M., Jr, Maeda, {. Yutaka, Jiang, G., Nepomuceno, L., Scanlan, T.S., Cohen, F.E., and Sladek, A.M. (2000). Analysis of Protein Dimerization and Ligand Binding of Orphan Receptor HNF4a. *302*, 831–851.

Bolotin, E., Liao, H., Ta, T.C., Yang, C., Hwang-Verslues, W., Evans, J.R., Jiang, T., and Sladek, F.M. (2010). Integrated approach for the identification of human hepatocyte nuclear factor 4alpha target genes using protein binding microarrays. Hepatology *51*, 642–653.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Jiang, G., Nepomuceno, L., Hopkins, K., and Sladek, F.M. (1995). Exclusive homodimerization of the orphan receptor hepatocyte nuclear factor 4 defines a new subclass of nuclear receptors. Mol. Cell. Biol. *15*, 5131–5143.

Kojetin, D.J., and Burris, T.P. (2014). REV-ERB and ROR nuclear receptors as drug targets. Nat. Rev. Drug Discov. *13*, 197–216.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Ochsner, S.A., Watkins, C.M., McOwiti, A., Xu, X., Darlington, Y.F., Dehart, M.D., Cooney, A.J., Steffen, D.L., Becnel, L.B., and McKenna, N.J. (2012). Transcriptomine, a web resource for nuclear receptor signaling transcriptomes. Physiol. Genomics *44*, 853–863.

Roshan-Moniri, M., Hsing, M., Butler, M.S., Cherkasov, A., and Rennie, P.S. (2014). Orphan nuclear receptors as drug targets for the treatment of prostate and breast cancers. Cancer Treat. Rev. *40*, 1137–1152.

Sladek, F.M. (2003). Nuclear receptors as drug targets: new developments in coregulators, orphan receptors and major therapeutic areas. Expert Opin. Ther. Targets *7*, 679–684.

Verzi, M.P., Shin, H., He, H.H., Sulahian, R., Meyer, C.A., Montgomery, R.K., Fleet, J.C., Brown, M., Liu, X.S., and Shivdasani, R.A. (2010). Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. Dev. Cell *19*, 713–726.

Viollet, B., Kahn, A., and Raymondjean, M. (1997). Protein kinase A-dependent phosphorylation modulates DNA-binding activity of hepatocyte nuclear factor 4. Mol. Cell. Biol. *17*, 4208–4219.

Wallerman, O., Motallebipour, M., Enroth, S., Patra, K., Bysani, M.S.R., Komorowski, J., and Wadelius, C. (2009). Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. Nucleic Acids Res. *37*, 7498–7508.

Chapter 4

Identification of Affinity Altering SNPs (aaSNPs) for PPARγ

Contributions from others:

Dr. Nina Titova – Extract preparation and application to PBMs

Abstract

Alzheimer’s disease is an age-associated neurodegenerative disease characterized by progressive loss of memory and cognition. Alzheimer’s is the most common form of dementia in North America with an estimated 5.4 million Americans afflicted in 2016 and is anticipated to grow to 13 million by the year 2050. Peroxisome proliferator-activated receptors (PPARs) are ligand-sensitive transcription factors and members of the nuclear receptor superfamily; they are promising drug targets for neurodegenerative disorders such as Parkinson’s, Alzheimer’s, Huntington’s, and ALS diseases. PPARγ is a member of the PPAR subfamily and has been shown to regulate lipid and glucose metabolism; agonists of PPARγ exhibit anti-inflammatory and antioxidant effects. With the growing rate of occurrence of Alzheimer’s disease, PPARγ agonists will be used against an increasingly wider population of individuals with rare and common genetic variations. In this chapter, we developed a high throughput DNA binding assay called protein binding microarrays (PBM) to better define the DNA sequences to which PPARs bind. We created a PPARγ PBM by data-mining single nucleotide polymorphisms (SNP) from PPARγ adipose tissue ChIP-seq and Alzheimer’s genome wide association studies (GWAS) in order to identify SNPs that alter the ability of PPARs to bind DNA, referred to as affinity altering SNPs (aaSNP). This PBM technology allowed us to identify >2,000 aaSNPs with a high degree of confidence (padj. ≤ 0.01), and a cluster of PPARγ aaSNPs within the *APOE* gene locus.

Introduction

The peroxisome proliferator-activated receptors (PPARα/δ/γ) are ligand-sensitive nuclear receptors (NR). PPARα/δ play a key role in fatty acid metabolism and energy homeostasis, while PPARγ plays a key role in insulin signaling and glucose metabolism. The PPARs have also been implicated in many human diseases, such as diabetes, cancer, lung disease, and several neurodegenerative disorders like Alzheimer's disease.

All three members of the PPAR family heterodimerize with retinoic acid receptor (RXR) to bind DNA sequences characterized as direct repeat 1 (DR1; AGGTCANAGGTCA) response elements; their endogenous ligands are lipid-derived substrates. In the absence of ligands, PPARs associate with co-repressor complexes to inhibit expression of target genes. Each member plays a role in fatty acid and glucose metabolism, but differ in their tissue-specificity and transcriptional activity. PPARα is primarily expressed in kidney, liver, muscle and heart tissues and is known to play a key role in fatty acid oxidation. In a fasted state, PPARα is activated by adipose-derived fatty acids promoting the synthesis of ketone bodies via fatty acid β-oxidation in the liver. PPARδ is ubiquitously expressed throughout the body where it promotes fatty acid metabolism and suppresses macrophage derived inflammation. PPARγ is primarily expressed in adipose tissue and is known to regulate adipocyte differentiation, fatty acid storage, and glucose metabolism.

PPARγ has been shown not only to be a master regulator of adipogenesis and lipid metabolism (Lee and Ge, 2014; Oger et al., 2014) but also a negative regulator of the cell cycle (Lin et al., 2007). PPARγ agonists have been shown to increase expression of tumor suppressor PTEN (Teresi et al., 2006), and activation of PPARγ inhibits proliferation of carcinoma cells (Borbath and Horsmans, 2008; Fukumoto et al., 2005; Grommes et al., 2006; Shappell et al., 2001). PPARγ agonists have also shown increased efficacy, partly due to anti-inflammatory effects, of various neurodegenerative diseases such as Parkinson's, Alzheimer's, and amyotrophic lateral sclerosis (ALS) (Gray et al., 2012; Heneka et al., 2011; Schintu et al., 2009).

Alzheimer's disease is a neurodegenerative disorder characterized by the gradual loss of memory and cognitive function (Salmon and Bondi, 2009; Serrano-Pozo et al., 2011). It is one of the most prevalent diseases in America affecting nearly 5.4 million adults and has become a serious health concern with the aging demographic and longer life spans world-wide (Lutz et al., 2008). There is growing evidence suggesting that individuals with type 2 diabetes have significantly increased risk of developing Alzheimer's, and *vice versa*. Inflammation, insulin resistance, and mitochondrial dysfunction are common pathological features of both Alzheimer's and type 2 diabetes. PPARγ is expressed at low levels in the brain in normal conditions, but there is evidence it can be up-regulated in Alzheimer's (de la Monte and Wands, 2006). Over the past decade PPARγ agonists have been used to treat the symptoms of Alzheimer's disease due to PPARγ's ability to increase insulin sensitivity and inhibit inflammation, and, as a result, improve cognition. While PPARγ agonists ameliorate Alzheimer's pathology by improving memory and cognition, they have not been able to cure the disease.

Materials and Methods

Preparation of nuclear extracts for Protein Binding Microarrays

Nuclear extracts were prepared as in Chapter 2. Extracts were prepared from COS-7 cells transiently co-transfected with Flag-tagged human PPAR (PPARα (NM_005036), PPARδ (NM_006238) or PPARγ (NM_005037)) and untagged human RXRα. Ligands were added to media 2h before harvesting: rosiglitizone (1μM, Cayman Chemicals) for PPARγ; GW7647 for PPARα (0.1μM, Cayman Chemicals) and GW0742 for PPARδ (0.1μM, Cayman Chemicals). The RXRa ligand, 9-cis retinoic acid (Sigma) was simultaneously also added to the cells at the same concentration as each PPAR ligand.

Protein Binding Microarrays

Protein Binding Microarrays (PBMs) were designed and processed as in Chapter 2 and in Bolotin et al. (Bolotin et al., 2010). Protease inhibitor (Sigma) and ligands were added at all incubation steps starting from hybridization. 1μM of each ligand was added to all washes. After purification, PPAR protein complexes were applied to arrays and incubated for 15h at 4°C, arrays were washed 3x for 2 min 30 sec each with PBS plus 0.1% Tween 20. Mouse monoclonal anti-Flag antibody (Ab) (Sigma, #F3165) diluted 1:100 in PBS buffer plus 2% non-fat milk, 0.1% Tween 20 were applied directly to the slide and incubated for 48h at 4°C, followed by a conjugated secondary Ab (GαM IgG [H+L] DyLight 550, Pierce #84540) diluted 1:50 (as described above) and then incubated for 4h at room temperature. Three washes, 2 min 30 sec each in PBS plus 0.1% Tween 20 were performed after each antibody incubation. The anti-Flag Ab is likely detecting PPAR:RXR heterodimeric complexes as PPAR in the absence of RXR does not bind DNA well.

PPARγ ChIP-SNP PBM Design

The PPARγ ChIP-SNP PBM was designed by extracting all common SNPs (1% of the population or 100% non-reference, dbSNP v142) from eight human PPARγ ChIP-seq datasets -- five human adipocyte tissue samples and three human adipocyte cell lines, two from SGBS cells and one from hASC cells (Mikkelsen et al., 2010; Schmidt et al., 2011; Soccio et al., 2011, 2015). Human adipose samples were all obese (BMI>30) females, aged 26-57 years, and two had pre-diabetes. ChIP peaks were filtered with length ≤ 800 nt and genetic variants were selected from a 170-nt window around the peak center or a 170-nt window in the middle of peak if no peak centers were reported. Reference and alternate alleles, up to four total alleles, were selected as 29-mers with 14-nt flanking genomic DNA around the variant to accommodate a full direct repeat 1 (DR1; 13nt, AGGTCAAAGGTCA) binding site. Two Alzheimer's genome wide association studies (GWAS) databases were also mined for genetic variants associated with Alzheimer's disease: The Late Onset Alzheimer's Disease (LOAD, 1e-5 pval) and International Genomics of Alzheimer's Project (IGAP, 1e-10 pval) totalling roughly 965 variants (Lambert et al., 2013; Naj et al., 2011). Finally, 600 *in vivo*-identified mouse aaSNPs for PPARγ were included in the PBM to compare to *in vitro* PPARγ binding (Soccio et al., 2015). These aaSNPs were derived from adipose PPARγ ChIP-seq datasets from two distantly related mouse strains that show differences in susceptibility to insulin resistance and obesity (C57Bl/6J and 129S1/SvlmJ). As a negative control, 400 random 29-mer DNA sequences generated as in Chapter 3 were included in the design. All alleles plus negative controls were spotted in quadruplicate on the slide, resulting in a one-million spot design.

Graphical and statistical analysis

Due to the technical variability of the PBM slides, some quality control and normalization steps were necessary. Slides were spatial normalized using the MANOR library in R. Averages and standard deviation were calculated across replicated spots. To identify test sequences with potential outliers, a coefficient of variation (cv) adjusted for small sample sizes was calculated with non-log values as follows:

$$\frac{sd}{avg} * (1 + \left(\frac{1}{4n}\right))$$

Any set of replicated spots with cv ≥ 0.5 are selected for outlier removal with a custom Python script. The distance from lowest to second highest score, and distance from highest to second lowest score are measured and compared within each group of four replicates. Outliers were removed with the following criteria; lowest value removed if first comparison was the largest, highest value removed if second comparison was the largest. Averages, standard deviation, and cv of the remaining three replicates are returned to the dataset.

To calculate binding levels of each test sequence in the experiment the entire population of negative controls are averaged together and used as a control population. Each individual test sequence is measured against the control with a one-tailed Student's T-test. Reported P-values are corrected for multiple hypothesis testing using the Benjamini-Hochberg ("fdr") option with the p.adjust function from the stats library in R. These values are reported as "padj." or adjusted P-values.

To calculate affinity altering SNPs (aaSNPs), the PBM binding score of the major and minor (ref and alt) allele averages were compared with a two-tailed Student's T-test, using standard deviation from control population as pooled variance. P-values were corrected with Benjamini-Hochberg option of the p.adjust function as above. To measure the effect size between the two alleles a Cohen's D value was calculated by dividing the difference of means by the standard deviation of the pooled negative controls.

All scatterplots were generated in R with ggplot2 library. All receptors applied within a single design were quantile normalized with 'preprocessCore' package in R, before plotting. Since aaSNPs were calculated prior to quantile normalization their significance was matched to quantile normalized values before highlighting on the scatterplots. Venn diagrams were generated with the 'VennDiagram' package in R.

Results

Identification of affinity altering SNPs (aaSNPs) in PPAR DNA binding sites

To identify genetic variants that impact the DNA binding affinity of the PPAR subfamily of NRs, we designed a 1 million (1M) spot PBM. Approximately 108,000 Common SNPs (≥1% in population or 100% non-reference) from dbSNP v142 were extracted from a 200-nt window (±100 nt) around the peak center of eight PPARγ ChIP-seq datasets: three from human adipocyte cell lines and five from human adipocyte tissue samples. Roughly 965 genetic variants associated with Alzheimer's were included from two Alzheimer's GWAS datasets, the Late Onset Alzheimer's Disease (LOAD; 1e-5 pval) and International Genomics of Alzheimer's Project (IGAP; 1e-10 pval). Finally, 600 *in vivo* mouse PPARγ aaSNPs (Soccio et al., 2015) were included as controls to compare to the *in vitro* binding. Each test sequence is replicated on the slide in quadruplicate to allow for the single removal of an outlier in case of technical variability of the slide (Fig 4.1A).

The PPAR subfamily of NRs (α,δ,γ) were applied to the PBM as heterodimers with RXRα. Results of the total number of significantly bound alleles (Binders, padj. ≤ 0.01, 14,470 to 24,079) and number of affinity altering SNPs (aaSNPs, padj. ≤ 0.01 and Cohen's ≥ 2, 654 to 2,108) are given in Fig 4.1B. PPARγ showed the highest ratio of aaSNPs/binders with 8.755% (Fig 4.1B), likely because the datasets mined were all PPARγ ChIP-seq with the exception of the GWAS variants. The aaSNPs for each receptor were compared against the other two family members: considerable overlap was found. PPARα and PPARδ had 256 common aaSNPs (39.14% of the 654 total aaSNPs for PPARα); PPARα and PPARγ had 214 common aaSNPs (32.7% of PPARα), while PPARγ and PPARδ share more than twice that at 507 (34.4% of PPARδ and 24.1% of PPARγ) (Fig 4.1B).

To visualize the aaSNPs distributed between NRs, scatterplots were generated showing the entire population of test sequences for each receptor and highlighted significant aaSNPs. PPARα and PPARδ show the most similar distribution of PBM binding scores across the entire design (Fig 4.2A). It is interesting to note that sequences with the highest PBM binding scores were bound equally well by both receptors, resulting in a large number of common aaSNPs (magenta spots). Both NRs show small groupings of unique aaSNPs in the range of (0,2) that were not bound well by the other receptor. PPARγ shows a very different binding profile compared to both PPARα and PPARδ as seen by the wide distribution of PBM binding scores (Fig 4.2B/C). Both PPARα- and PPARδ-specific aaSNPs show large clusters binding moderately well (0,2) on the x-axis with very poor binding for PPARγ on the y-axis. Nonetheless, even with these PPAR comparisons many of the considerable number of common aaSNPs have high PBM binding scores for both receptors (Fig 4.2B/C). The close similarity in DNA binding specificity for PPARα and PPARδ, as well as the greater difference with PPARγ, could be due to differences in the respective DNA binding domains (DBD) -- PPARα differs from PPARδ by nine amino acids while both PPARα and PPARδ differ from PPARγ by 11 aa (Fig 4.S1). Interestingly, most of the differences with PPARγ are in the second zinc finger which is involved in heterodimerization (Tsai and O'Malley, 1994). We have shown previously that just one amino acid change in a critical region of a NR DBD can lead to considerable changes in binding of a large number of sequences that are easily detected in the PBMs (Fang et al., 2012).

PBM aaSNP identification is validated by in vivo results

To measure the ability of the high-throughput PBM methodology to identify aaSNPs *in vitro* that might be relevant *in vivo*, 600 *in vivo* mouse PPARγ aaSNPs were included in the design. Adipose PPARγ ChIP-seq datasets from two distantly related mouse strains that show differences in susceptibility to insulin resistance and obesity (C57Bl/6J and 129S1/SvlmJ) were compared to identify allele-specific binding that disrupted predicted PPARγ binding *in vivo* (Soccio et al., 2015). Reference and alternate alleles along with 14-nt flanking sequences from the mouse reference genome (mm9) were spotted on the PPARγ ChIP-SNP-PBM. The human PPARγ PBM successfully identified 125 (21.18%) of the *in vivo* mouse PPARγ aaSNPs (padj. ≤ 0.01 and Cohen's ≥ 2) (Fig 4.3A). This is a considerable success rate considering that the PBM was probed with human PPARγ expressed in monkey kidney cells (COS-7) while the *in vivo* aaSNPs were identified in the context of mouse adipose tissue, although the human and mouse PPARγ are 99% identical and the DBD is 100% identical.

In contrast, human PPARα and PPARδ identified fewer of the murine PPARγ aaSNPs: 36 (6.1%) and 61 (10.33%), respectively.

Visualizing the shared mouse PPARγ aaSNPs for each receptor on scatterplots of PBM binding scores for the human receptors shows that many of the PPARα-specific mm9 aaSNPs (yellow lines) had very high binding scores for PPARγ, but did not show enough of a difference between the alleles to be considered a shared aaSNP (Fig 4.3B).

Considering the source of these aaSNPs was from PPARγ ChIP peaks it is not surprising to find so many PPARγ-specific aaSNPs essentially non-binding for PPARα as seen by the cluster of vertical blue aaSNPs in the left side of the plot. Similar trends in binding can also be seen between PPARγ and PPARδ (Fig 4.3C). In contrast, PPARα and PPARδ, with such similar binding profiles, show many more shared aaSNPs (magenta lines) as well as many unique aaSNPs (yellow and red lines) along the diagonal with trajectories similar to the shared aaSNPs but not quite reaching the threshold of significance.

Alzheimer's GWAS aaSNPs

Between the two Alzheimer's GWAS studies there was a total of 965 genetic variants that were tested in the PBM for their ability to disrupt DNA binding of all three PPAR receptors. Using relatively strict significance thresholds (padj. ≤ 0.01 and Cohen's ≥ 2) we successfully identified 45 aaSNPs (PPARα: 5; PPARδ: 17; PPARγ: 23), of which seven were shared with more than one receptor, resulting in a total of 36 unique aaSNPs that alter the ability of one or more PPAR to bind DNA (4.6% of the 965 total) (Fig 4.4A). Annotating each variant to find the nearest TSS we noticed that 19 of these

GWAS aaSNPs were located on chromosome 19, and of those 19, there were 10 aaaSNPs within a 100 kb window of *APOE*, the gene most highly linked to Alzheimer's (Yu et al., 2014) (Fig 4.4B). When we overlay these data with histone marks from the ENCODE project, three of these variants fall within H3K4Me1 marks, which are often found near regulatory elements (Fig 4.4B, blue arrows). Looking more closely at the two aaSNPs closest to the *APOE* TSS we find that the PPARγ-specific aaSNP rs440446, which falls within the first intron of *APOE,* also sits directly within H3K4Me1, H3K27Ac, and H3K4Me3 histone marks that are often found near active regulatory elements and promoters (Fig 4.4C). The sequence of the minor allele (C) of rs440446 bound with a relatively high PBM score of 3.8, while the major allele (G) did not (score = 0.24) (padj 0.001, Cohen's D 4.42). This suggests that 37.3% of the population carrying the minor allele may have increased PPARγ binding at the *APOE* locus, resulting in greater expression of *APOE* and hence less Alzheimer's. Indeed, Transcriptomine shows that pioglitazone, a ligand for PPARγ, increases the expression of *Apoe* 7.2-fold while rosiglitazone, another PPARγ ligand, increases it 5.9-fold in mouse adipocytes; rosiglitazone also increased expression of *APOE* in human dendritic cells by 3.6-fold. Whether rs440446 plays a role in the PPARγ-directed regulation of *APOE* in humans remains to be determined. Transcriptomine also shows that *Apoe* expression is decreased in a PPARδ KO, indicating that *Apoe* is also a target of PPARδ: our PBM data suggests that it might be a direct target.

We next examined whether either of the aaSNPs (rs35568738 or rs440446) is listed as an eQTL in GTEx. While rs35568738 is not an eQTL in GTEx v7, rs440446 is indeed shown as an eQTL with effect size (Slope) of -0.23 and a p-value of 1.1e-10 (Fig 4.5). The alternate allele (G) has a lower level of expression than the reference allele (C) in skin, consistent with PPARγ binding better to the C allele (Fig 4.4). While *APOE* expression in the skin is not likely to be relevant to Alzheimer's, this result nonetheless shows that an aaSNP (rs440446) that is linked to Alzheimer's decreases PPARγ binding *in vitro* and correlates with reduced expression of *APOE* in human samples. The only caveat is that we do not know if the altered expression in the skin is due to changes in PPARγ binding in that tissue -- but PPARγ as well as *APOE* are expressed in the skin but only at very low levels.

Discussion

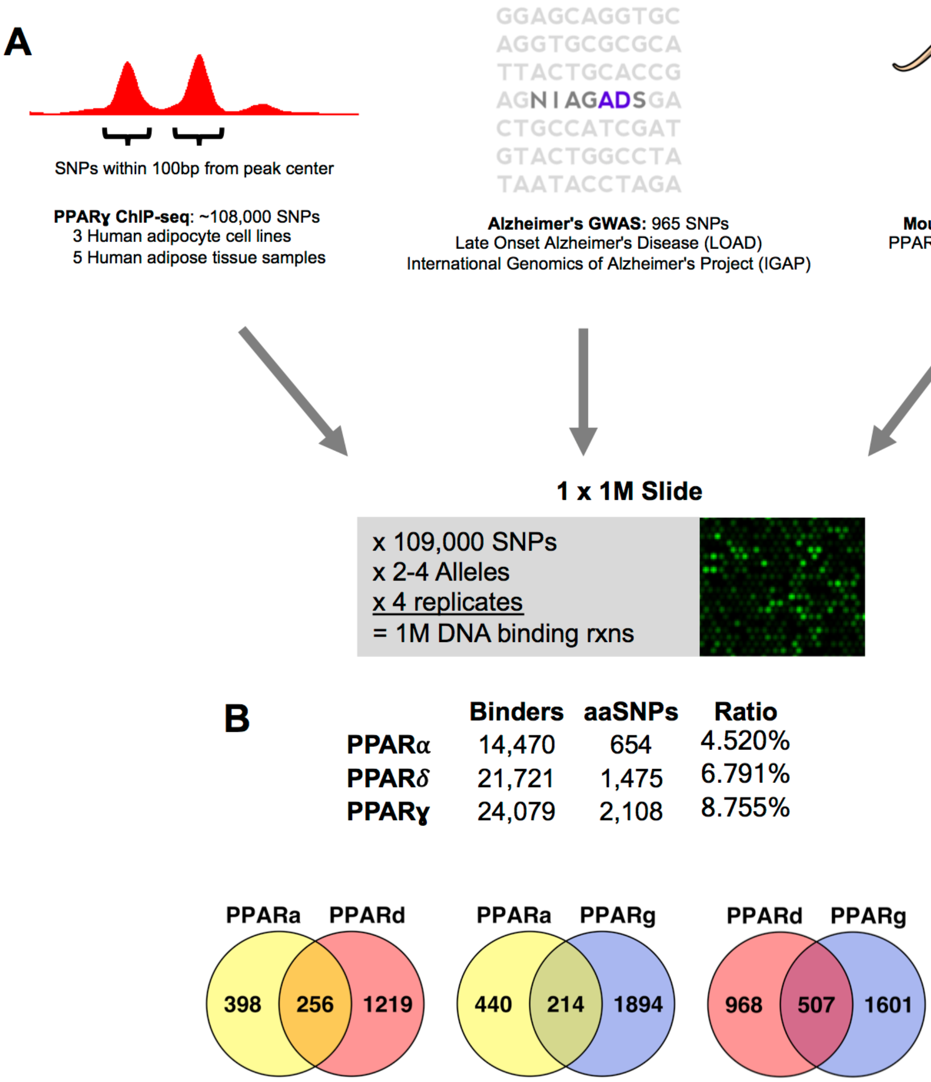
PPARγ is a promising drug target to treat neurodegenerative disease, especially one as prevalent as Alzheimer's. Currently, more than 5 million adults in the United States are diagnosed with Alzheimer's which means that therapeutics administered against the disease will be taken by a broad range of individuals with rare and common genetic variations. We already have a basic understanding of the effect of PPARγ agonists in Alzheimer's disease as well as thousands of SNPs identified by GWAS studies associated with the disease. Having a better understanding of how common and rare genetic variations impact PPARγ DNA binding will have implications not only for determining who is most susceptible to getting Alzheimer's but also for who is most likely to respond to PPARγ agonists. The knowledge of which genetic variants are capable of disrupting PPARγ recruitment to promoters and enhancers will help further the field of personalized medicine and set the stage for revealing how NR therapeutics result in altered phenotypes in individuals with Alzheimer's disease.

Our results show that while all three members of the PPAR family of NRs (PPARα/δ/γ) are affected by aaSNPs found in PPARγ ChIP peaks, there is a clear distinction between the overall DNA binding affinity of PPARγ compared to the other two members, PPARα and PPARδ. These receptors have highly conserved DNA binding domains (DBD) with high conservation of the first zinc finger domain, typically involved in DNA binding (Fig 4.S1). However, PPARγ contains a two amino acid difference in the second zinc finger that was found some time ago to be involved in heterodimerization which could impact DNA binding (Tsai and O'Malley, 1994). PBMs could be used to determine whether those amino acids are indeed involved in DNA binding specificity across thousands of sequences, as we did previously for HNF4α (Fang et al., 2012).

The analysis of *in vivo* derived aaSNPs from differential murine PPARγ ChIP peaks identified fewer aaSNPs *in vitro* with the PBM technology than expected. Obviously, there other conditions present in the adipose tissue samples from C57BL/6J and 129S1/SvlmJ that could be contributing to differences in PPARγ chromatin binding that are not represented in our PBMs, including DNA methylation and other epigenetic factors. In fact, many murine aaSNPs that failed to be identified with the PPARγ-ChIP-SNP PBM show low PBM binding scores for both alleles, suggesting that some of the altered PPARγ ChIP peaks may have been due to PPARγ recruitment to the DNA via other TFs. It would be of interest to examine other TFs known to work with PPARγ to regulate adipogenesis, such as RXRα and C/EBPα (Siersbæk et al., 2010), for DNA binding affinity on the PPARγ ChIP-SNP PBM to see if any additional murine aaSNPs can be verified.

The analysis of GWAS-identified SNPs associated with Alzheimer's resulted in 3.73% of the variants showing the capability to alter DNA binding affinity of any PPAR isoform. In comparison to Disease SNP-PBM results from Chapter 3, these results are in line with the expected ratio of aaSNPs to binders identified from datasets with no prior *in vivo* binding data. Out of the 36 total aaSNPs identified, a total of 19 variants were associated with chromosome 19 and 9 variants associated with chromosome 11. While the localization bias of these data may be a result of the GWAS studies identifying clusters of variants of regions associated with Alzheimer's disease phenotypes, the fact that so many of them are capable of disrupting PPAR DNA binding affinity may not be coincidental.

While these data suggest that PPARγ aaSNPs may play a role in Alzheimer's disease, specifically involving regulation of *APOE*, more *in vivo* analysis will be needed to causally associate these aaSNPs with PPARγ transcriptional activity. The first step in an analysis of this nature would be to utilize CRISPR technology with human cell lines to sequence and point mutate the aaSNP of interest, followed by ChIP and luciferase assays to show that the aaSNP does in fact impact PPARγ DNA binding in normal cell conditions and that as a result expression levels of *APOE* would indeed be affected.



	Binders	aaSNPs	Ratio
PPARα	14,470	654	4.520%
PPARδ	21,721	1,475	6.791%
PPARɣ	24,079	2,108	8.755%

Figure 4.1

Figure 4.1. Design of PPARγ ChIP-SNP PBM

(*A*) PPARγ ChIP-SNP PBM design. SNPs were extracted from PPARγ ChIP-seq datasets derived from 5 adipocyte samples and 3 adipose cell lines, along with genetic variants from two Alzheimer's GWAS studies, and 600 PPARγ *in vivo* mouse aaSNPs. All common variants (2-4 alleles) from dbSNP v142 spotted in quadruplicate for a 1 million spot design. (*B*) *Top*, all three human PPARs were applied to this design (PPARα/δ/γ) as NEs from COS-7 cells ectopically expressing PPAR and RXRα; shown are total number of significantly bound alleles (padj. ≤ 0.01) and total number of affinity altering SNPs (aaSNP, padj. ≤ 0.01 & Cohen's ≥ 2). Cohen's D statistic was calculated using the standard deviation of 400 random controls. The ratio of aaSNPs to total binders is also reported. *Bottom*, Venn diagrams comparing aaSNPs identified for each NR. A difference of binding between reference and alternate alleles is counted as a single aaSNP measurement (padj. ≤ 0.01 & cohen's ≥ 2).

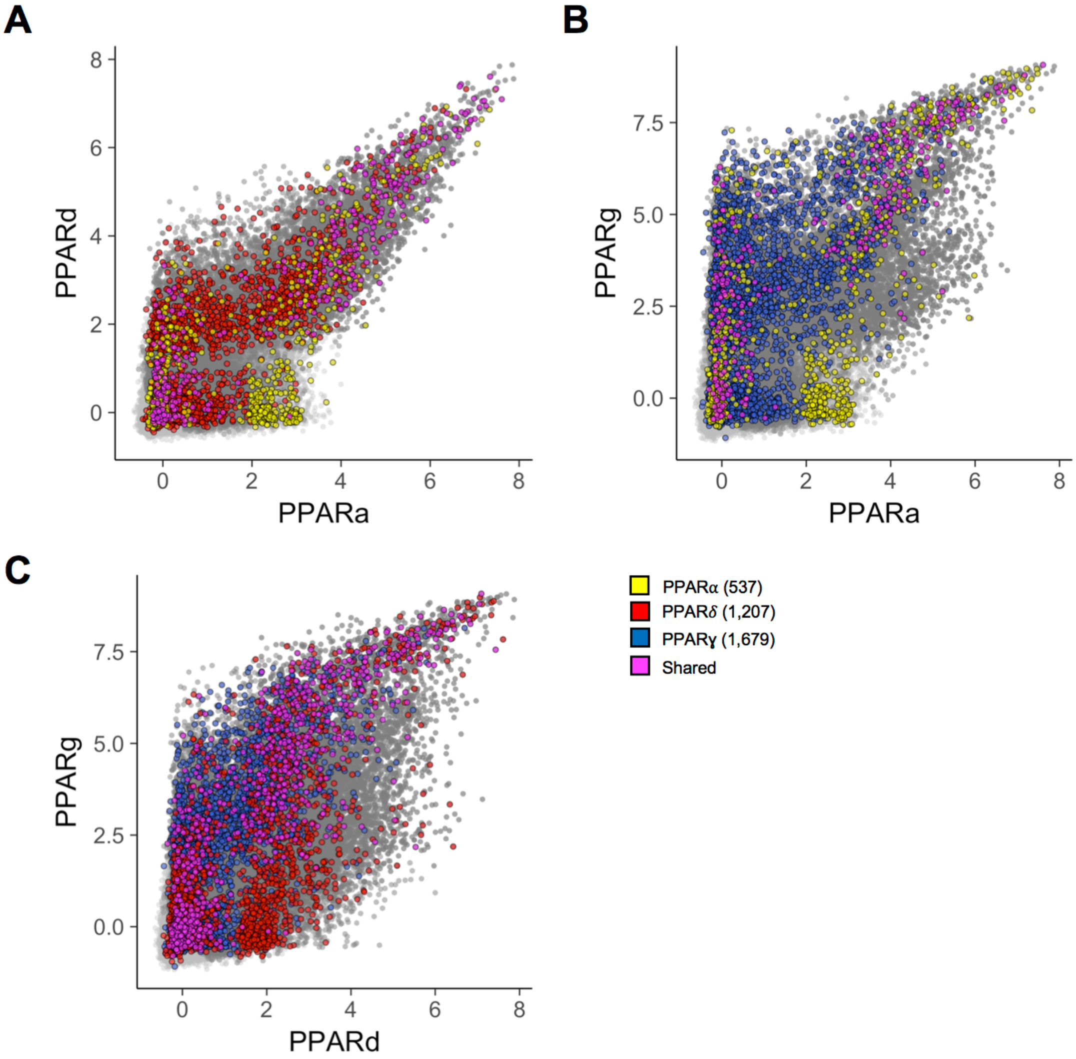

A
PPARd
8
6
4
2
0
0
2
4
6
8
PPARa
B
PPARg
7.5
5.0
2.5
0.0
0
2
4
6
8
PPARa
C
PPARg
7.5
5.0
2.5
0.0
0
2
4
6
8
PPARd
PPARα (537)
PPARδ (1,207)
PPARγ (1,679)
Shared

Figure 4.2

Figure 4.2. Visualizing aaSNPs within the PPAR family

(*A*) Scatterplots of quantile normalized PBM binding scores. Every spot is the average PBM binding score of a single allele. All binders (padj. ≤ 0.01) plotted in dark grey. Best and worst binding alleles for aaSNPs are highlighted designated in legend. Shared aaSNPs are highlighted in magenta.

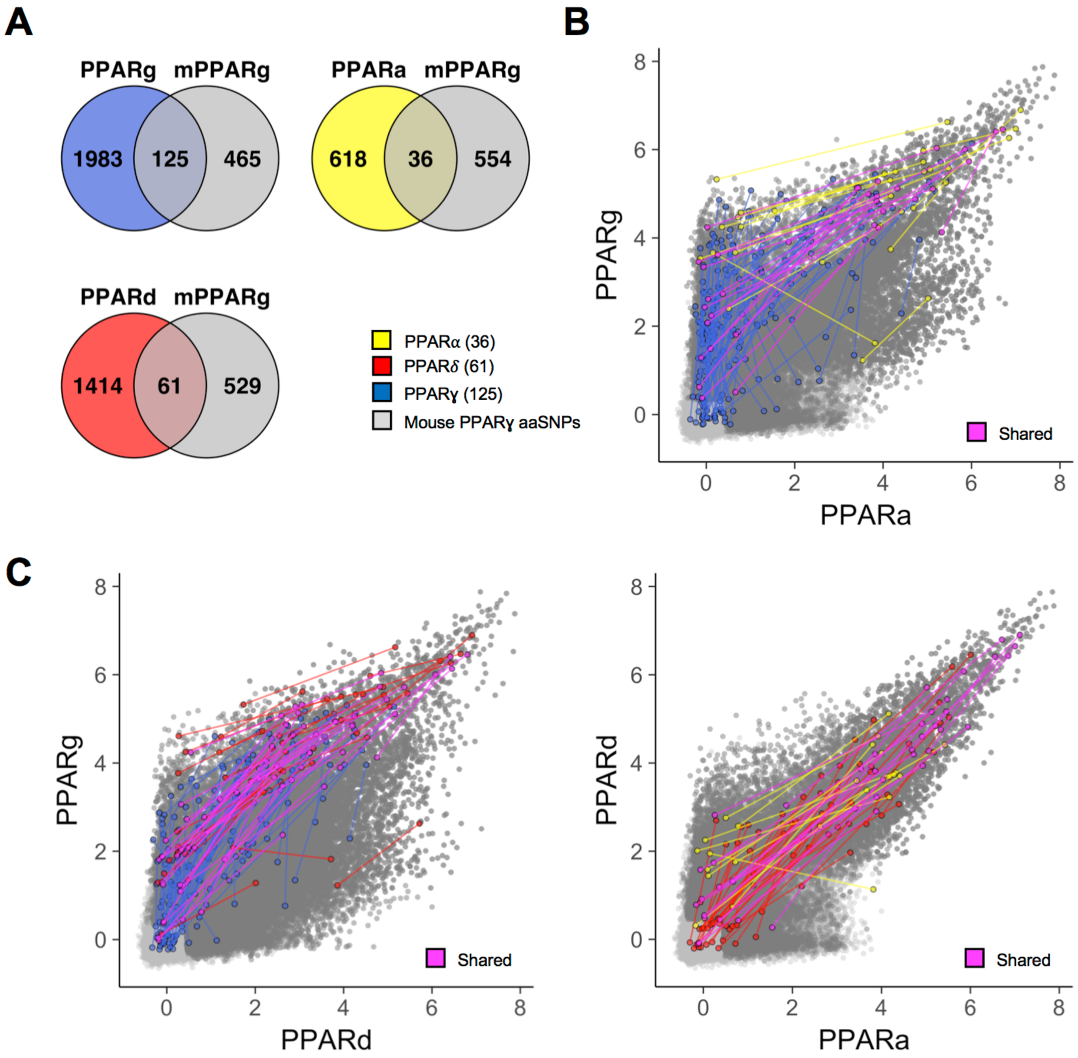

A
PPARg
mPPARg
1983
125
465
PPARa
mPPARg
618
36
554
PPARd
mPPARg
1414
61
529
PPARα (36)
PPARδ (61)
PPARɣ (125)
Mouse PPARɣ aaSNPs
B
PPARg
0
2
4
6
8
PPARa
Shared
C
PPARg
PPARd
Shared
PPARd
PPARa
Shared

Figure 4.3

Figure 4.3. *In vivo* aaSNPs verified with *in vitro* PBM methods

(*A*) Venn diagrams showing total aaSNPs (padj. ≤ 0.01 and Cohen's ≥ 2) for each receptor compared to all mouse PPARγ (mm9) *in vivo* aaSNPs from (Soccio et al., 2015). (*B*) Scatterplot of quantile normalized PBM binding scores with highlighted aaSNPs found in the mm9 *in vivo* dataset. Lines link best and worst alleles of each aaSNP. Shared aaSNPs are highlighted in magenta. (*C*) As in (*B*) but with different PPAR comparisons.

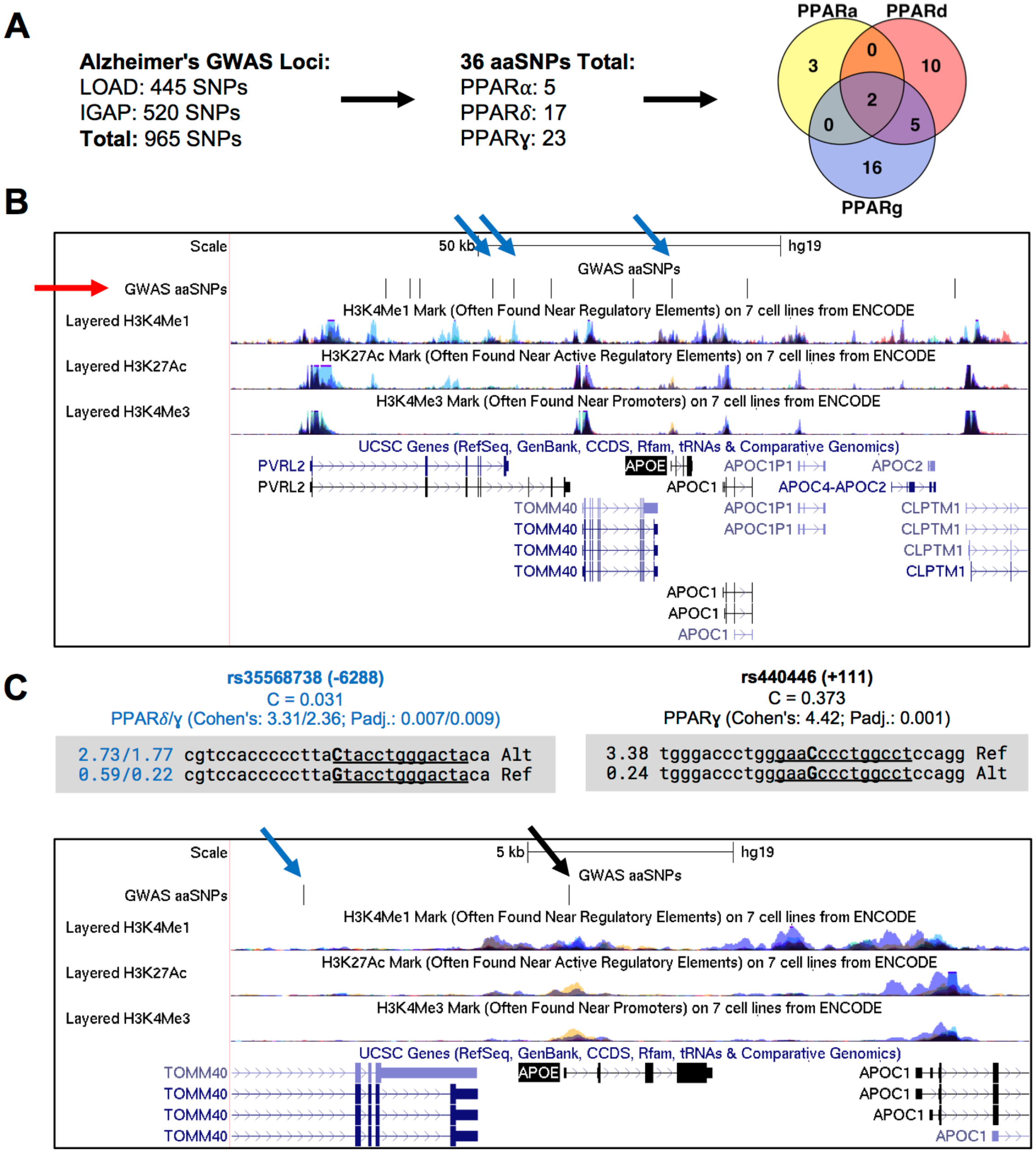

A
Alzheimer's GWAS Loci:
LOAD: 445 SNPs
IGAP: 520 SNPs
Total: 965 SNPs
36 aaSNPs Total:
PPARα: 5
PPARδ: 17
PPARγ: 23
PPARa
PPARd
PPARg
3
0
10
2
0
5
16
B
Scale
50 kb
hg19
GWAS aaSNPs
GWAS aaSNPs
Layered H3K4Me1
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PVRL2
PVRL2
APOE
APOC1P1
APOC2
APOC1
APOC4-APOC2
TOMM40
TOMM40
TOMM40
TOMM40
APOC1P1
APOC1P1
CLPTM1
CLPTM1
CLPTM1
CLPTM1
APOC1
APOC1
APOC1
C
rs35568738 (-6288)
C = 0.031
PPARδ/γ (Cohen's: 3.31/2.36; Padj.: 0.007/0.009)
2.73/1.77 cgtccacccccttaCtacctgggactaca Alt
0.59/0.22 cgtccacccccttaGtacctgggactaca Ref
rs440446 (+111)
C = 0.373
PPARγ (Cohen's: 4.42; Padj.: 0.001)
3.38 tgggaccctgggaaCccctggcctccagg Ref
0.24 tgggaccctgggaaGccctggcctccagg Alt
Scale
5 kb
hg19
GWAS aaSNPs
GWAS aaSNPs
Layered H3K4Me1
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
TOMM40
TOMM40
TOMM40
TOMM40
APOE
APOC1
APOC1
APOC1
APOC1

Figure 4.4

Figure 4.4. Alzheimer's GWAS aaSNPs disrupt PPAR binding the *APOE* gene locus (*A*) Counts of total aaSNPs (padj. ≤ 0.01 and Cohen's ≥ 2) identified from GWAS datasets, and Venn diagram of shared and unique variants between PPARs. (*B*) UCSC Genome browser view of *APOE* gene locus displaying locations of significant GWAS aaSNPs identified in this study in the top track (red arrowhead), three layered histone mark tracks across 7 cell lines from the ENCODE project, and UCSC gene list with *APOE* highlighted in black. (*C*) UCSC Genome browser view as in (*B*) of 20 kb window around *APOE* TSS. Highlighted in blue an aaSNP for PPARδ and PPARγ. Shown is the dbSNP id and distance to TSS along aaSNP values for PPARδ and PPARγ respectively. Shown below are PBM binding scores and test sequences for major and minor alleles, and underlined is the best predicted DR1 binding site. Similarly, highlighted in black is a PPARγ-specific aaSNP.

Single-Tissue eQTLs for 19_45409167_C_G_b37
Data Source: GTEx Analysis Release V7 (dbGaP Accession phs000424.v7.p2)

rs440446

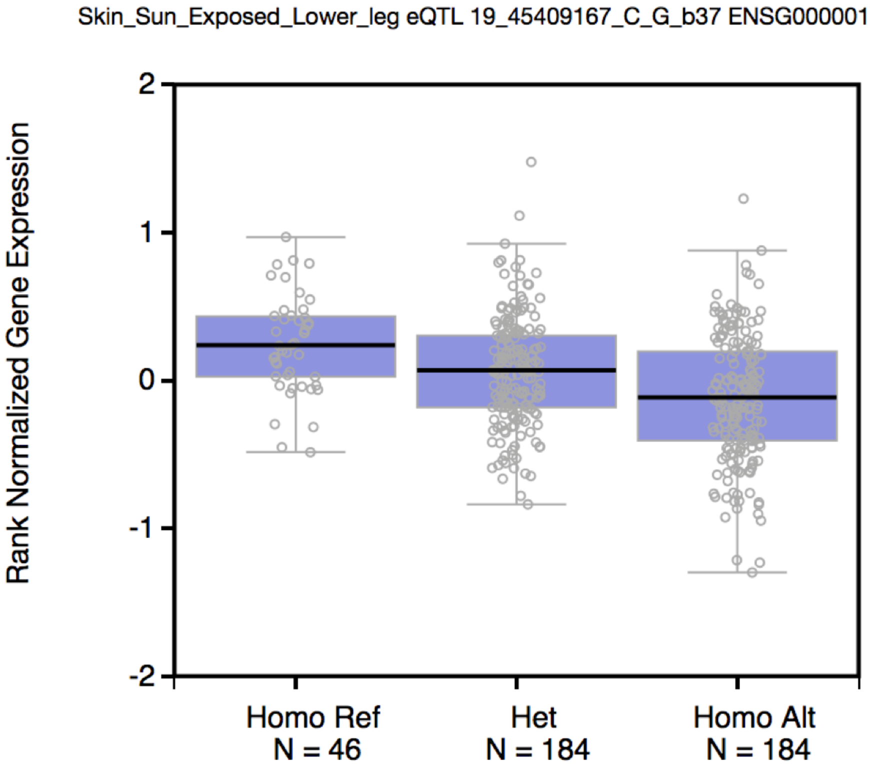
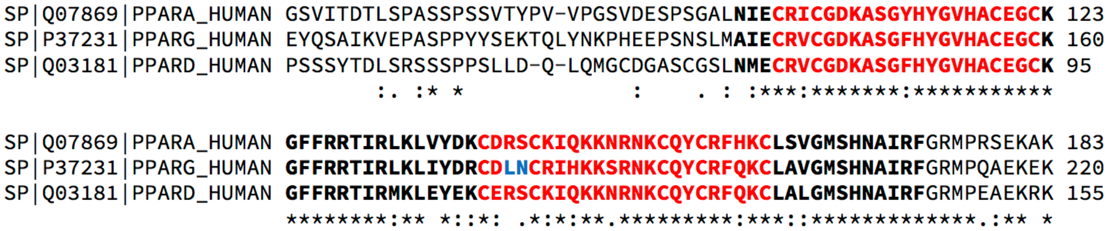


Figure 4.5

Figure 4.5. Alzheimer's GWAS aaSNP rs440446 is an eQTL in GTEx.

Screenshot from GTEx showing rs440446 as an eQTL in skin. No other tissue was found to have a significant eQTL for this SNP.



```
SP|Q07869|PPARA_HUMAN GSVITDTLSPASSPSSVTYPV-VPGSVDESPSGALNIECRICGDKASGYHYGVHACEGCK 123
SP|P37231|PPARG_HUMAN EYQSAIKVEPASPPYYSEKTQLYNKPHEEPSNSLMAIECRVCGDKASGFHYGVHACEGCK 160
SP|Q03181|PPARD_HUMAN PSSSYTDLSRSSSPPSLLD-Q-LQMGCDGASCGSLNMECRVCGDKASGFHYGVHACEGCK 95
                             :. :* *             :    . : :***:*******:***********

SP|Q07869|PPARA_HUMAN GFFRRTIRLKLVYDKCDRSCKIQKKNRNKCQYCRFHKCLSVGMSHNAIRFGRMPRSEKAK 183
SP|P37231|PPARG_HUMAN GFFRRTIRLKLIYDRCDLNCRIHKKSRNKCQYCRFQKCLAVGMSHNAIRFGRMPQAEKEK 220
SP|Q03181|PPARD_HUMAN GFFRRTIRMKLEYEKCERSCKIQKKNRNKCQYCRFQKCLALGMSHNAIRFGRMPEAEKRK 155
                      ********:** *::*: .*:*:**.*********:***::*************.:** *
```

Figure 4.S1

Figure 4.S1. Alignment of DNA Binding Domain of human PPARs.

Shown is the conserved 64-amino acid (aa) DNA binding domain of human PPARγ (NP_001341596), PPARα (NP_001001928) and PPARδ (NP_001165289). Underline, DNA recognition sequence in first zinc finger. Red, aa that differ among the NRs. The DNA recognition sequence is critical for distinguishing the half-site specificity of HNF4α, RXRα and COUP-TF (Fang et al., 2012). It is identical amongst all three human PPARs suggesting that one or more of the other variations noted in red are responsible for the differences in DNA binding specificity noted in the PBMs.

References

Bolotin, E., Liao, H., Ta, T.C., Yang, C., Hwang-Verslues, W., Evans, J.R., Jiang, T., and Sladek, F.M. (2010). Integrated approach for the identification of human hepatocyte nuclear factor 4alpha target genes using protein binding microarrays. Hepatology *51*, 642–653.

Borbath, I., and Horsmans, Y. (2008). The Role of PPARgamma in Hepatocellular Carcinoma. PPAR Res. *2008*, 209520.

Fang, B., Mane-Padros, D., Bolotin, E., Jiang, T., and Sladek, F.M. (2012). Identification of a binding motif specific to HNF4 by comparative analysis of multiple nuclear receptors. Nucleic Acids Res. *40*, 5343–5356.

Fukumoto, K., Yano, Y., Virgona, N., Hagiwara, H., Sato, H., Senba, H., Suzuki, K., Asano, R., Yamada, K., and Yano, T. (2005). Peroxisome proliferator-activated receptor delta as a molecular target to regulate lung cancer cell growth. FEBS Lett. *579*, 3829–3836.

Gray, E., Ginty, M., Kemp, K., Scolding, N., and Wilkins, A. (2012). The PPAR-gamma agonist pioglitazone protects cortical neurons from inflammatory mediators via improvement in peroxisomal function. J. Neuroinflammation *9*, 63.

Grommes, C., Landreth, G.E., Sastre, M., Beck, M., Feinstein, D.L., Jacobs, A.H., Schlegel, U., and Heneka, M.T. (2006). Inhibition of in vivo glioma growth and invasion by peroxisome proliferator-activated receptor gamma agonist treatment. Mol. Pharmacol. *70*, 1524–1533.

Heneka, M.T., Reyes-Irisarri, E., Hüll, M., and Kummer, M.P. (2011). Impact and Therapeutic Potential of PPARs in Alzheimer’s Disease. Curr. Neuropharmacol. *9*, 643–650.

Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease. Nat. Genet. *45*, 1452–1458.

Lee, J.-E., and Ge, K. (2014). Transcriptional and epigenetic regulation of PPARγ expression during adipogenesis. Cell Biosci. *4*, 29.

Lin, M.S., Chen, W.C., Bai, X., and Wang, Y.D. (2007). Activation of peroxisome proliferator-activated receptor gamma inhibits cell growth via apoptosis and arrest of the cell cycle in human colorectal cancer. J. Dig. Dis. *8*, 82–88.

Lutz, W., Sanderson, W., and Scherbov, S. (2008). The coming acceleration of global population ageing. Nature *451*, 716–719.

Mikkelsen, T.S., Xu, Z., Zhang, X., Wang, L., Gimble, J.M., Lander, E.S., and Rosen, E.D. (2010). Comparative epigenomic analysis of murine and human adipogenesis. Cell *143*, 156–169.

de la Monte, S.M., and Wands, J.R. (2006). Molecular indices of oxidative stress and mitochondrial dysfunction occur early and often progress with severity of Alzheimer's disease. J. Alzheimers. Dis. *9*, 167–181.

Naj, A.C., Jun, G., Beecham, G.W., Wang, L.-S., Vardarajan, B.N., Buros, J., Gallins, P.J., Buxbaum, J.D., Jarvik, G.P., Crane, P.K., et al. (2011). Common variants in MS4A4/MS4A6E, CD2uAP, CD33, and EPHA1 are associated with late-onset Alzheimer's disease. Nat. Genet. *43*, 436–441.

Oger, F., Dubois-Chevalier, J., Gheeraert, C., Avner, S., Durand, E., Froguel, P., Salbert, G., Staels, B., Lefebvre, P., and Eeckhoute, J. (2014). Peroxisome proliferator-activated receptor γ regulates genes involved in insulin/insulin-like growth factor signaling and lipid metabolism during adipogenesis through functionally distinct enhancer classes. J. Biol. Chem. *289*, 708–722.

Salmon, D.P., and Bondi, M.W. (2009). Neuropsychological assessment of dementia. Annu. Rev. Psychol. *60*, 257–282.

Schintu, N., Frau, L., Ibba, M., Caboni, P., Garau, A., Carboni, E., and Carta, A.R. (2009). PPAR-gamma-mediated neuroprotection in a chronic mouse model of Parkinson's disease. Eur. J. Neurosci. *29*, 954–963.

Schmidt, S.F., Jørgensen, M., Chen, Y., Nielsen, R., Sandelin, A., and Mandrup, S. (2011). Cross species comparison of C/EBPa and PPARg profiles in mouse and human adipocytes reveals interdependent retention of binding sites. BMC Genomics *12*, 152.

Serrano-Pozo, A., Frosch, M.P., Masliah, E., and Hyman, B.T. (2011). Neuropathological alterations in Alzheimer disease. Cold Spring Harb. Perspect. Med. *1*, a006189.

Shappell, S.B., Gupta, R.A., Manning, S., Whitehead, R., Boeglin, W.E., Schneider, C., Case, T., Price, J., Jack, G.S., Wheeler, T.M., et al. (2001). 15S-Hydroxyeicosatetraenoic Acid Activates Peroxisome Proliferator-activated Receptor g and Inhibits Proliferation in PC3 Prostate Carcinoma Cells1. Cancer Res. *61*, 497–503.

Siersbæk, R., Nielsen, R., and Mandrup, S. (2010). PPARγ in adipocyte differentiation and metabolism – Novel insights from genome-wide studies. FEBS Lett. *584*, 3242–3249.

Soccio, R.E., Tuteja, G., Everett, L.J., Li, Z., Lazar, M.A., and Kaestner, K.H. (2011). Species-specific strategies underlying conserved functions of metabolic transcription factors. Mol. Endocrinol. *25*, 694–706.

Soccio, R.E., Chen, E.R., Rajapurkar, S.R., Safabakhsh, P., Marinis, J.M., Dispirito, J.R., Emmett, M.J., Briggs, E.R., Fang, B., Everett, L.J., et al. (2015). Genetic Variation Determines PPARγ Function and Anti-diabetic Drug Response In Vivo. Cell *162*, 33–44.

Teresi, R.E., Shaiu, C.-W., Chen, C.-S., Chatterjee, V.K., Waite, K.A., and Eng, C. (2006). Increased PTEN expression due to transcriptional activation of PPARγ by Lovastatin and Rosiglitazone. Int. J. Cancer *118*, 2390–2398.

Tsai, M.J., and O'Malley, B.W. (1994). Molecular mechanisms of action of steroid/thyroid receptor superfamily members. Annu. Rev. Biochem. *63*, 451–486.

Yu, J.-T., Tan, L., and Hardy, J. (2014). Apolipoprotein E in Alzheimer's disease: an update. Annu. Rev. Neurosci. *37*, 79–100.

Chapter 5

Conclusion

The primary focus of this dissertation is on how genetic variation, single nucleotide polymorphisms (SNPs) and alternative promoter usage, can impact nuclear receptor (NR) function. There is a vast amount of variant-phenotype association data available, but many of these variants are noncoding and we do not know how many of them may influence the regulation of gene expression (Maurano et al., 2012; Ward and Kellis, 2012). Expression quantitative trait loci (eQTL) studies get us one step closer to understanding how these variants associate with disease or phenotypes (Schadt et al., 2008). Gene expression levels between individuals and cell types are regulated by transcription factors (TF) through sequence-specific interactions with genomic DNA. While chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) allows for the identification of genome-wide occupancy of a TF, it can often be challenging to identify the true binding site within a single peak. Attempting to identify genetic variants that can disrupt or alter DNA binding affinity of a TF via ChIP-seq experiments would require a large number of samples and would be a time-consuming process. Protein binding microarrays (PBM) are high-throughput DNA binding assays where custom-designed microarrays are extended to double-stranded oligonucleotides and probed for TF binding directly to the DNA on the slide with a fluorophore-conjugated antibody.

Another level of genetic diversity with impact on tissue-specific gene expression between cell types is the use of alternative promoters. The primary role of alternate promoters is thought to be the control of gene expression under different cellular conditions, including tissue specific gene-expression (Davuluri et al., 2008). However, very little is known about the physiological functions of the different proteins resulting from the alternative promoter usage.

Nuclear receptors (NR) are ligand-sensitive TFs that regulate a wide array of biological processes including development, metabolism, and circadian rhythms. It is not surprising then that they also play a role in many diseases including obesity, diabetes, cancer, atherosclerosis, and inflammation. NRs may interact with many variations of a consensus DNA response element (AGGCTA or AGAACA) throughout the genome as monomers, homodimers, or heterodimers in a wide array of conformations, including direct or inverted repeats with anywhere from 0 to 6 nucleotides (nts) as spacers between each element. At these response elements, NRs recruit transcriptional co-regulators which interact with the general transcription machinery to either repress or activate their target genes.

Since NR activity is modulated by the presence of ligands, naturally ligand synthesis and degradation play a crucial role in tissue-specific hormonal signaling and gene expression. Ligands must be synthesized and delivered throughout the body and the degradation of ligands helps to limit both the duration and the intensity of the NR-ligand response. The enzymes that regulate these processes are the cytochrome P450s (Cyp) enzymes and they play a major role in the oxidative metabolism of cholesterol, steroids, bile acids, and fatty acids (Furge and Guengerich, 2006; Nebert and Russell, 2002).

The A/B domain of NRs is the most variable domain within the superfamily, with no conservation in either length or structure, and is thought to be highly flexible and unstructured in the absence of binding partners (Chandra et al., 2008; Wärnmark et al., 2001). This flexibility has led to difficulties in deciphering the structure of A/B domains, which remain poorly understood.

Hepatocyte nuclear factor 4 alpha (HNF4α) is a liver-enriched transcription factor and a member of the NR superfamily (Sladek et al., 1990). HNF4α is expressed in the liver, kidney, colon, pancreas, stomach, and intestine. It is highly expressed in the liver where it is best known as a master regulator of liver-specific gene expression (Bolotin et al., 2010; Odom et al., 2004) and is essential for adult and fetal liver function. The human *HNF4A* and mouse *Hnf4a* genes are highly conserved and both are regulated by alternative promoters, the proximal P1 and distal P2 promoters. In the adult liver the P1 promoter is the only active promoter, while during fetal development both P1 and P2 promoters are active (Torres-Padilla et al., 2001). P1-HNF4α is expressed in the liver, small intestine, colon, and kidney while P2-HNF4α is expressed in the fetal liver, pancreas, stomach, small intestine, and colon. While it has been shown that P1-HNF4α acts as a tumor suppressor in the liver (Hatziapostolou et al., 2011; Walesky and Apte, 2015), the specific roles of the isoforms remain unclear.

Isoforms from the alternate promoters of HNF4α share >90% homology with each other and have identical DNA-binding and ligand-binding domains. The main difference between P1 and P2 isoforms are alternative first exons that result in an altered A/B domain and the loss of the AF-1 domain for the P2 isoforms (Briançon and Weiss, 2006; Torres-Padilla et al., 2002). P2-HNF4α is not normally expressed in the adult liver, thus, to study the role of P1- and P2-HNF4α in the mouse liver we used genetically engineered (exon swap) mice that express exclusively the P1- or the P2-HNF4α isoforms (Briançon and Weiss, 2006).

In Chapter 2, we investigate the roles of HNF4α isoforms resulting from alternative promoters in a series of genome-scale experiments using exon-swap mice. RNA-seq analysis showed that a 16 amino acid change in the A/B domain of HNF4α can drastically alter the transcriptome of the adult liver, with hundreds of significantly dysregulated. The P2-HNF4α profile showed a loss of expression of CAR (*Nr1i3*) and dysregulated of dozens of phase I, II, and III enzymes. The implications of these differences can be seen in altered lipid and steroid hormone metabolism. In comparing the α7HMZ gene expression profile with HNF4α KO we show that many of the differentially expressed genes found in α7HMZ livers are due to reduced activation potential of P2-HNF4α, most likely due to the loss of AF-1. In comparing with fetal liver expression and liver cancer cell line Hepa1-6 we show that P1- and P2-HNF4α preferentially regulate genes in fetal and liver cancer. ChIP-seq results show that P1- and P2-HNF4α isoforms have nearly identical DNA binding affinities with only roughly 600 (1.5%) differential ChIP peaks. Protein binding microarrays further confirmed these findings highlighting only minor changes specificity of spacer nucleotide DR1 preferred

P2-HNF4α. *De novo* motif calling of uniquely bound peaks revealed a set of TFs that may help orchestrate altered transcriptional activity between P1- and P2-HNF4α isoforms. Protein-protein interaction from rapid immunoprecipitation and mass spectrometry of endogenous proteins (RIME) highlights unique interactions with other NRs and interactions between BMAL (*Arntl*) and CLOCK (*Clock*) suggest unique roles of HNF4α isoforms in circadian rhythms. Additionally, we observe fewer genes dysregulated throughout the day in α7HMZ livers compared to WT suggesting that P1- and P2-HNF4α play different roles in the maintenance of circadian rhythms.

In Chapter 3, we utilized PBMs to investigate the impact of common genetic variation on the DNA binding of nuclear receptors (NR). By probing DNA binding affinity on genomic DNA flanking genetic variants found in the promoters of disease-associated genes, we were able to identify hundreds of affinity altering SNPs (aaSNPs). As expected, NRs that heterodimerize shared the largest number of aaSNPs between them, while NRs that bind completely different response elements (HNF4α, DR1; GR, IR3), displayed very little similarity in PBM binding scores and almost no shared aaSNPs (2). To identify more physiologically relevant aaSNPs a PBM was designed by extracting common variants near the center of ChIP-seq peaks for HNF4α and RXRα. The results of this design show that while data-mining *in vivo* DNA binding assays identifies more aaSNPs, >1,000 for HNF4α2, the usefulness of the slide is limited to the NRs the slide was design for, and any other NRs that may heterodimerize or co-occupy those binding sites. While both designs were successful in identifying many aaSNPs for many NRs, we lacked any physiological relevance for disease or phenotype. The genotype-tissue expression (GTEx) project was data-mined for significant eQTLs in the liver associated with changes in gene expression and we successfully identified thousands of aaSNPs *in vitro*. By cross-referencing these data with NR regulatory networks from the nuclear receptor signaling atlas (NURSA) tool, Transcriptomine, we were able to find hundreds of aaSNPs associated with changes in gene expression of a target gene for the NR they are affecting.

With a similar approach in Chapter 4, we also showed the power of the PBMs to analyze 100,000 genetic variants from PPARγ ChIP peaks, and 1,000 GWAS identified variants associated with Alzheimer's, a neurodegenerative disease commonly treated with PPARγ agonists. The results show that PPARα and PPARδ are more similar in DNA binding affinities than PPARγ is to either, despite having highly conserved DBDs. Analysis of *in vivo* derived murine PPARγ aaSNPs only verified 21.11% at our most strict cut-offs. Relaxing the effect size and p-value threshold can improve this rate to nearly 43.55%. Analysis of the Alzheimer's GWAS SNPs reveals a 100 kb window around *APOE* enriched with aaSNPs for PPARγ, one of which sits in a regulatory region inside the first intron.

The work in this dissertation shows that alternative promoters may play more important roles in genomic diversity than simply providing tissue- or condition-specific expression of TFs. We show that a small (16 amino acid) difference in the A/B domain of HNF4α, can impact the expression of hundreds of genes. The implications of this suggest that the role of the P2 promoter is not simply to control tissue-specific HNF4α expression, but to express a form of HNF4α with an alternate transcriptional profile.

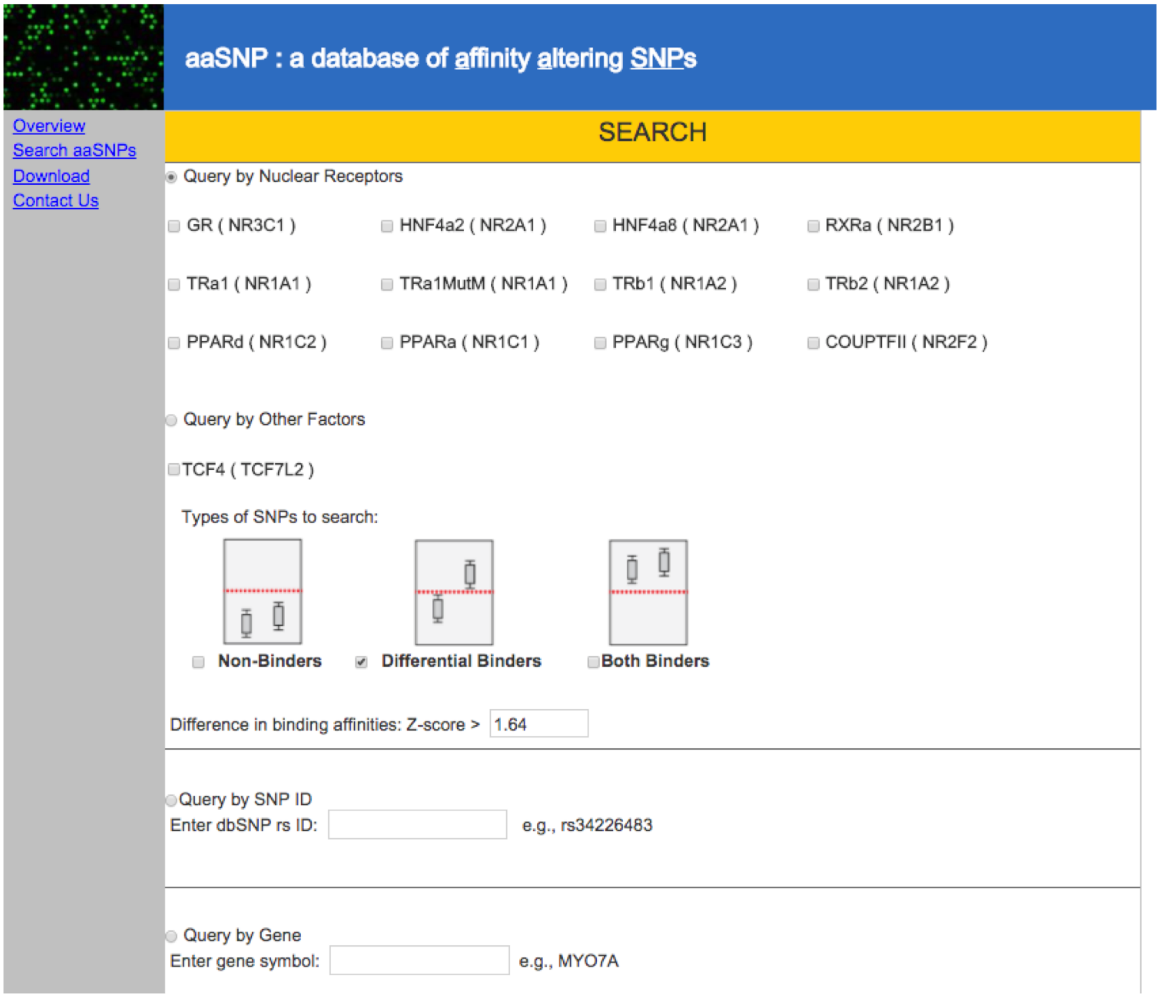
We also show that the high-throughput PBM DNA binding assays are a very powerful tool for measuring DNA binding affinity for a given TF across hundreds of thousands of oligonucleotide probes in a single experiment. Data derived from these experiments can be used to elucidate preferred DNA binding sequences of a TF, as seen in Chapter 2, or to identify SNPs with the capability to disrupt the DNA binding potential of a TF, as seen in Chapters 3 & 4. In total, 14,000 aaSNPs were identified from four 1 million spot PBMs (500,000 total SNPs tested). While GWAS studies have the capabilities to identify thousands of variants associated with disease and phenotype, our PBM analysis shows that a large proportion of these variants may not significantly affect the DNA binding capabilities of many NRs. While we have not extensively probed for DNA binding affinity of all human NRs, we have tested a large number that bind DR1 sequences. Many of these NRs have shown a wide range of DNA binding affinities, even within the same NR subfamily as seen by the diversity between the PPARs in Chapter 4. Testing the rest of the NR family would likely reveal tens of thousands more aaSNPs all while helping to further define the DNA binding specificities of NR subfamilies.

Similarly, this approach can be used for any class of TF with western blot quality antibodies and can be incorporated with many TF regulatory networks.

We propose that aaSNPs could be extremely an important aspect of personalized medicine, and medical genetics, by allowing insight into how common drug targets like NRs (and other TFs) can be differentially recruited to enhancers and promoters in two different individuals and the potential impact on gene expression this may have. In Chapter 3 we have shown that a rare or common genetic variation in an enhancer or promoter that results in a change of DNA binding affinity for a NR may result in a loss of expression in a nearby target gene. As the number of individuals with diseases treated with NR agonist and antagonists increases we greatly increase to population of genetic variants that these medications are being exposed to. Should any of these variants successfully alter the intended NR regulatory networks by altered DNA binding affinity to key response elements, we would expect for there to off-target or unexpected consequences. By incorporating known aaSNPs into NR regulatory networks we start to understand how these altered networks might act upon NR-mediated activation or repression. We hope that this information could be useful in determining the risk of administering drugs targeting NRs before they are prescribed to avoid potentially dangerous side effects.

aaSNP Database

All of the PBM and aaSNP data derived from this dissertation can be accessed online via the aaSNP Database (http://nrmotif.ucr.edu/aaSNP/) hosted by the Sladek Lab. Here we allow researchers to search, interact with, and download aaSNP results. Search parameters can be set to look for aaSNPs by NR, genetic variant, or gene symbol (nearest TSS to SNP, or eGene for GTEx derived data) (Fig 5.1). Search results will display basic information about the SNP such as ID, chromosomal location, and nearest TSS, along with information about the test sequence, PBM binding score, and aaSNP effect size and adjusted p-value (Fig 5.2).



aaSNP : a database of affinity altering SNPs

Overview
Search aaSNPs
Download
Contact Us

SEARCH

● Query by Nuclear Receptors

□ GR (NR3C1)	□ HNF4a2 (NR2A1)	□ HNF4a8 (NR2A1)	□ RXRa (NR2B1)
□ TRa1 (NR1A1)	□ TRa1MutM (NR1A1)	□ TRb1 (NR1A2)	□ TRb2 (NR1A2)
□ PPARd (NR1C2)	□ PPARa (NR1C1)	□ PPARg (NR1C3)	□ COUPTFII (NR2F2)

○ Query by Other Factors

□ TCF4 (TCF7L2)

Types of SNPs to search:

□ **Non-Binders** ☑ **Differential Binders** □ **Both Binders**

Difference in binding affinities: Z-score > 1.64

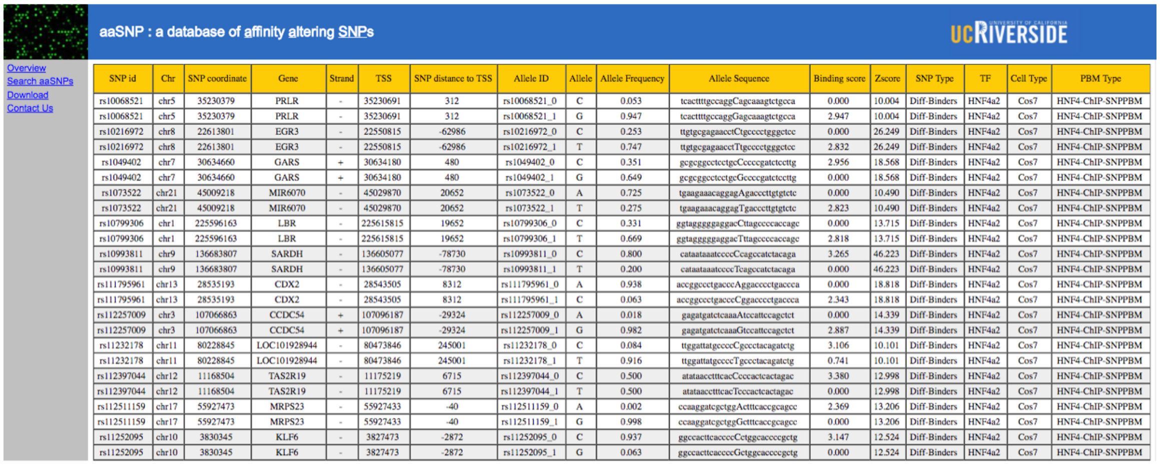
○ Query by SNP ID
Enter dbSNP rs ID: e.g., rs34226483

○ Query by Gene
Enter gene symbol: e.g., MYO7A

Figure 5.1

Figure 5.1. aaSNP Database Search Form

Shown here is a screenshot from December 2017 of the aaSNP Database hosted by the Sladek Lab. Current aaSNP datasets are searchable by multiple NRs (filtering on significance), SNP/GTEx id to identify all NRs disrupted, and by gene symbol to identify aaSNPs nearest a specific transcription start site.



aaSNP : a database of affinity altering SNPs

SNP id	Chr	SNP coordinate	Gene	Strand	TSS	SNP distance to TSS	Allele ID	Allele	Allele Frequency	Allele Sequence	Binding score	Zscore	SNP Type	TF	Cell Type	PBM Type
rs10068521	chr5	35230379	PRLR	-	35230691	312	rs10068521_0	C	0.053	tcacttttgccaggCagcaaagtctgcca	0.000	10.004	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs10068521	chr5	35230379	PRLR	-	35230691	312	rs10068521_1	G	0.947	tcacttttgccaggGagcaaagtctgcca	2.947	10.004	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs10216972	chr8	22613801	EGR3	-	22550815	-62986	rs10216972_0	C	0.253	ttgtgcgagaacctCtgcccctgggctcc	0.000	26.249	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs10216972	chr8	22613801	EGR3	-	22550815	-62986	rs10216972_1	T	0.747	ttgtgcgagaacctTtgcccctgggctcc	2.832	26.249	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs1049402	chr7	30634660	GARS	+	30634180	480	rs1049402_0	C	0.351	gcgcggcctcctgcCccccgatctccttg	2.956	18.568	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs1049402	chr7	30634660	GARS	+	30634180	480	rs1049402_1	G	0.649	gcgcggcctcctgcGccccgatctccttg	0.000	18.568	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs1073522	chr21	45009218	MIR6070	-	45029870	20652	rs1073522_0	A	0.725	tgaagaaacaggagAgacccttgtgtctc	0.000	10.490	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs1073522	chr21	45009218	MIR6070	-	45029870	20652	rs1073522_1	T	0.275	tgaagaaacaggagTgacccttgtgtctc	2.823	10.490	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs10799306	chr1	225596163	LBR	-	225615815	19652	rs10799306_0	C	0.331	ggtagggggaggacCttagccccaccagc	0.000	13.715	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs10799306	chr1	225596163	LBR	-	225615815	19652	rs10799306_1	T	0.669	ggtagggggaggacTttagccccaccagc	2.818	13.715	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs10993811	chr9	136683807	SARDH	-	136605077	-78730	rs10993811_0	C	0.800	cataataaatccccCcagccatctacaga	3.265	46.223	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs10993811	chr9	136683807	SARDH	-	136605077	-78730	rs10993811_1	T	0.200	cataataaatccccTcagccatctacaga	0.000	46.223	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs111795961	chr13	28535193	CDX2	-	28543505	8312	rs111795961_0	A	0.938	accggccctgacccAggacccctgaccca	0.000	18.818	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs111795961	chr13	28535193	CDX2	-	28543505	8312	rs111795961_1	C	0.063	accggccctgacccCggacccctgaccca	2.343	18.818	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs112257009	chr3	107066863	CCDC54	+	107096187	-29324	rs112257009_0	A	0.018	gagatgatctcaaaAtccattccagctct	0.000	14.339	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs112257009	chr3	107066863	CCDC54	+	107096187	-29324	rs112257009_1	G	0.982	gagatgatctcaaaGtccattccagctct	2.887	14.339	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs11232178	chr11	80228845	LOC101928944	-	80473846	245001	rs11232178_0	C	0.084	ttggattatgccccCgccctacagatctg	3.106	10.101	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs11232178	chr11	80228845	LOC101928944	-	80473846	245001	rs11232178_1	T	0.916	ttggattatgccccTgccctacagatctg	0.741	10.101	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs112397044	chr12	11168504	TAS2R19	-	11175219	6715	rs112397044_0	C	0.500	atataacctttcacCcccactcactagac	3.380	12.998	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs112397044	chr12	11168504	TAS2R19	-	11175219	6715	rs112397044_1	T	0.500	atataacctttcacTcccactcactagac	0.000	12.998	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs112511159	chr17	55927473	MRPS23	-	55927433	-40	rs112511159_0	A	0.002	ccaaggatcgctggActttcaccgcagcc	2.369	13.206	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs112511159	chr17	55927473	MRPS23	-	55927433	-40	rs112511159_1	G	0.998	ccaaggatcgctggGctttcaccgcagcc	0.000	13.206	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs11252095	chr10	3830345	KLF6	-	3827473	-2872	rs11252095_0	C	0.937	ggccacttcaccccCctggcaccccgctg	3.147	12.524	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM
rs11252095	chr10	3830345	KLF6	-	3827473	-2872	rs11252095_1	G	0.063	ggccacttcaccccGctggcaccccgctg	0.000	12.524	Diff-Binders	HNF4a2	Cos7	HNF4-ChIP-SNPPBM

Figure 5.2

Figure 5.2. aaSNP Database Results Page

Shown here is the results page from December 2017 of an aaSNP search for significant variants affecting HNF4α2. Shown are SNP id, chromosome, SNP location, and nearest gene information including symbol, strand, and distance to TSS. Information related to the aaSNP show an allele specific id used on our PBMs, the allele that was tested along with allelic frequency, and the test sequence, PBM binding score and Z-score of the aaSNP.

References

Bolotin, E., Liao, H., Ta, T.C., Yang, C., Hwang-Verslues, W., Evans, J.R., Jiang, T., and Sladek, F.M. (2010). Integrated approach for the identification of human hepatocyte nuclear factor 4alpha target genes using protein binding microarrays. Hepatology *51*, 642–653.

Briançon, N., and Weiss, M.C. (2006). In vivo role of the HNF4α AF-1 activation domain revealed by exon swapping. EMBO J. *25*, 1253–1262.

Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T.P., and Rastinejad, F. (2008). Structure of the intact PPAR-gamma-RXR- nuclear receptor complex on DNA. Nature *456*, 350–356.

Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T.H.-M. (2008). The functional consequences of alternative promoter use in mammalian genomes. Trends Genet. *24*, 167–177.

Furge, L.L., and Guengerich, F.P. (2006). Cytochrome P450 enzymes in drug metabolism and chemical toxicology: An introduction. Biochem. Mol. Biol. Educ. *34*, 66–74.

Hatziapostolou, M., Polytarchou, C., Aggelidou, E., Drakaki, A., Poultsides, G.A., Jaeger, S.A., Ogata, H., Karin, M., Struhl, K., Hadzopoulou-Cladaras, M., et al. (2011). An HNF4α-miRNA inflammatory feedback circuit regulates hepatocellular oncogenesis. Cell *147*, 1233–1247.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Nebert, D.W., and Russell, D.W. (2002). Clinical importance of the cytochromes P450. Lancet *360*, 1155–1162.

Odom, D.T., Zizlsperger, N., Benjamin Gordon, D., Bell, 1. George W., Rinaldi, N.J., Murray, H.L., Volkert, 1. Tom L., Schreiber, J., Alexander Rolfe, P., Gifford, D.K., et al. (2004). Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. Science *303*.

Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biol. *6*, e107.

Sladek, F.M., Zhong, W.M., Lai, E., and Darnell, J.E., Jr (1990). Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. Genes Dev. *4*, 2353–2365.

Torres-Padilla, M.E., Fougere-Deschatrette, C., and Weiss, M.C. (2001). Expression of HNF4a isoforms in mouse liver development is regulated by sequential promoter usage and constitutive 3 end splicing. Mech. Dev. *109*, 183–193.

Torres-Padilla, M.E., Sladek, F.M., and Weiss, M.C. (2002). Developmentally regulated N-terminal variants of the nuclear receptor hepatocyte nuclear factor 4alpha mediate multiple interactions through coactivator and corepressor-histone deacetylase complexes. J. Biol. Chem. *277*, 44677–44687.

Walesky, C., and Apte, U. (2015). Role of hepatocyte nuclear factor 4α (HNF4α) in cell proliferation and cancer. Gene Expr. *16*, 101–108.

Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. *30*, 1095–1106.

Wärnmark, A., Wikström, A., Wright, A.P.H., Gustafsson, J.-Å., and Härd, T. (2001). The N-terminal Regions of Estrogen Receptor α and β Are Unstructured in Vitro and Show Different TBP Binding Properties. J. Biol. Chem. *276*, 45939–45944.