Title

Investigation of Transcription Factor Binding Sequences and Target Genes using Protein Binding Microarrays

Author
Bolotin, Eugene Leonidovich

Publication Date
2010

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Investigation of Transcription Factor Binding Sequences and Target Genes Using Protein Binding Microarrays

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

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Eugene Leonidovich Bolotin

March 2010

Dissertation Committee:
Dr. Frances M. Sladek, Chairperson
Dr. Frank Sauer
Dr. Tao Jiang

The Dissertation of Eugene Leonidovich Bolotin is approved

Frank Sauer

Tao Jiang

Frances M. Sladek, Chairperson

University of California, Riverside

Acknowledgments

I am grateful to all the people who have supported me through this endeavor. I would like to especially thank my advisor Dr. Frances M. Sladek, without whom this dissertation would not have been possible. I thank my co-advisors Dr. Frank Sauer and Dr. Tao Jiang for insightful advice and continuous support. Additionally, I would like to thank my friend Dr. John Ta for providing many fun times and invaluable reagents. I would also like to thank my family and friends who got me through many tough times.

ABSTRACT OF THE DISSERTATION

Investigation of Transcription Factor Binding Sequences and Target Genes Using Protein Binding Microarrays

by

Eugene Leonidovich Bolotin

Doctor of Philosophy, Graduate Program in Genetics, Genomics and Bioinformatics
University of California, Riverside, March 2010
Frances M. Sladek, Chairperson

This dissertation describes the investigation of binding rules and DNA binding sequences for several transcription factors (TFs). We develop Protein Binding Microarrays (PBMs) to study the interactions between TFs and DNA *in vitro* and we use a support vector machine (SVM) algorithm to capture these interactions *in silico*. We then apply this methodology to study the binding of TFs to promoters and repetitive sequences in a genomewide fashion.

In Chapter 2, we thoroughly investigate HNF4α/DNA binding interactions using PBMs. We investigate binding specificities for various isoforms and species of HNF4α. We then use PBMs to rank ~ 4,000 HNF4α binding sequences in order of binding affinity. Using this training set we identify/predict novel HNF4α binding sequences and rules, and from these rules we generate a model for HNF4α binding. We then use this large dataset, in combination with ChIP-on-chip and RNAi followed by an expression profiling to identify hundreds of novel HNF4α direct target genes.

In Chapter 3, we identify HNF4α association with Alu repeats, a novel finding. We investigate HNF4α binding to Alu sequences in *in vitro* and *in vivo* in the promoters of HNF4α-regulated genes, and thus reveal a novel association between HNF4α and Alu repeats.

Finally in Chapter 4, we leverage the PBM technology to investigate the binding properties of transcription factors COUP-TF2 and TCF-1. We identify many sequences that bind both HNF4α and TCF-1 and those bind both HNF4α and COUP-TF2. This finding suggests competition between these TFs on the promoters of their target genes. Additionally, we investigate the effect of coregulator PGCα and the effect of the endogenous ligand, linoleic acid,on HNF4α DNA binding.

This study significantly advances our knowledge of binding sequences, binding motifs, target genes, and transcriptional regulation for several transcription factors, HNF4α, COUP-TF2 and TCF-1. It also sheds light on evolution of HNF4α binding sequences through Alu repetitive elements. Finally, it provides a powerful framework for the comprehensive investigation of transcriptional regulation in mammalian systems for other transcription factors.

Contents

List of Figures

List of Tables

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.

Philosophy of Science §1
ALBERT EINSTEIN

Chapter 1

Introduction

1.1 Eukaryotic Transcriptional Regulation

Transcription is essential eukaryotic for cell function. Transcription is quite simply the generation of an RNA transcript copy from a DNA template. All transcription in a cell is accomplished by RNA polymerases. There are three distinct RNA polymerases in eukaryotic genome: RNA polymerase I transcribes ribosomal RNA used in translation; RNA polymerase II transcribes mainly protein coding messenger RNA, regulatory micro RNA, and some small RNAs; RNA polymerase III transcribes mainly transfer RNA and 5s subunit of ribosomal RNA both essential in translation. In this dissertation, we are interested in the regulation of protein coding genes, and thus are studying the regulation of RNA polymerase II [20].

As mentioned, there is a great diversity of RNA types, such as messenger RNA (mRNA), micro RNA (miRNA), small nuclear RNA (snRNA) and many others [42, 43]. These RNAs could very different functions. For example; mRNA is translated into proteins; miRNA is used directly in gene regulation to repress transcription; and tRNA is used as an adapter between RNA and protein in translation. For a long time it was thought that only small regions of the genome were transcribed. However, new findings show that most of the human genome is transcribed emphasizing the ubiquity of RNA [67]. Not surprisingly, transcription of critically important genes is under tight regulatory control in a cell. The location of a transcript in the cell, time of transcription, and the amount of transcript produced is tightly regulated, and critical for cell function.

Regulation can happen at every step of transcription; initiation, elongation, and termination. During initiation, RNA polymerase II and general transcription factors (GTFs) assemble at the transcription start site (TSS +1). During elongation, the RNA transcription machinery synthesizes the RNA transcript. During termination, the RNA is freed from the RNA polymerase and the transcription machinery dissociates from the DNA template. After the transcript is made other mechanisms regulate the stability, splicing, localization and other properties of the transcript [39].

Transcriptional initiation takes place at +1, in the promoter region of the gene. The promoter consists of two regions; the core promoter, typically ~100 bp upstream and downstream of TSS, and a proximal promoter that is ~1-2 kb upstream of TSS. The core promoter is bound by GTFs while proximal promoter is bound by specific transcription factors (TFs). Additionally, another regulatory region called the enhancer can be located hundreds or thousands of kb away from the TSS and can influence transcription. Enhancers, like proximal promoters, bind TFs and can activate or repress transcription. TFs differ from other DNA binding proteins in that they can influence transcription, by having either activating or repressive effect on RNA polymerase II complex. They often act in a combinatoral fashion on the promoter by recruiting various proteins and enzymes that have many types of activities such as histone acetyltransferase (HAT) and histone deacetylase (HDAC) activity. The grand total of the transcriptional machinery on the promoter is what regulates transcription initiation [32].

When the mechanisms of transcriptional control break down, the cell can experience many effects such as: inefficiency in function, production of harmful proteins, apoptosis, out of control proliferation, and even cell death. Often these breakdowns can affect the whole organism leading to disease and even death. In this dissertation we are interested in the regulation of transcription initiation. Many mechanisms for the regulation of transcription initiation have been identified, but undeniably a major factor, from bacteria to vertebrates, are the TFs.

1.1.1 Overview of Transcription Initiation

In eukaryotic cells, unlike prokaryotic cells, RNA polymerase II cannot transcribe genes without *cis* acting regulatory elements or TFs [18]. The transcription machinery is complex and involves coordination of RNA polymerase and GTFs, and specific TFs. GTFs are single protein factors such as TFIIA and TFIIB as well as multi protein complexes such as TFIID, TFIIF and TFIIH, that are required transcription [33, 58]. GTFs are found at every transcriptional event, while different TFs are recruited depending on promoter type, the cell type, and the state of the cell. TFs bind to the *cis* region of the transcribed gene and initiate or repress transcription by recruiting the basal transcription machinery (i.e. GTFs and the RNA polymerase II complex), or preventing it from binding to DNA (Fig. 1.1).

TFs do not initiate transcription on their own. They function by recruiting various coactivators and corepressors of transcription. Co-activators and corepressors affect the recruitment of RNA polymerase indirectly, by chemically modifying histones in the promoter regions. Depending on the type and amount of modification, the chromatin condenses, repressing transcription, or decondenses allowing transcription to proceed. The Mediator complex is another important player in the transcriptional regulation and it serves as a link between RNA polymerase II complex to the TFs, however the exact mechanism of how it influences initiation of eukaryotic transcription is unclear [10, 34, 35].

Regulation of transcriptional initiation is one of the ways the cell regulates the amount of the appropriate transcript at the appropriate time, and it is the way organisms regulate tissue specific, temporal, and spacial transcription. Since TFs take such a central stage in gene regulation and protein composition of the cell, their importance cannot be underestimated [20, 32, 39, 49, 52].

1.2 Chromatin Organization and Transcriptional Regulation

Chromatin organization has a tremendous impact on the regulation of transcription initiation. Chromatin is the highly regulated nucleoprotein complex, composed primarily DNA wound around a histone octamer and any other protein that binds DNA. Chromatin is typically divided into heterochromatin or euchromatin. Heterochromatin is considered "silenced" for transcription while euchromatin is considered actively transcribed. However, recent research has shown that the regulation of chromatin is much more complex

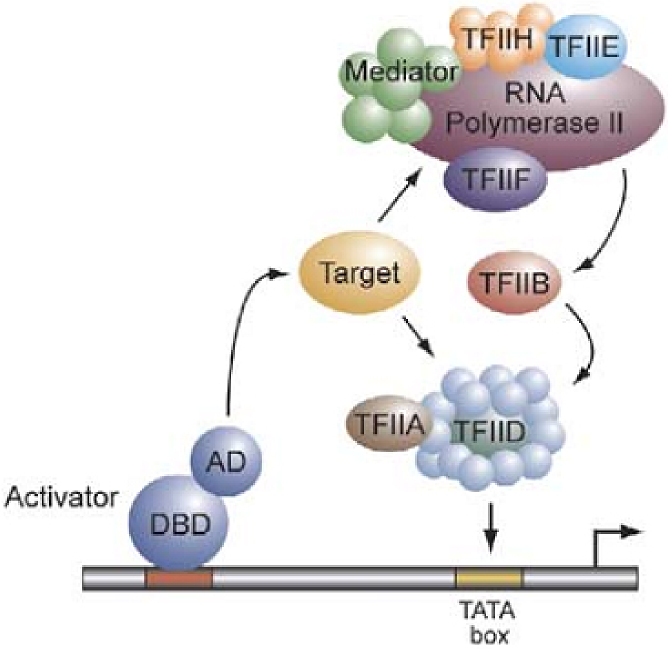


Figure 1.1. Schematic diagram of the complicated interplay of transcription machinery of gene regulation adapted from [18]. DBD is the DNA binding domain and AD is the activation domain of a given TF.

and is dependent on a “histone code.” The histone code is a complex combinatorial post-translational modification of histone tails, which can either tighten or relax the DNA wrapped around the histones. While the relaxed state facilitates RNA polymerase binding and activation of transcription, the tightened state prevents the polymerase complex from binding and represses transcription. Modifications that have been shown to be involved in transcriptional activation are acetylation of histones H3 and H4 as well as trimethylation of histone H3 on the lysine residue [52]. Transcriptional repression has been correlated with a loss of acetylation of histones of H3 and H4 as well as methylation of histone H3 at lysine 9 residue. Recent studies suggest that regulation of transcription by chromatin is much more complex than some of the examples mentioned in this paragraph and many other histone post-translational modifications are involved in chromatin organization [8].

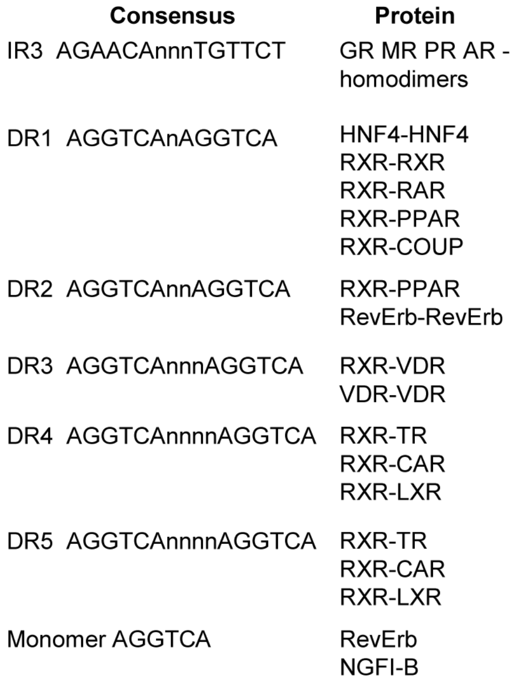
1.2.1 Variables Affecting Initiation of Transcription by TFs

Many variables affect the initiation of transcription: the type of TF bound, the number of such TFs, their distance from the TSS, their direction relative to the promoter, their composition (single protein, homodimer, heterodimer), the structural conformation of the TF, any ligands that bind the TF, and possibly other undiscovered mechanisms [1]. TFs binds a specific DNA sequence (i.e.) transcription factor binding site (TFBS), frequently in the promoter region of a gene or in a nearby enhancer element, although functional *cis* acting TF binding site as far as 100kb away from the TSS have been reported [56]. The binding sequence is typically a sequence of DNA nucleotides usually between 4 to 25 bases long [66]. A binding motif is a simple model or a general representation of the sequence to which a TF binds, typically displayed as a consensus sequence or a position weight matrix (PWM), as identified through comparative analysis of known TFBS (Figs 1.2, 1.5). In short it represents the similarities between the types of DNA sequences that a TF binds. Often when represented as motifs, similarities between binding sequences become apparent, such as “core positions” that are often considered more important in TF/DNA interactions, or highly variable positions that are considered less important, but not necessarily so. Because a given TF may bind thousands of different binding sequences with differential affinity, an *in silico* prediction analysis of a TFBS is challenging. In addition, the relatively small (~4-25 nucleotide) size of these motifs and their low information content creates a high probability of occurrence of a particular motif in the genome purely by chance and therefore without biological significance [57].

1.3 Nuclear Receptors

Nuclear receptors (NRs) are a superfamily of conserved TFs that share the architecture of a very well conserved DNA binding domain (DBD) and a ligand binding domain (LBD) as well as presence of a less conserved activation function domain (AF-1). They differ from most TFs in that the majority are able to bind small lipophilic molecular ligands. Upon ligand binding most NRs recruit coactivators and activate transcription. However, there are some exceptions. Endogenous ligands have been identified for several "orphan" NRs (such as COUP-TFs) and some have no DNA binding domain (SHP, DAX). Additionally some NRs are known to repress transcription instead of acting as activator (such as Rev-erb). Since NR ligands are used for communication between organs, regulation of inflammation and many other functions in the organism, the NRs provide an important link between transcription and physiology. Since together the NR family regulates thousands of genes, it is not surprising that mutations in the coding regions of NRs have been linked to variety of diseases. Because ability of the NRs to bind small molecules and change the transcriptional state of the cell, many successful drugs have been developed to modulate their activity. It is estimated that more than 13% of total drugs on the market today target NRs, constituting a multibillion dollar industry [48]. Not surprisingly, considerable efforts have been extended to elucidate mechanisms of action of NRs and to identify their target genes. NR family members are phylogenetically related and their DBDs are conserved.

Because of that conservation their DNA binding motifs appear to be very similar [41]. The first binding sequence for NR was identified using a DNAse protection assay in 1984 for glucocorticoid receptor (GR) NR3C1 [30]. The sequence was subsequently characterized as a palindrome consisting of an inverse repeat with a spacing of 3 nucleotide referred to as an IR3 [62]. Further studies by Ron Evans' group [61] and Glass et al. [17] determined that another class of NRs binds to Direct Repeats (DRs) with various spacings (0 to 5) described as "DR" rules. Hence, the receptors have been divided into two conserved subfamilies for which the ligands are also similar, and based on the structure of their ligands are categorized into "steroid" and "non-steroid" groups. The steroid group includes, but is not limited to glucocorticoid receptor (GR) NR3C1, estrogen receptors ERα NR3A1 and ERβ NR3A1, progesterone receptor (PR) NR3C3 , mineralocorticoid receptor (MR) NR3C3, androgen receptor (AR) NR3C4. The steroid receptors bind IR repeats with various spacers. The non-steroid group binds DR repeats with AGGTCA half site and includes, but is not limited to retinoid acid receptors (RXRs) NR2B1 NR2B2 NR2B3, peroxisome proliferator-activated receptors (PPARs) NR1C1 NR1C2 NR1C33, liver X receptors (LXRs) NR1H3 NR1H3, farnesoid X receptor (FXR) NR1H4, pregnane X receptor (PXR) NR1L2 and HNF4s [31]. Ability of RXR commonly forms a heterodimer with other receptors and depending on the pairing prefers a variety of spacer lengths (Fig. 1.2).



Consensus	Protein
IR3 AGAACAnnnTGTTCT	GR MR PR AR - homodimers
DR1 AGGTCAnAGGTCA	HNF4-HNF4 RXR-RXR RXR-RAR RXR-PPAR RXR-COUP
DR2 AGGTCAnnAGGTCA	RXR-PPAR RevErb-RevErb
DR3 AGGTCAnnnAGGTCA	RXR-VDR VDR-VDR
DR4 AGGTCAnnnnAGGTCA	RXR-TR RXR-CAR RXR-LXR
DR5 AGGTCAnnnnAGGTCA	RXR-TR RXR-CAR RXR-LXR
Monomer AGGTCA	RevErb NGFI-B

Figure 1.2. Summary of the binding "rules" for a subset of nuclear receptors. IR3 is shown for steroid nuclear receptors, ER is a special case of a non steroid receptor binding to IR3 with AGGTCA half site. DRx for the non steroid nuclear receptors are shown with varying spacing. Some of the NRs can bind as monomers. Figure adapted from [31].

1.4 Background for Hepatocyte Nuclear Factor 4α

1.4.1 General Importance

HNF4α was chosen as the primary focus for this dissertation because it possesses several important properties including: high conservation, physiological and disease relevance, and large database of known target genes and binding sites. These properties make HNF4α a perfect candidate for investigation of its binding sites, and target genes. Not only there was a large number of known TFBS for HNF4α at the start of this project, which facilitated the testing of a novel technology, but there was an even larger group of potential binding sequences and regulated genes that remained to be discovered. Furthermore, the high degree of conservation of HNF4α allowed us to make hypothesis across multiple species, and its high physiological relevance would allow us and others to use the TFBS investigation to potentially contribute to our understanding of disease.

1.4.2 Structure and Function

Hepatocyte nuclear factor 4 alpha, *HNF4A*, (HNF4α)(NR2A1) is a member of the superfamily of NRs and is a liver-enriched TF that is also expressed in the kidney, pancreas, intestine, colon and stomach [5]. Originally identified based on its ability to bind DNA response elements in the human apolipoprotein C3 (*APOC3*) and mouse transthyretin (*TTR*) promoters [54], HNF4α has since been shown to play a critical role in both the development of the embryo and the adult liver [22, 64]. HNF4α is highly conserved across species and found in wide variety of organisms from mammals to insects, with 100% amino acid conservation of DNA binding domain across mammalian species. It is thus far has been found in every multicellular animal examined. [53]. It regulates a wide range of target genes and is involved in a variety of biological processes such as: transport, metabolism, and development. It is especially important in hepatocyte differentiation and normal adult liver function [22, 36].

Since ligands play a critical role in the function of NRs, a concentrated effort has been applied to identifying an endogenous ligand for HNF4α. These efforts have been met with success and HNF4α has been found to be reversibly bound to linoleic acid (LA). Unfortunately, the ligand's functional significance is still not very well understood [68].

Thus, HNF4α has been categorized into the "enigmatic adopted" orphans category, a group of NRs for which the endogenous ligand has been identified, but the function of the ligand remains unclear [55].

1.4.3 Mutations in *HNF4A*

HNF4α is a master regulator of liver function [5, 29, 47]. Not surprisingly, many mutations in its coding region have been linked to variety of liver and pancreas-related diseases. For instance, mutations in the HNF4α DBD and LBD have been linked to early-onset type 2 diabetes of the young (MODY1) [45, 53]. Multiple mutations (and SNPs) associated with diabetes have been mapped to the coding region of *HNF4A* and in the P2 but not the P1 promoter, consistent with the P2 promoter driving expression of HNF4α in the beta cells of the pancreas [5]. Additionally, mutations in the TFBS of HNF4α in the promoter of Factor VII and Factor IX have been found to be involved in involved in hemophilia, which emphasizes underlying the importance of cataloging HNF4α regulatory elements [9, 51, 53].

1.4.4 HNF4α Structure and Isoforms

HNF4α is similar to other NRs in having six modular domains. The domains are named A-F, and are broadly characterized by function [19]. The A/B domain is important for recruiting co-activators of transcription; C domain, is a zinc finger DBD; D is a hinge region that plays a role in DNA binding; E domain is involved in ligand binding and protein dimerization; and F domain is important in repression of transcription (Fig. 1.3).

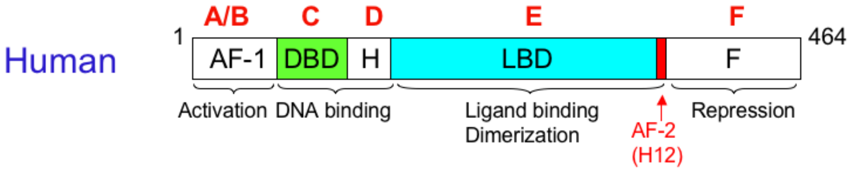


Figure 1.3. Shown is HNF4α2 isoform. At the top are the classical domains; at the bottom are the functions. Many co-activators (p300, CBP. SRC1. GRIP1), mediator components, general transcription factors (TFIIB, TBP, TAFs, PC4, ADA2) and transcriptional activators (Smad3/4) have been found to interact with the AF-1 region. The zinc finger (Zn++) plus hinge (H) region is sufficient for DNA binding although the LBD provides the major dimerization motifs in helices 9 and 10. The AF-2 (helix 12) is absolutely required for transactivation and interaction with various co-activators. Other transcriptional activators (HNF1, p53, SHP, SREBPs, COUP-TFs, Sp1) and co-regulators (PGC1, p300, CBP, GRIP1, Src1, ACTR), as well as the co-repressor SMRT, also interact with the LBD. The F domain represses transcription, but on its own does not interact with the co-repressor SMRT. Adapted from [5].

The *HNF4A* gene exhibits tissue-dependent alternate splicing and has nine proposed isoforms which vary in both the C-and N-terminal regions. The isoforms show physiologically important phenotypic differences in transgenic isoform-specific mice [7]. Alternate splicing in HNF4α is developmentally regulated by two promoters, P1 and P2. The primary isoforms dependent on the P1 promoter are HNF4α1 and HNF4α2, while the primary isoforms regulated by the P2 promoter are HNF4α7 and HNF4α8 (Fig. 1.4). Recent findings show that there are some subtle yet important differences in effects of HNF4α1 and HNF4α7 mainly on lipid metabolism possibly through the apolipoprotein gene family [7].

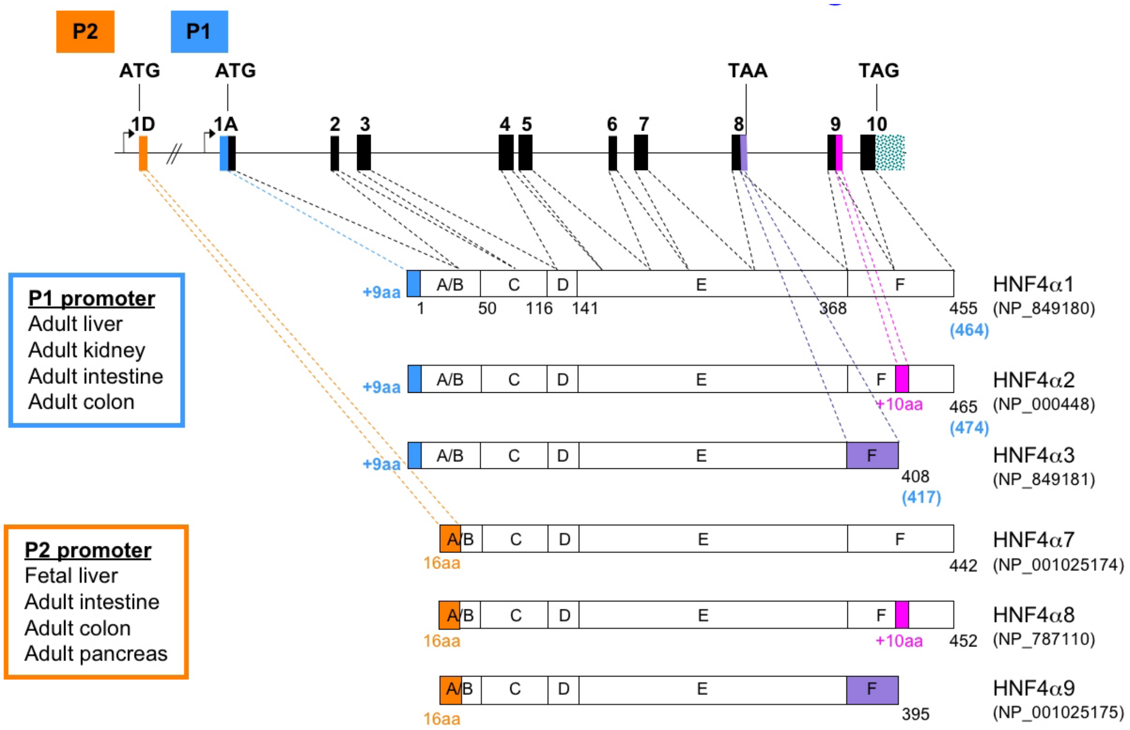


Figure 1.4. Domains of HNF4α. The human *HNF4A* gene spans ~74 kb and contains two promoters that drive the expression of at least 6 splice variants. Numbering is based on the original amino acid sequence. A conserved alternate translation start site in the P1 promoter is shown in blue. Letters refer to the functional domains of nuclear receptors. Indicated are the major tissues in which the P1 and P2 promoters are expressed [21]; fetal kidney, gut and stomach, as well as the visceral endoderm also express HNF4α although promoter usage has not been established [13]. Stippled bars refer to untranslated regions. Exons 1C and 1B were originally proposed to create an insertion in the A/B domain giving rise to isoforms HNF4α4/5/6 [12, 15], although use of that exon in the full length protein is now in question [21, 24](G. Ryffel, personal communication). The mouse *Hnf4a* gene has a similar structure and expression pattern, with minor variations; the mouse HNF4α2 protein is ~96% percent identical to human. Additional P2-driven isoforms have been recently reported [25]. Adapted from [5].

1.4.5 HNF4α DNA Binding

HNF4α binds DNA exclusively as a homodimer [4, 27]. The canonical HNF4α consensus sequence consists of the half site AGGTCA with one nucleotide spacer (referred to as a DR1, AGGTCAxAGGTCA) [28]. The binding can be described as a position weight matrix (PWM), a logo where the relative size of a nucleotide letter corresponds to its relative frequency at a fixed position with respect to other nucleotides at that position (Fig. 1.5). HNF4α shares its consensus binding sequence with phylogenetically related NR family members such as COUP-TF, HNF4γ, RXR and others. Recently the human HNF4α DBD bound to DNA was crystallized [38]. The crystal structure shows that HNF4α prefers the right half-site of the DR1 to the left half-site. However, the crystal structure alone is not enough to predict HNF4α binding. In order to predict binding for HNF4α, an extensive dataset of HNF4α binding sequences and their affinities is needed. We are fortunate to have a large amount of preliminary binding data for HNF4α (>217), from the literature and gel shift assays conducted in our lab (see Table 6.1). However, even this amount is not enough to accurately identify all binding sequences for HNF4α. Additionally, the ≈ 217 identified sequences were derived in a biased fashion on the first identified HNF4α binding sites, and following the discovery of the direct repeat rules for NR DNA binding on the DR1 consensus [61]. Therefore, it is possible that if surveyed in an unbiased fashion another distinct binding consensus might emerge. Furthermore, the total number of possible 13-mer sequences that HNF4α can potentially bind is much greater than 217 (4^{13} ~67 million), and whereas HNF4α will certainly not bind all potential 13-mers, the total number of DNA sequences that will bind HNF4α is anticipated to be in the 10,000's. Additionally, even though HNF4α shares its consensus sequence with other NRs; there must be critical differences in binding between them, otherwise the cell would not be able to differentiate the binding of these TFs.

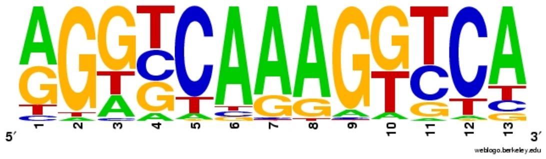


Figure 1.5. PWM of HNF4α binding from derived from 217 binding sequences from the literature and EMSA done by Chuhu Yang (Table 6.1)

1.5 Methods of TFBS Prediction

1.5.1 *In silico* methods

Because of the importance of transcriptional regulation many *in silico* methods have been devised to predict TFBS throughout the genome. Despite their differences, they can be divided into two major categories. The first group of methods are algorithms that predict binding sites *ab initio*, or without prior knowledge of what the sequences are or what kind of TF binds to them. These methods are referred to in the machine learning field as "unsupervised learning methods." Usually such methods look for short, over represented sequences in the genome or across several genomes. These TFBS are often "weighted" by their occurrences in promoters of genes with similar functions, conserved regions, enriched in ChIP assays and others. Such methods are Gibbs sampling, word matching, sequence conservation or a combination of any of the above [11]. The second type of methods utilize an initial experimentally derived training set of motifs for a given TF. These methods are referred to as "supervised learning methods." Given a training set, various models can be built to search for more potential TFBS. Examples of these methods are position weight matrices (PWM), Markov Chains (MC), and information content based methods [57].

Both supervised and unsupervised approaches are excellent for generating potential binding sites, but each has unique advantages and major disadvantages. The advantage of *ab initio* unsupervised learning methods methods is their ability to discover completely novel binding sequences and motifs, while their disadvantage is their inability to distinguish functional from nonfunctional TFBS or to identify which TF, if any, binds a specific TFBS. The advantage of supervised methods is their ability to predict potential TFBS for a given TF. The disadvantage is their large >30% false positive rate for simple methods like PWM [11, 70] and, if used on a genomic scale an overwhelming amount of false positive matches are that produced. This disadvantege can be overcome by increasing the size of the training set to create a more complex model. However, most TFs have less than 20 literature derived TFBS, and even the most well studied TF, SP1, has ~200, limiting the complexity of the model that could be fit to the datasets as evidenced by Transfac database [65]. This forces a smaller model that often does not take into account interdependences between positions of a sequence and cannot accurately model binding [59].

1.5.2 Support Vector Machine

One method that deserves more attention is Support Vector Machine (SVM), developed by Vladimir Vapnic at AT&T for optical character recognition [6]. It belongs to the supervised learning group of prediction/machine learning algorithms because it needs a rather extensive training set [46]. It is currently underused for TFBS prediction because of its large training set requirement. However, it is one of the few methods that takes into account interdependences between binding positions. SVM is a group of methods that has been developed over the last several decades and is currently an extremely robust and reliable set of methods for classifying complex datasets. Thus, the SVM approach is perfectly suited for TFBS prediction, as will be demonstrated later. Briefly, SVM is a form of multidimensional regression classifier that uses a hyperplane to separate the two data sets; in sequence recognition context, binding from non binding sequences. The SVM attempts to optimally separate binding from non binding sequences by maximizing the distance between them and the hyperplane. It has been further extended to multiple classification and to continuous regression which allows for prediction of continuous values as opposed to classification. SVM have been successfully used in biological sequence analysis when classification of large datasets is required and a large training set is available, and has been found to have <10% error rate, which makes them three to five times more effective than PWM which typically has >30% error rate, for sequence prediction [23, 60, 71].

1.6 Methods of Binding Determination

1.6.1 Overview of Technologies

Over the years many experimental methods were proposed for calculating affinities of DNA/TF interactions or ranking sequences by their binding affinity. The classical method is the electrophoretic mobility shift assay (EMSA) [16]. EMSA is an assay that visualizes the migration of a DNA/protein complex through the native polyacrylamide gel. Additional confirmation is usually conducted with an antibody to the protein of interest to control for non specific binding; if binding is specific, an additional band is then observed corresponding to protein/antibody/DNA complex that migrates slower due to the larger size. While EMSA is the current "gold standard" for determination of TF/DNA interactions, it is cumbersome, and not amendable to high throughput screening. The maximum number of sequences compared in a single experiment is dependent on the number of wells in a gel and seldom exceeds 20. Since TF can bind hundreds if not thousands of different sequences additional methods for determining these interactions have been devised. These methods include SELEX, surface plasmon resonance, Protein Binding Microarrays (PBMs), micro-cantilever technologies and others [26, 40, 69]. Out of all of these methods, Protein Binding Microarrays are the most mature platform in being being high throughput, relatively inexpensive, easy to perform, highly reproducible, and highly correlated to binding affinities [2, 3].

1.6.2 Protein Binding Microarray (PBMs)

PBM technology was first pioneered by the Udalova group at Oxford in 2004 [37] and later developed further by the Bulyk group at Harvard [2, 3, 44]. In PBMs, double stranded DNA oligonucleotides $\sim$ 50-60 base pairs are immobilized on a glass slide. The sequence of the 10-20 base pairs closest to the glass slide are identical between all the oligonucleotides. The last 20-30 bases are varied in various ways in an attempt to capture the full range of possible binding sequences. There are many ways to manufacture PBMs, but they are typically ordered from manufacturers of conventional single-stranded arrays and are made double-stranded by polymerase extension reaction [2], although it is also possible to generate them using oligonucleotide self complementarity [63]. After the arrays have been manufactured, the proteins of interest are over expressed/purified and hybridized to the array. The TF is allowed to hybridize to the DNA and subsequently the non bound protein is washed off. The concentration of the TF bound to each spot is measured using immunofluorescence to the immuno-specific tag on the TF or directly by anti-TF antibody. The power of PBMs is that they are able to rank the potential TFBS in the order of the approximate relative binding affinities, and are well correlated to EMSA results [2, 14]. Typically 15,000 to 50,000 oligonucleotides could be measured at the same time, although the next generation Agilent arrays can accommodate up to 1,000,000 oligonucleotide spots.

1.6.3 Bulyk et al. PBM vs Bolotin et al. PBM

In this dissertation we will extensively use PBM in order to identify binding sequences for the TFs HNF4α, COUP-TFII, LEF/TCF, and to investigate interactions between HNF4α and the coactivator PGC1α. There are certain major differences between PBMs developed by Bulyk and others and the ones developed in this work. Because the aims of other groups was to make "universal" PBM to identify binding sequences of as many TFs as budget, and manpower allows, their strategy is to make arrays with every possible of 8nt ($4^8 = 65,536$) sequence represented. In order to do that they use a combinatorial method called "de Brujin" overlapping sequences [50]. They then use a form of clustering to generate a "Z-score" based on the average binding intensity of all the sequences containing the sequence of interest [2]. The disadvantage of this approach is the inability to resolve motifs longer than 8nt in length. Because distance to the glass slide greatly affects sequence binding and multiple proteins could bind to a single sequence, their arrays have a relatively large amount of noise and hence they are more suited to identifying PWMs than accurately ranking individual sequences. In our studies we have a different aim. We are more interested in accurately ranking sequences for which a PWM is known for a specific TF. Additionally, our motifs are 13-15 nt in length, so we cannot use the 8nt universal design. Therefore in this study we have used the strategy of starting with a known consensus, literature-derived, and ChIP-on-chip mined sequences that would potentially bind the TF in question (i.e. HNF4α). We then varied those sequences to generate a total of 3,000 unique sequences, replicated five times each on the array. Thus, the arrays are custom designed to rank sequences that HNF4α would potentially prefer. We then used an SVM to discover "rules" for HNF4α binding to derive subsequent generations of PBMs. A complete description of the methods is given in Chapter 2. Clearly, the problem of determining TFBS and target genes for any TF is quite complex and far from solved. Complete knowledge of TFBS for every TF is highly valuable. It would allow us to potentially predict gene regulation "de novo" for any genome and allow us to identify regulatory regions and modules. This information in turn will result in a greater understanding of gene regulation in a healthy organism and disregulation that leads to disease, paving the way for potential new cures to diseases. This dissertation will aim to create a comprehensive framework for studying TFBS and target genes, focusing on the proteins, statistical and experimental techniques introduced here.

Bibliography

[1] M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117:185–198, 2004.

[2] M. F. Berger and M. L. Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc*, 4:393–411, 2009.

[3] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, and M. L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–1435, Nov 2006.

[4] A. A. Bogan, Q. Dallas-Yang, J. Ruse, M. D., Y. Maeda, G. Jiang, L. Nepomuceno, T. S. Scanlan, F. E. Cohen, and F. M. Sladek. Analysis of protein dimerization and ligand binding of orphan receptor HNF4alpha. *J Mol Biol*, 302:831–51, 2000.

[5] E. Bolotin, J. Schnabl, and F. Sladek. HNF4A (Homo sapiens). *Transcription Factor Encyclopedia http://www.cisreg.ca/tfe*, 2009, 2008.

[6] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA, 1992. ACM.

[7] N. Brianon and M. C. Weiss. In vivo role of the HNF4alpha AF-1 activation domain revealed by exon swapping. *EMBO J*, 25(6):1253–1262, Mar 2006.

[8] E. I. Campos and D. Reinberg. Histones: annotating chromatin. *Annu Rev Genet*, 43:559–599, 2009.

[9] J. A. Carew, E. S. Pollak, S. Lopaciuk, and K. A. Bauer. A new mutation in the HNF4 binding region of the factor VII promoter in a patient with severe factor VII deficiency. *Blood*, 96:4370–2, 2000.

[10] J. Chen, H. K. Kinyamu, and T. K. Archer. Changes in attitude, changes in latitude: nuclear receptors remodeling chromatin to regulate transcription. *Mol Endocrinol*, 20(1):1–13, Jan 2006.

[11] M. K. Das and H.-K. Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8 Suppl 7:S21, 2007.

[12] T. Drewes, S. Senkel, B. Holewa, and G. U. Ryffel. Human hepatocyte nuclear factor 4 isoforms are encoded by distinct and differentially expressed genes. *Mol Cell Biol*, 16(3):925–931, Mar 1996.

[13] S. A. Duncan, K. Manova, W. S. Chen, P. Hoodless, D. C. Weinstein, R. F. Bachvarova, and J. E. Darnell. Expression of transcription factor HNF-4 in the extraembryonic endoderm, gut, and nephrogenic tissue of the developing mouse embryo: HNF-4 is a marker for primary endoderm in the implanting blastocyst. *Proc Natl Acad Sci U S A*, 91(16):7598–7602, Aug 1994.

[14] T. Egener, E. Roulet, M. Zehnder, P. Bucher, and N. Mermod. Proof of concept for microarray-based detection of DNA-binding oncogenes in cell extracts. *Nucleic Acids Res*, 33(8):e79, 2005.

[15] H. Furuta, N. Iwasaki, N. Oda, Y. Hinokio, Y. Horikawa, K. Yamagata, N. Yano, J. Sugahiro, M. Ogata, H. Ohgawara, Y. Omori, Y. Iwamoto, and G. I. Bell. Organization and partial sequence of the hepatocyte nuclear factor-4 alpha/MODY1 gene and identification of a missense mutation, R127W, in a Japanese family with MODY. *Diabetes*, 46(10):1652–1657, Oct 1997.

[16] M. M. Garner and A. Revzin. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res*, 9(13):3047–3060, Jul 1981.

[17] C. K. Glass, J. M. Holloway, O. V. Devary, and M. G. Rosenfeld. The thyroid hormone receptor binds with opposite transcriptional effects to a common sequence motif in thyroid hormone and estrogen response elements. *Cell*, 54(3):313–323, Jul 1988.

[18] M. R. Green. Eukaryotic transcription activation: right on target. *Mol Cell*, 18:399–402, 2005.

[19] M. Hadzopoulou-Cladaras, E. Kistanova, C. Evagelopoulou, S. Zeng, C. Cladaras, and J. A. Ladias. Functional domains of the nuclear receptor hepatocyte nuclear factor 4. *J Biol Chem*, 272:539–550, 1997.

[20] S. Hahn. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol*, 11(5):394–403, May 2004.

[21] L. W. Harries, J. M. Locke, B. Shields, N. A. Hanley, K. P. Hanley, A. Steele, P. R. Njlstad, S. Ellard, and A. T. Hattersley. The diabetic phenotype in HNF4A mutation

carriers is moderated by the expression of HNF4A isoforms from the P1 promoter during fetal development. *Diabetes*, 57(6):1745–1752, Jun 2008.

[22] G. P. Hayhurst, Y. H. Lee, G. Lambert, J. M. Ward, and F. J. Gonzalez. Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. *Mol Cell Biol*, 21:1393–403, 2001.

[23] D. T. Holloway, M. Kon, and C. Delisi. Machine learning for regulatory analysis and transcription factor target prediction in yeast. *Syst Synth Biol*, 1(1):25–46, Mar 2007.

[24] J. Huang, V. Karakucuk, L. L. Levitsky, and D. B. Rhoads. Expression of HNF4alpha variants in pancreatic islets and Ins-1 beta cells. *Diabetes Metab Res Rev*, 24(7):533–543, Oct 2008.

[25] J. Huang, L. L. Levitsky, and D. B. Rhoads. Novel P2 promoter-derived HNF4alpha isoforms with different N-terminus generated by alternate exon insertion. *Exp Cell Res*, 315(7):1200–1211, Apr 2009.

[26] F. Huber, M. Hegner, C. Gerber, H.-J. Gntherodt, and H. P. Lang. Label free analysis of transcription factors using microcantilever arrays. *Biosens Bioelectron*, 21(8):1599–1605, Feb 2006.

[27] G. Jiang, L. Nepomuceno, K. Hopkins, and F. M. Sladek. Exclusive homodimerization of the orphan receptor hepatocyte nuclear factor 4 defines a new subclass of nuclear receptors. *Mol Cell Biol*, 15:5131–43, 1995.

[28] G. Jiang and F. M. Sladek. The DNA binding domain of hepatocyte nuclear factor 4 mediates cooperative, specific binding to DNA and heterodimerization with the retinoid X receptor alpha. *J Biol Chem*, 272:1218–25, 1997.

[29] K. H. Kaestner. Making the liver what it is: the many targets of the transcriptional regulator HNF4alpha. *Hepatology*, 51(2):376–377, Feb 2010.

[30] M. Karin, A. Haslinger, H. Holtgreve, R. I. Richards, P. Krauter, H. M. Westphal, and M. Beato. Characterization of DNA sequences through which cadmium and glucocorticoid hormones induce human metallothionein-IIA gene. *Nature*, 308(5959):513–519, 1984.

[31] S. Khorasanizadeh and F. Rastinejad. Nuclear-receptor interactions on DNA-response elements. *Trends Biochem Sci*, 26(6):384–390, Jun 2001.

[32] F. Koch, F. Jourquin, P. Ferrier, and J.-C. Andrau. Genome-wide RNA polymerase II: not genes only! *Trends Biochem Sci*, 33(6):265–273, Jun 2008.

[33] R. D. Kornberg. Mediator and the mechanism of transcriptional activation. *Trends Biochem Sci*, 30(5):235–239, May 2005.

[34] W. L. Kraus and J. Wong. Nuclear receptor-dependent transcription with chromatin. Is it all about enzymes? *Eur J Biochem*, 269(9):2275–2283, May 2002.

[35] K. C. Lee and W. L. Kraus. Nuclear receptors, coactivators and chromatin: new approaches, new insights. *Trends Endocrinol Metab*, 12(5):191–197, Jul 2001.

[36] J. Li, G. Ning, and S. A. Duncan. Mammalian hepatocyte differentiation requires the transcription factor HNF-4alpha. *Genes Dev*, 14:464–74, 2000.

[37] J. Linnell, R. Mott, S. Field, D. P. Kwiatkowski, J. Ragoussis, and I. A. Udalova. Quantitative high-throughput analysis of transcription factor binding specificities. *Nucleic Acids Res*, 32:e44, 2004.

[38] P. Lu, G. B. Rha, M. Melikishvili, G. Wu, B. C. Adkins, M. G. Fried, and Y. I. Chi. Structural basis of natural promoter recognition by a unique nuclear receptor, HNF4alpha. Diabetes gene product. *J Biol Chem*, 283:33685–97, 2008.

[39] W. M. Macfarlane. Demystified.... Transcription. *Mol Pathol*, 53(1):1–7, Feb 2000.

[40] J. Majka and C. Speck. Analysis of protein-DNA interactions using surface plasmon resonance. *Adv Biochem Eng Biotechnol*, 104:13–36, 2007.

[41] D. J. Mangelsdorf, C. Thummel, M. Beato, P. Herrlich, G. Schtz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon, and R. M. Evans. The nuclear receptor superfamily: the second decade. *Cell*, 83(6):835–839, Dec 1995.

[42] J. S. Mattick and I. V. Makunin. Small regulatory RNAs in mammals. *Hum Mol Genet*, 14 Spec No 1:R121–R132, Apr 2005.

[43] J. S. Mattick and I. V. Makunin. Non-coding RNA. *Hum Mol Genet*, 15 Spec No 1:R17–R29, Apr 2006.

[44] S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young, and M. L. Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 36(12):1331–1339, Dec 2004.

[45] M. A. Navas, E. J. Munoz-Elias, J. Kim, D. Shih, and M. Stoffel. Functional characterization of the MODY1 gene mutations HNF4(R127W), HNF4(V255M), and HNF4(E276Q). *Diabetes*, 48:1459–65, 1999.

[46] W. S. Noble. What is a support vector machine? *Nat Biotechnol*, 24(12):1565–1567, Dec 2006.

[47] D. T. Odom, N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303:1378–81, 2004.

[48] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins. How many drug targets are there? *Nat Rev Drug Discov*, 5(12):993–996, Dec 2006.

[49] B. Panning and D. J. Taatjes. Transcriptional regulation: it takes a village. *Mol Cell*, 31(5):622–629, Sep 2008.

[50] A. A. Philippakis, A. M. Qureshi, M. F. Berger, and M. L. Bulyk. Design of compact, universal DNA microarrays for protein binding microarray experiments. *J Comput Biol*, 15(7):655–665, Sep 2008.

[51] M. J. Reijnen, F. M. Sladek, R. M. Bertina, and P. H. Reitsma. Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. *Proc Natl Acad Sci U S A*, 89:6300–3, 1992.

[52] A. Saunders, L. J. Core, and J. T. Lis. Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol*, 7(8):557–567, Aug 2006.

[53] F. Sladek and S. Seidel. Hepatocyte nuclear factor 4alpha. In T. Burris and E. McCabe, editors, *Nuclear Receptors and Genetic Diseases*, pages 309–361. Academic Press, London, 2001.

[54] F. M. Sladek, W. M. Zhong, E. Lai, and J. Darnell, J. E. Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. *Genes Dev*, 4:2353–65, 1990.

[55] J. Sonoda, L. Pei, and R. M. Evans. Nuclear receptors: decoding metabolic disease. *FEBS Lett*, 582(1):2–9, Jan 2008.

[56] M. Stam, C. Belele, J. E. Dorweiler, and V. L. Chandler. Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. *Genes Dev*, 16(15):1906–1918, Aug 2002.

[57] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16:16–23, 2000.

[58] M. C. Thomas and C.-M. Chiang. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol*, 41(3):105–178, 2006.

[59] A. Tomovic and E. J. Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics*, 23(8):933–941, Apr 2007.

[60] M. Towsey, P. Timms, J. Hogan, and S. A. Mathews. The cross-species prediction of bacterial promoters using a support vector machine. *Comput Biol Chem*, 32(5):359–366, Oct 2008.

[61] K. Umesono, K. K. Murakami, C. C. Thompson, and R. M. Evans. Direct repeats as selective response elements for the thyroid hormone, retinoic acid, and vitamin D3 receptors. *Cell*, 65(7):1255–1266, Jun 1991.

[62] P. Walker, J. E. Germond, M. Brown-Luedi, F. Givel, and W. Wahli. Sequence homologies in the region preceding the transcription initiation site of the liver estrogen-responsive vitellogenin and apo-VLDLII genes. *Nucleic Acids Res*, 12(22):8611–8626, Nov 1984.

[63] C. L. Warren, N. C. S. Kratochvil, K. E. Hauschild, S. Foister, M. L. Brezinski, P. B. Dervan, G. N. Phillips, and A. Z. Ansari. Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci U S A*, 103(4):867–872, Jan 2006.

[64] A. J. Watt, W. D. Garrison, and S. A. Duncan. HNF4: a central regulator of hepatocyte differentiation and function. *Hepatology*, 37:1249–53, 2003.

[65] E. Wingender. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform*, 9(4):326–332, Jul 2008.

[66] E. Wingender, P. Dietze, H. Karas, and R. Knuppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 24:238–41, 1996.

[67] G. K. Wong, D. A. Passey, and J. Yu. Most of the human genome is transcribed. *Genome Res*, 11(12):1975–1977, Dec 2001.

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

[69] Y. Zhao, D. Granas, and G. D. Stormo. Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12):e1000590, Dec 2009.

[70] Q. Zhou and J. S. Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–916, Apr 2004.

[71] A. Zien, G. Rtsch, S. Mika, B. Schlkopf, T. Lengauer, and K. R. Mller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, Sep 2000.

Whenever people agree with me I always feel I must be wrong.

Oscar Wilde

Chapter 2

Integrated Approach for the Identification of Human HNF4α Target Genes Using Protein Binding Microarrays

*Text for this chapter has been taken from a manuscript of the same title, published in Hepatology. 2010 Feb **51** (2):642-53 and coauthored with Hailing Liao, Tuong Chi Ta, Chuhu Yang, Wendy Hwang-Verslues, Jane R. Evans, Tao Jiang, and Frances Sladek*

2.1 Abstract

Hepatocyte nuclear factor 4 alpha (HNF4α), a member of the nuclear receptor superfamily, is essential for liver function and linked to several diseases including diabetes, hemophilia, atherosclerosis and hepatitis. While many DNA response elements and target genes have been identified for HNF4α, the complete repertoire of binding sites and target genes in the human genome is unknown. Here, we adapt protein binding microarrays (PBMs) to examine the DNA binding characteristics of two HNF4α species (rat and human) and isoforms (HNF4α2 and HNF4α8) in a high throughput fashion. We identified ~1,400 new binding sequences and used this dataset to successfully train a Support Vector Machine (SVM) model that predicts an additional ~10,000 unique HNF4α binding sequences; we also identify new rules for HNF4α DNA binding. We performed expression profiling of an HNF4α RNAi knockdown in HepG2 cells and compared the results to a search of the promoters of all human genes with the PBM and SVM models, as well as published genome-wide location analysis. Using this integrated approach, we identified ~240 new direct HNF4α human target genes, including new functional categories of genes not typically associated with HNF4α, such as cell cycle, immune function, apoptosis, stress response and other cancer-related genes. In conclusion, we report the first use of PBMs with a full length liver-enriched transcription factor and greatly expand the repertoire of HNF4α binding sequences and target genes, thereby identifying new functions for HNF4α. We also establish a web-based tool, HNF4 Motif Finder, that can be used to identify potential HNF4α binding sites in any sequence.

2.2 Introduction

Hepatocyte nuclear factor 4α, HNF4α (HNF4A) is a member of the nuclear receptor superfamily of ligand-dependent transcription factors (NR2A1) and a liver-enriched transcription factor (TF) that is also expressed in the kidney, pancreas, intestine, colon and stomach [5]. Originally identified based on its ability to bind DNA response elements in the human apolipoprotein C3 (*APOC3*) and mouse transthyretin (Ttr) promoters [37], HNF4α has since been shown to play a critical role in both the development of the embryo and the adult liver [18, 40]. Mutations in the *HNF4A* coding sequence and promoter regions are linked to Maturity Onset Diabetes of the Young 1 (MODY1) [17], and mutations in HNF4α response elements have been directly linked to disease, most notably in genes encoding blood coagulation factors in hemophilia and in HNF1α in MODY3 [13, 35, 36] . Through classical promoter analysis, functional HNF4α binding sites have been identified in >140 genes, including those involved in the metabolism of glucose, lipids and amino acids, as well as xenobiotics and drugs [5, 16, 40] (see Supplemental Table 6.1 for a listing of those genes). However, recent genome-wide location analyses suggest that the number of HNF4α targets may be much greater (>1000) based on widespread binding of HNF4α to promoter regions [31, 32, 34], although it is not known how many of those are functional targets. A more comprehensive list of direct HNF4α targets was recently made even more critical with our finding that HNF4α binds an exchangeable ligand and hence may be a potential drug target [43].

HNF4α binds DNA exclusively as a homodimer [4, 21]. The canonical HNF4α consensus sequence consists of the half site AGGTCA with one nucleotide spacer (referred to as a DR1, AGGTCAxAGGTCA) [23] . Whereas the number of experimentally verified HNF4α binding sequences is sizeable (>217) (see Supplemental Tables 6.1 and 6.2), they were derived in a biased fashion building on the first HNF4α binding sites [37], and subsequently on the direct repeat rules for nuclear receptor DNA binding [23] . Furthermore, the total number of 13-mer permutations is much greater than 217 (413 ~ 67 million), and whereas HNF4α will certainly not bind all potential 13-mers, the total number of DNA sequences that will bind HNF4α is anticipated to be in the 10,000's. Since the presence of one or more HNF4α response elements in the promoter region of a gene is a prerequisite for classification as a direct HNF4α target, it is desirable to accurately predict all the HNF4α binding sites throughout the genome in an unbiased fashion.

Recent genome-wide technologies, most notably genome-wide location analysis (i.e., chromatin immunoprecipitation (ChIP) followed by tiling arrays, ChIP-chip) and expression profiling, have greatly accelerated the identification of target genes for many TFs, including HNF4α. However, as powerful as those technologies are, they provide information only about the state of the cells used in the assay, not about any other physiological state. Furthermore, expression profiling cannot indicate whether a gene is a direct or an indirect target and ChIP does not provide any information about whether the gene is expressed by the bound TF. And neither assay allows one to precisely identify the sequence to which the TF binds. The third tool in the genomic arsenal - computational prediction of target genes - is curiously less developed than the other two. While many attempts have been made at predicting TF binding sites, including our own for HNF4α [14] , this approach still suffers from a lack of sizeable datasets of verified binding sites. To improve the prediction of potential HNF4α target genes, we adapted the protein binding microarray (PBM) technology

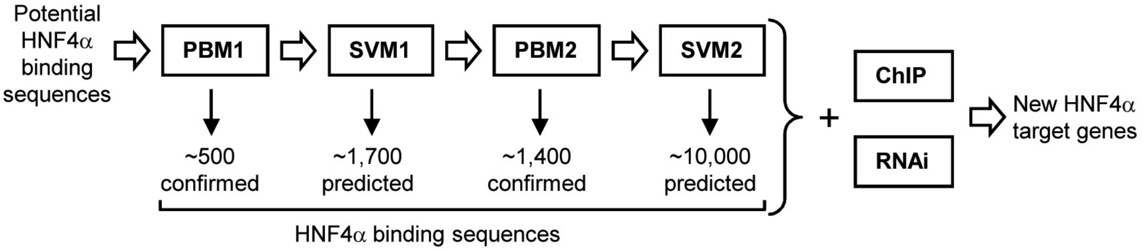


Figure 2.1. Overview of workflow. Known and predicted HNF4α binding sequences (217 sequences from the literature, sites predicted by the Markov model and ChIP-chip analysis, and random controls) were printed on the first generation protein binding microarray (PBM1) and incubated with minimally processed crude nuclear extracts from COS-7 cells transfected with full length HNF4α Results from the initial screen were used to train the support vector machine (SVM1), resulting in 1,700 predicted HNF4α binding sequences that were printed onto a second generation PBM (PBM2), etc. Searches of human promoters using PBM/SVM results were cross referenced with results from RNAi expression profiling and CHIP-chip to identify new HNF4α targets.

to rank thousands of HNF4α sequences based on their relative binding affinities using full length protein expressed in mammalian cells. We compare two species of HNF4α (rat and human) and two tissue-specific isoforms (HNF4α2 and HNF4α8). Additionally, we use a Support Vector Machine (SVM), a powerful machine learning model to predict additional HNF4α binding sequences with high accuracy. Finally, we combine the PBM and SVM binding site searches with expression profiling performed here and ChIP-chip performed by others to identify ~240 new direct target genes of HNF4α in cells of hepatic origin (see Fig. 2.1 for an overview).

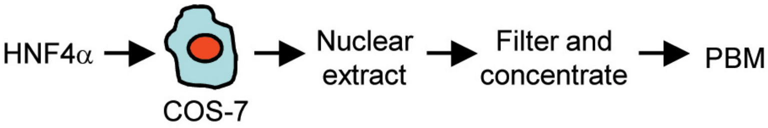


Figure 2.2. Minimally processed crude nuclear extracts were used on PBM. This processing allows the HNF4α to be maintained in similar to *in vivo* conditions.

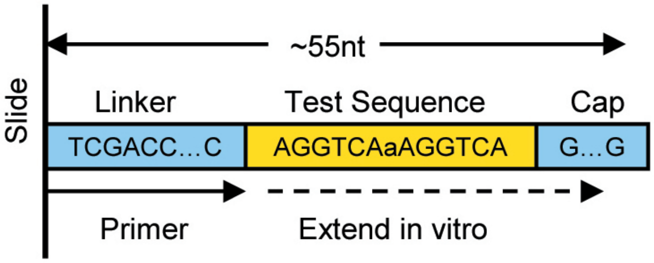


Figure 2.3. Oligo design. Single-stranded oligonucleotides with a common linker, site-specific and a G/C-rich cap region printed on the PBM were extended in vitro in the presence of Cy3-dUTP. Test sequence or variable region is indicated in yellow.

2.3 Materials and Methods

2.3.1 Preparation of HNF4α Proteins in COS-7 cells

Nuclear extracts were prepared from COS-7 cells transiently transfected with HNF4a expression vectors as previously described [21]. Mock-transfected samples contained no DNA. Crude nuclear extracts were filtered and concentrated using Microcon Ultracel YM-30 filter (Millipore, Bedford, MA) and applied directly to the PBM (Fig. 2.2 , except for purified samples that were immunoprecipitated from the crude extracts with the α445 antibody [37] (Fig. 2.6 and then peptide-eluted.

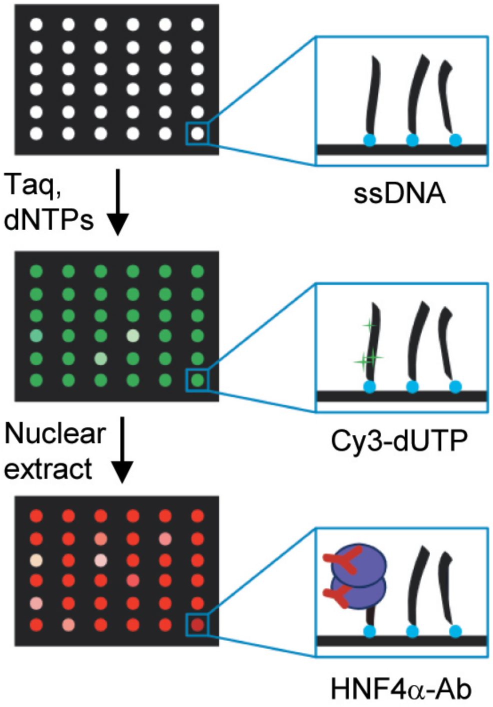


Figure 2.4. Overview of experimental steps. Single stranded DNA arrays are ordered from manufacturer, converted to double stranded using polymerase primer extension with Cy-dUTP incorporation, incubated with extracts containing HNF4α and visualized by immunoflouorescence.

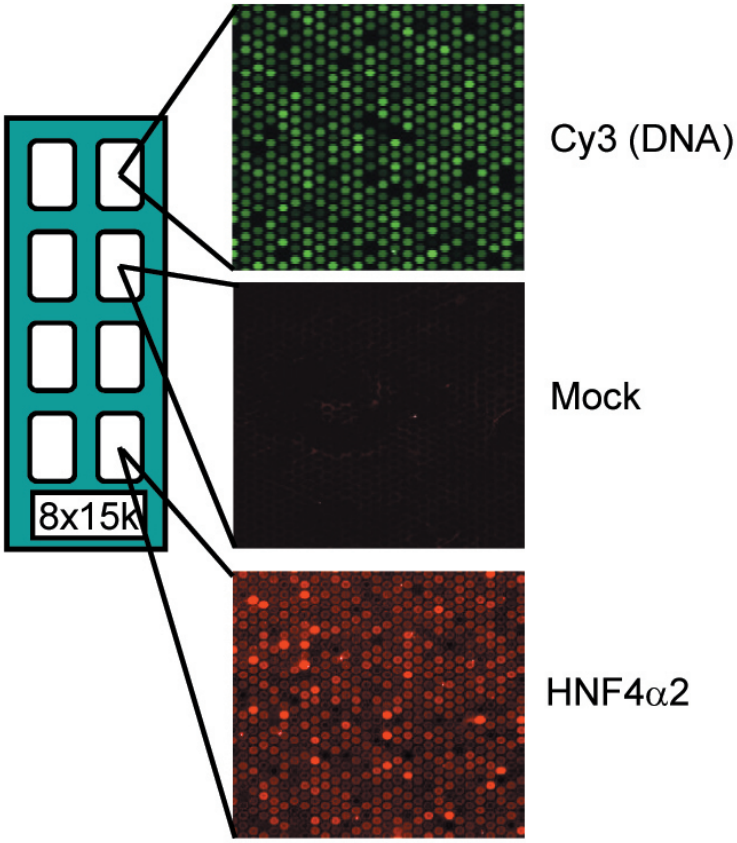


Figure 2.5. Typical PBM results. Double-stranded DNA with Cy3 incorporated (top panel), mock-transfected cells lacking HNF4α (middle panel) and extracts containing HNF4α with fluorescent signal proportional to the binding affinity (bottom panel). 8x15k, Agilent microarray slide with 8 replicate subarrays with ~3000 unique sequences each spotted 5 times (~15,000 spots) per subarray. Supplemental Figure 2.25 shows that non transfected COS-7 cells do not express HNF4α and that the antibody used to detect HNF4α in the PBM is completely specific. Supplemental Figure 2.26 shows a linear relationship between Cy3 incorporation and the number of A's in the extended sequence.

2.3.2 Reagents

The expression vector pMT7.rHNF4α2 containing wild type (wt) rat HNF4α2 (NM_022180), the predominant isoform in liver, has been previously described [43], as have the vectors containing the human HNF4α2 (NM_000457) (pcDNA3.1.hHNF4α2) and HNF4α8 (pcDNA3.1.hHNF4α8) [9, 12] and the affinity purified antibody to the very C-terminus of HNF4α (α445) [37]. Mouse monoclonal antibodies to the N-terminus (αNTD) (PP-K9218-00) and C-terminus of HNF4α2 (αCTD) (PP-H1415-00) and secondary antibodies conjugated to Cy5 (Northern Lights 637 Fluorochrome-labeled donkey anti-mouse or anti-rabbit, #NL008 or NL005, respectively) were purchased from R&D Systems (Minneapolis, MN).

2.3.3 Cell Culture Conditions

Monkey kidney cells (COS-7, ATCC #CRL-1651) were maintained in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% Bovine Calf Serum (BCS). Human hepatocellular carcinoma/hepatoblastoma cell line (HepG2, ATCC# HB-8065) were maintained in DMEM containing 10% Fetal Bovine Serum (FBS), 1% non-essential amino acids, and 1% sodium pyruvate. All cell lines were supplemented with 1% penicillin and streptomycin and maintained in a 5% CO2 incubator at 37 °C.

2.3.4 Protein Binding Microarray

(PBM) design and primer extension. Custom 8x15k arrays of single-stranded 42- to 51-mer oligonucleotides were manufactured by Agilent Technologies (Santa Clara, CA). Both

PBM1 and PBM2 contained 3000 unique sequences replicated five times. Each sequence begins with a 27-mer linker (5'-TCGACCGATACTCTAATCTCCCTAGGC-3') followed by a variable region of 5 to 14 nucleotides and a 5-nt cap (5'-GCGCG-3') (Fig. 2.3). PBM1 contained a combination of sequences collected from the literature, mined from several ChIP-chip datasets [31, 34], predicted by previously developed models [14], and created from variations on the consensus 5'-AGGTCAaAGGTCA-3'. Random controls and Sp1 sites were included on the array to account for nonspecific and indirect binding (for a complete list of sequences on PBM1, see Supplemental Table 6.3). PBM2 contained sequences derived from PBM1, sequences obtained from SVM1 (see below) searches on human promoter regions, and the ChIP-chip data set [31] (for a complete list of sequences on PBM2, see Supplemental Table 6.4). The primer extension reaction was performed using Sequenase 2.0 (USB, Cleveland, OH), dUTP-Cy3 (GE Healthcare, Waukesha, WI), and a common primer (5'-TCGACCGATACTCTAATCTCCC-3') as previously described [7] except for the following modifications: the hybridization chamber was inverted and incubated at 85 °C for 10 min, followed by 75 °C for 10 min, 65 C for 10 min, 50 °C for 10 min, and 55 °C for 90 min. The chamber was disassembled, washed extensively in PBS (pH 7.5) and air dried at room temperature. The dUTP incorporation was visualized as described below.

2.3.5 PBM Application

PBMs were pre-moistened in PBS plus 0.01% Triton-X 100 for 1 min and blocked for 1 hr with PBS plus 2% nonfat dry milk, subsequently washed for 10 min in PBS plus 0.1% Tween-20 and then incubated for 1 hr with protein binding solution (16 mM HEPES, pH

7.8,120 mM KCl, 8 mM EDTA, 8 mM EGTA, 1% Tween-20, 0.25 ng/l of poly-dIdC) and 500 ng HNF4α in crude nuclear extracts from transfected COS-7 cells (see above). The arrays were washed with PBS plus 0.1% Tween-20 for 5 min at low speed on a shaking platform and incubated with primary antibody (αNTD, αCTD or α445) diluted 1:100 in PBS, 2% milk, 0.1% Tween-20 overnight at room temperature. Following incubation with secondary antibody (GaM or GaR conjugated to Cy5) at 1:50 in the same buffer for 1 hr, the arrays were washed 3x in PBS plus 0.1% Tween-20 for 5 min, then 3 min in PBS, air dried and scanned using a GenePix Axon 4000B scanner (Molecular Devices, Sunnyvale, CA) at 543 nm (Cy3) dUTP and 633 nm (Cy5 conjugated secondary antibody). All PBS and milk was filtered through 0.45 mm filters (Corning, Lowell, MA); all other reagents were filtered through 0.33 mm filters (M2135, MoBiTec, Gottingen, Germany). All washes and incubations were performed in an Agilent hybridization chamber at room temperature (27 °C).

2.3.6 PBM Analysis and Data Normalization

All PBMs were scanned using a GenePix Axon 4000B scanner (Molecular Devices, Sunnyvale, CA) at 543 nm (Cy3) to check for even primer extension, and 633 nm (Cy5) to quantify protein binding. Scanning was performed with a 5 mm resolution at optimal laser intensity and at near-saturation of the highest intensity spots. Images were saved as lossless TIFF files and quantified using GenePix 6.0 software (Molecular Devices). Aberrant spots were manually flagged and removed from subsequent analysis. Background-subtracted mean intensities were calculated for remaining spots. The signals were gradient-corrected using Micro-Array NORmalizatation of array-CGH data (MANOR) implemented in R [30] as needed. Cross- and intra-array normalization was performed using quantile normalization [6], enabling comparison between independent experiments. Replicates for each probe were averaged and their coefficient of variation (CV) was calculated. Only probes with a CV less than 0.3 were used for the training set.

2.3.7 SVM Training and Sequence Analysis

The training data from PBM1 was generated by averaging six correlated arrays (Supplemental Table 6.3 and Fig. 2.27); the sequences were ranked based on their relative intensity. A kernel-based support vector machine (KSVM) function from Kernlab package in R with Laplace dot kernel was used to train the SVM1 model [25]. SVM1 was applied in a sliding window approach to classify the 13-mer sequences in all of the annotated human promoters (UCSC hg18) and the ChIP-chip dataset into ”binding” and ”nonbinding” categories [31]. The threshold was empirically adjusted until the false positive rate (~5%) and false negative rate (~5%) were simultaneously achieved in the 10-fold cross validation test. The top predicted binding sequences with a promoter score >0.41 and a ChIP-chip score > 0.25 (resulting in ~1,700 and ~1,500 sequences, respectively) were selected for PBM2 (see Supplemental Table 6.4). A second SVM, which is called SVM2 and uses the regression mode, was trained on three averaged PBM2 experiments Fig. 2.28 and achieved a high correlation with PBM2 in the 10-fold cross validation test (R^2 = 0.75) (Fig. 2.13). The SVM2 model was also used to search the promoter regions (-2kb to +1 kb relative to the transcription start site, +1) of all annotated genes in the human genome (UCSC hg18) following a sliding window approach.

2.3.8 RNA Interference and Expression Profiling Analysis

RNA interference (RNAi) against HNF4α1 was performed in HepG2 cells using siRNAs corresponding to nucleotides +179 to +197 of human *HNF4A* (NM_178849, sense siRNA: 5'-UGUGCAGGUGUUGACGAUGdTdT-3', antisense siRNA 5'-CAUCGUCAACACCUGCACAdTdT-3') purchased from Dharmacon Research, Inc. (An NCBI Blast search indicated that this sequence is unique to *HNF4A*). Approximately 24 hr prior to transfection, cells were plated at a density of ~1.5 x105 or ~2.5 x105 cells/well in a 12- or 6-well plate, respectively (~50-70% confluency) without antibiotics. The siRNAs (100 or 200 pmol, respectively) were introduced into the cells using TransIT-TKO transfection reagent purchased from Mirus Bio Corporation (Madison, WI). Each experiment included a control containing just the transfection reagent as well as a nonspecific siRNA control against firefly luciferase (PGL3: sense siRNA 5'-CUUACGCUGAGUACUUCGAdTdT-3'; antisense siRNA 5'-dTdTGAAUGCGACUCAUGAAGCU). To verify HNF4α protein levels, cells were lysed in RIPA buffer purchased from Santa Cruz Biotechnology and analyzed by SDS-PAGE followed by immunoblotting (Supplemental Fig. reftable:s5). Total RNA was extracted using Trizol Reagent (Invitrogen, Carlsbad, CA) and reverse transcribed using the Reverse Transcription System kit (Promega, Madison, WI). PCR amplification was performed in the linear range using a PTC-100 TM programmable thermal controller (MJ Research, Inc., Hercules, CA). One-fifth of each reaction was electrophoresed on a 2% agarose gel, stained with ethidium bromide and visualized by UV light (see Supplemental Table 6.6 and 6.7 for a list of PCR primers used). Expression profiling analysis was performed with

Affymetrix oligonucleotide arrays (HGU133 Plus 2.0) using RNA from control (PGL3 siRNA) or treated (HNF4α siRNA) HepG2 cells (48 h, 200 pmol siRNA per well of 6-well plate, introduced by Lipofectamine 2000 (Invitrogen)). Arrays were hybridized in biological replicate by the UCR Genomics Core Instrumentation Facility. Results were analyzed by Bioconductor LIMMA package [15].

2.3.9 Gel Shift Conditions and ChIP

Electrophoretic mobility shift analysis (EMSA, or gel shift) was carried out for Figures 2.17, 2.30 and 2.30 essentially as previously described [43] using crude nuclear extracts from COS-7 cells transfected with HNF4α expression vectors and 32P-radiolabelled probes as indicated. Unlabeled competitor oligonucleotides were added prior to the addition of the HNF4α protein in the indicated amounts α445 antibody specific to HNF4α was added ∼15 min after the HNF4α protein. In Figure S7C, competitor oligonucleotides were prepared by USB Sequenase 2.0 (USB) extension as was done for the PBMs and the shift conditions were modified to mimic the PBM conditions (most notably addition of Tween-20 to 0.01%). See Supplemental Table 6.15 for sequences of all oligonucleotides used in gel shift assays. ChIP was preformed as described in [19].

2.4 Results

2.4.1 Protein Binding Microarray (PBM) using full length HNF4α in crude nuclear extracts.

PBMs are a high throughput *in vitro* DNA binding assay that allow for the examination of TF binding to thousands of unique sequences in a single experiment [3]. Recently, PBMs have been used to define the DNA binding specificity of large classes of TFs [1, 2] and have been shown to correlate well with gel shift results [27]. Whereas as others have pioneered the technology using the DNA binding domain (DBD) of TFs purified from bacteria, here we adapt the PBM technology to more closely approximate physiological conditions. Since HNF4α has a very strong dimerization domain outside of the DBD and a very low affinity for DNA when expressed in bacteria [4, 20, 22], we ectopically expressed full length, native HNF4α in COS-7 cells and prepared minimally processed nuclear extracts (Fig. 2.2) that we then applied directly to a PBM specifically designed for HNF4α (Fig. 2.3, 2.4). The PBM was developed with a highly specific antibody to the C-terminus of HNF4α (see Supplemental Figure 2.25), allowing us to examine a completely native TF. The full length HNF4α protein in the crude extracts yielded an excellent signal with a range of intensities, while extracts from mock-transfected cells yielded no reproducible signals (Fig.2.5).

2.4.2 Reproducibility and Utility of Adapted PBM.

We compared two species (rat and human) and two isoforms of HNF4α (HNF4α2 and HNF4α8), as well as antibodies that recognized different regions of HNF4α (Fig. 2.6).

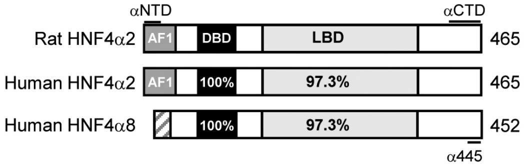


Figure 2.6. Diagram of HNF4α splice variants used in PBM indicating percent amino acid identity in conserved regions. DBD, DNA binding domain; LBD, Ligand Binding Domain; AF1, Activation Function 1. The regions of the protein detected by the monoclonal antibodies (αNTD, amino-terminal HNF4α antisera; αCTD, carboxy-terminal HNF4α antisera) and the affinity purified polyclonal antibody α445 are indicated. (See Materials and Methods for additional details on plasmids and antibodies.)

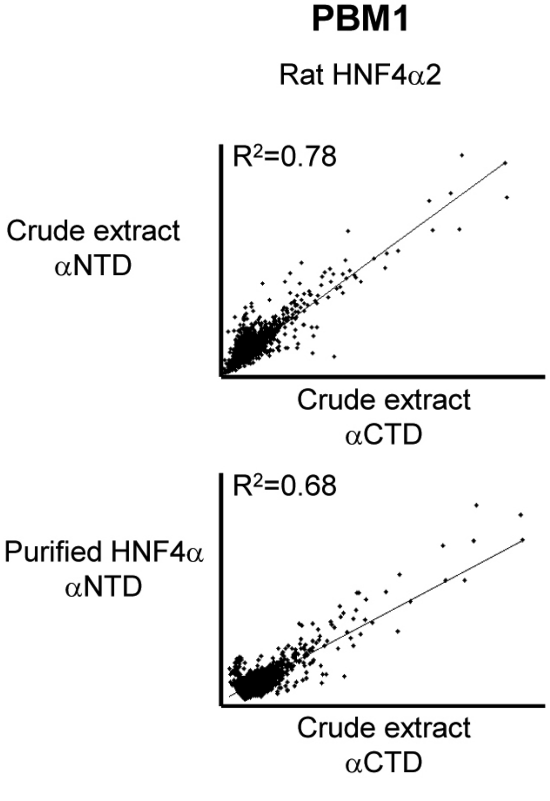


Figure 2.7. Scatter plot of individual spot intensities showing correlation between PBM1 using rat HNF4α2 protein and the αNTD and αCTD antibodies (top panel) as well as purified HNF4α2 versus crude nuclear extracts (bottom panel).

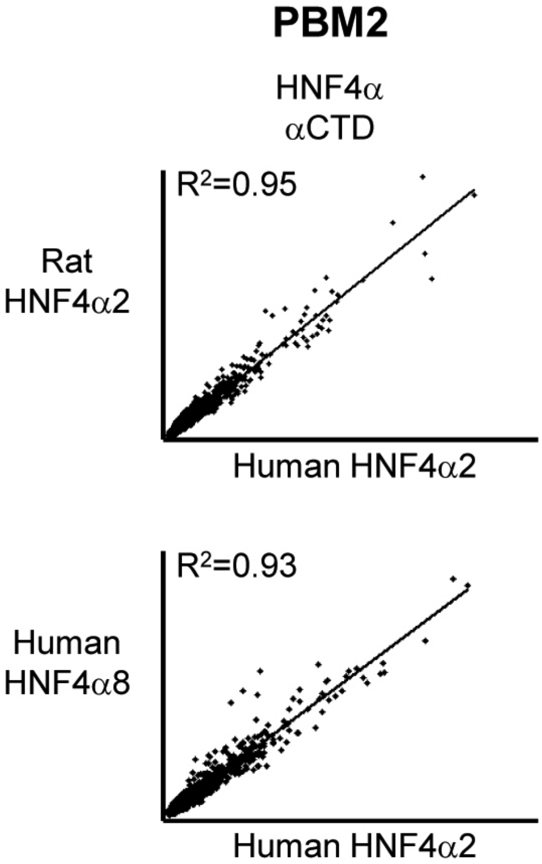


Figure 2.8. Scatter plot of PBM2 results as in **B** comparing different HNF4α isoforms from different species. See Supplemental Figures 2.27 and 2.28 for scatter plot matrices of PBM1 and PBM2 from 9 experiments.

There was an excellent correlation between replicate arrays in the first generation PBM (PBM1) using crude nuclear extracts, regardless of antibody used ($R^2 = 0.78$), and results with affinity purified protein were very similar to those with crude extracts ($R^2 = 0.68$) (Fig. 2.7). In a second generation of the PBM (PBM2), different HNF4α isoforms (HNF4α2 vs. HNF4α8) and species (human vs. rat) also produced excellent correlations (R2 > 0.9), indicating that these isoform and species differences do not influence the binding of HNF4α to DNA. This is not surprising considering that the DBD is identical in these constructs (Fig. 2.6).

2.4.3 Accuracy of PBM and SVM.

PBM1 identified ~500 new HNF4α binding sequences with the DR1-derived sequences exhibiting the best binding affinities relative to negative controls (p < 8.274x10-12) (Fig. 2.9). Sequences derived from ChIP-chip analysis bound roughly as well as the DR1 variants. In PBM2, an additional ~1,000 novel sequences that strongly bind HNF4α were identified, including sequences identified by SVM1. The signal-to-noise ratio (literature-derived vs. random sites) was also significantly improved in PBM2 due to optimization of the binding conditions (p < 2.6 x10-11 vs. p < 2.6 x10-16, respectively, using the Student t-test) (Fig. 2.10). The PBM2 results also correlated very well with gel shift results (Fig. 2.11). Additionally, SVM2 derived from PBM2 predicted binding sequences with a high degree of accuracy ($R^2 = 0.76$) (Fig. 2.12).

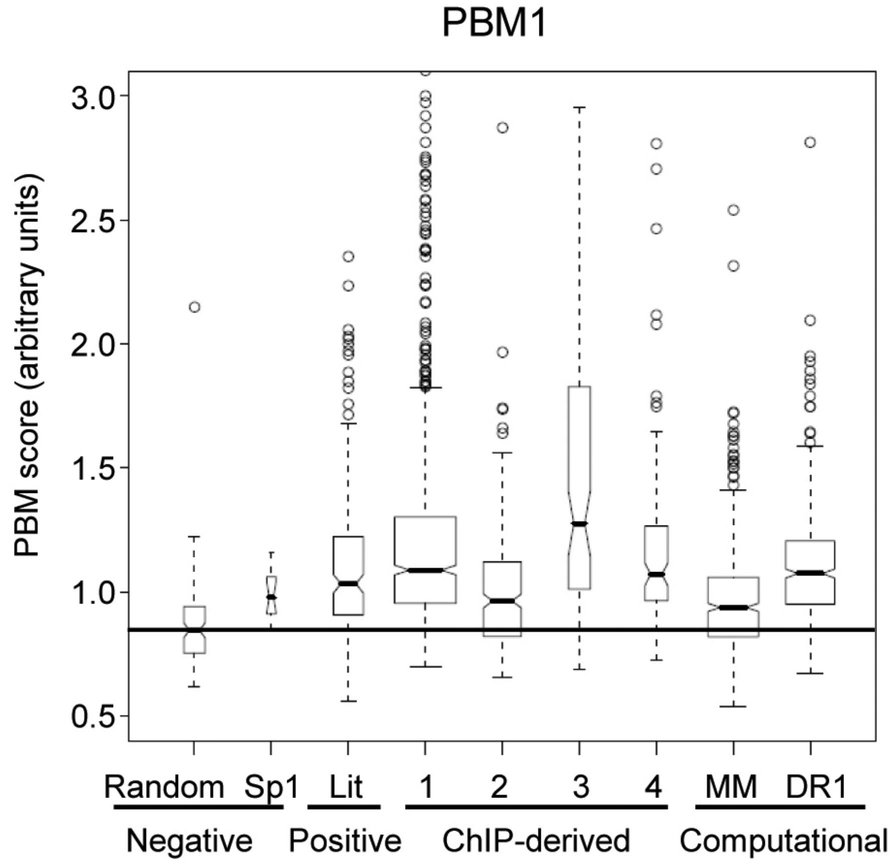


Figure 2.9. Box plot of sequence categories represented on PBM1 and corresponding PBM score from 6 independent arrays with each sequence spotted 5 times. Box width indicates the relative number of sequences per category. Non-overlapping box plot notches strongly indicate that the medians significantly differ ($p < 0.05$). Boxes and whiskers (dashed line) represent quartiles of binding scores for each sequence category. Line, median of random sequences. Negative controls: randomly generated 13-mers; known Sp1 sites derived from the literature. Positive controls, 217 known HNF4α binding sites from the literature (Lit) (Supplemental Tables 6.1 and 6.2). ChIP-derived, binding sites derived from published HNF4α ChIP-chip data: 1, from Odom et al. supplement [31] ; 2, from Rada-Iglesias et al. supplement [34] ; 3, our analysis of Odom et al. data using Bioprospector software; 4, our analysis of Odom et al. data using AlignACE software. Computational, binding sequences derived from our permutated Markov model (MM) [14] and permutations of the DR1 consensus sequence (DR1).

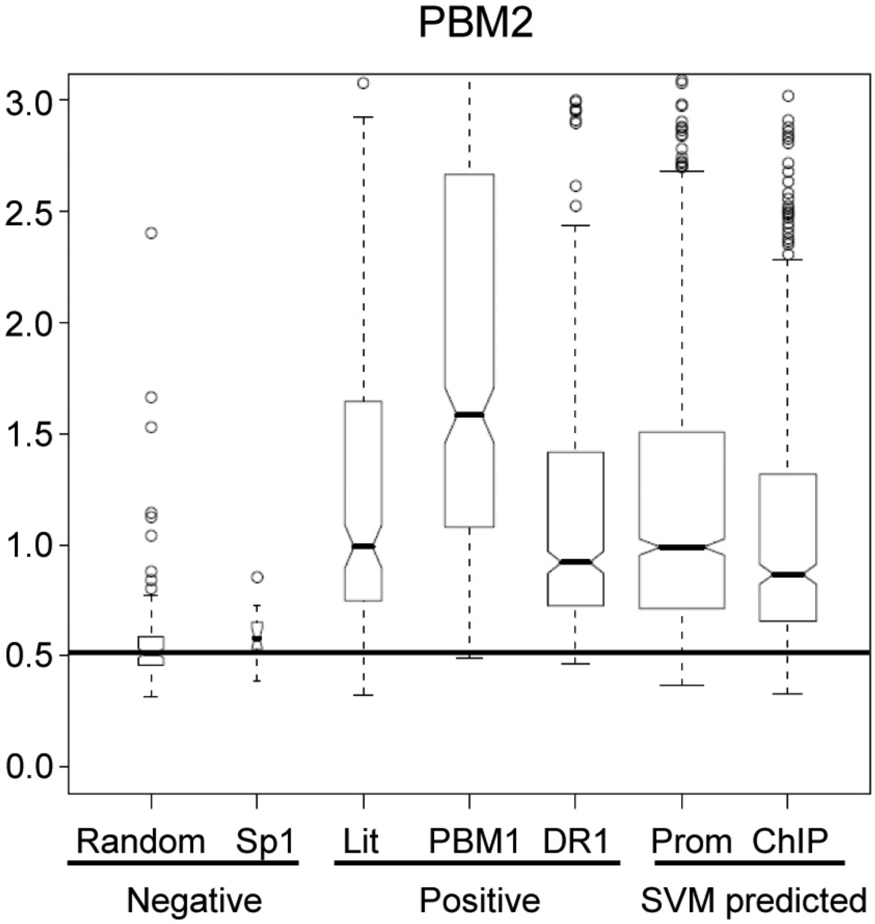


Figure 2.10. (B) Box plot of sequence categories represented in PBM2 (3 independent arrays) as in A. PBM1, best 500 sequences from PBM1; SVM predicted, sequences from SVM1 search of promoter regions of all annotated human genes (Prom) and ChIP-chip data (ChIP) [31] . For a complete list of all the sequences on PBM1 and PBM2 and binding scores see Supplemental Tables 6.3 and 6.4.

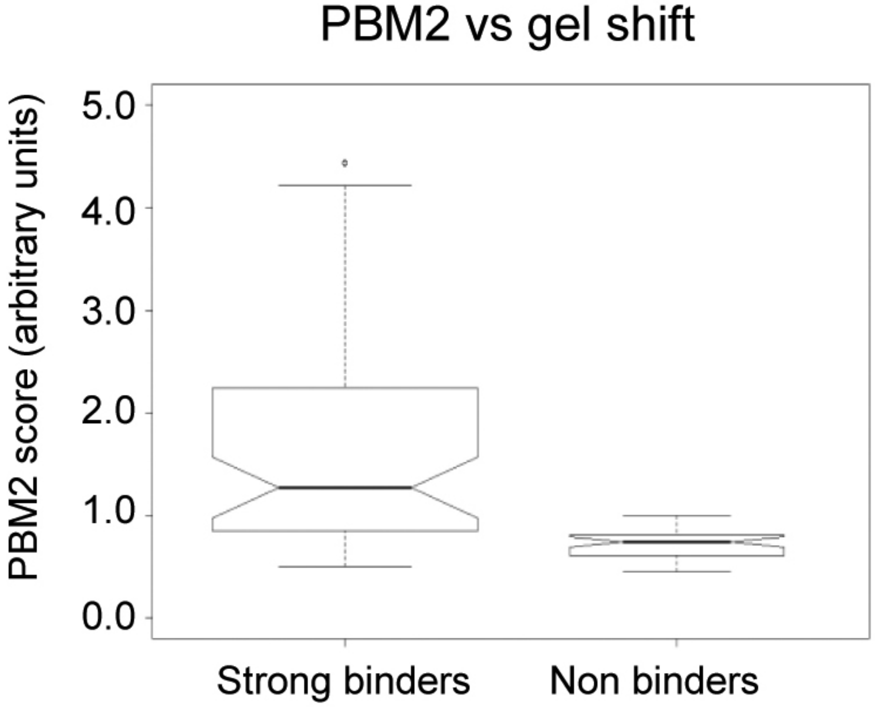


Figure 2.11. PBM2 vs gel shift. Box plot of PBM2 results versus results from ~100 gel shift experiments showing a statistically significant difference (Student t-test, $p<0.00622$) between strong binders and non or very weak binders.

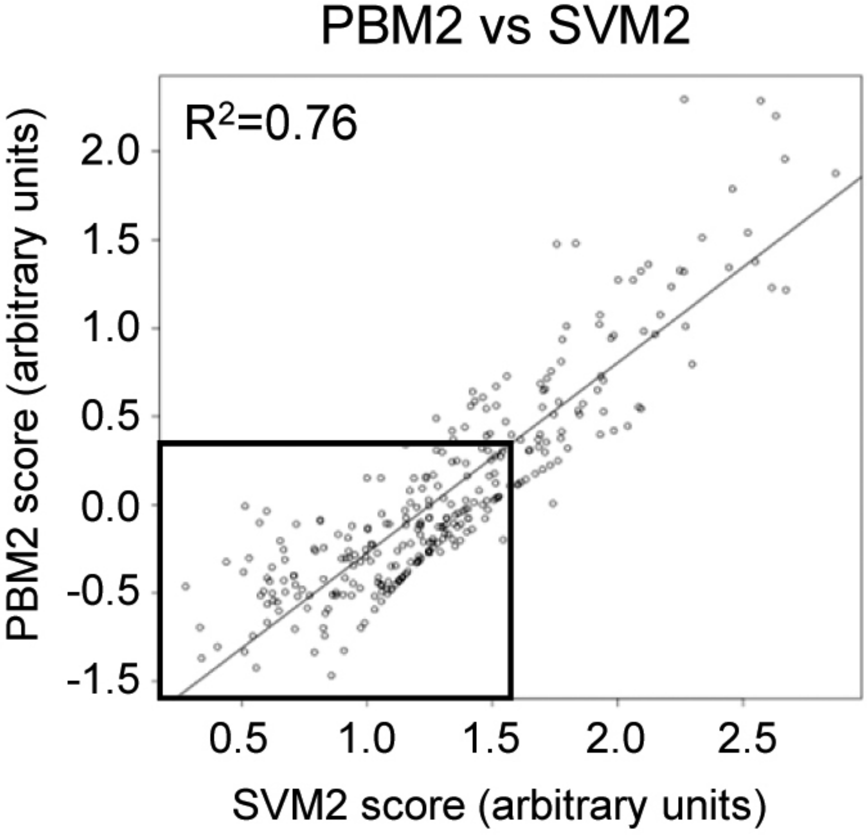


Figure 2.12. Scatter plot of log(PBM2) intensity compared to SVM2 score of one of the 10-fold cross validation results used to evaluate the predictive power of SVM2. A cutoff of an SVM2 score > 1.51, corresponding to 3 standard deviations from the mean of random controls, was used to identify binding sequences in subsequent analyses.

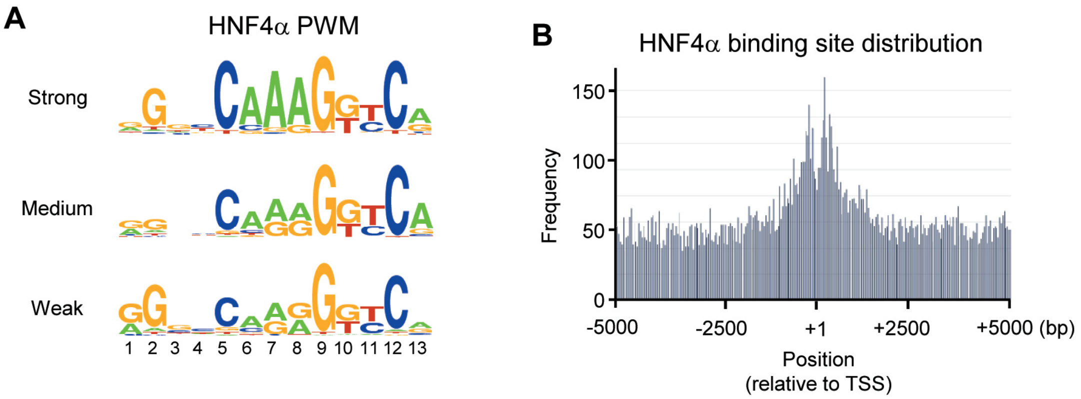


Figure 2.13. Position weight matrix (PWM) for HNF4α binding sequence motif and HNF4α binding site distribution. (A) Position weight matrix (PWM) of HNF4α binding sequences derived from PBM2. All sequences with relative binding affinity at least 2 standard deviations above the mean of the random controls were divided into 3 groups of ~450 each – strong, medium and weak - and used to generate the PWMs [10]. (B) Distribution of potential HNF4α binding sites around the transcription start site (TSS, +1) of all human promoters (UCSC hg18) as determined by an exact match search with PBM2 results. Sites are over represented in the -1 kb to +1 kb region. (See Supplemental Fig. 2.31 for PWM and gel shifts of noncanonical binding sites detected in the PBM.)

2.4.4 Identification of New "Rules" for HNF4α DNA binding by PBM.

Even though position weight matrices (PWMs) do not capture the interdependence between the positions in a motif as do PBMs and SVMs, they are useful for describing motifs. Interestingly, the PWM of the ~450 sequences that yielded the greatest binding intensity in PBM2 ("strong binders") did not strictly follow the DR1 rule of AGGTCAxAGGTCA. Rather, a core sequence of CAAAG is the most prominent feature, with the classical AGGTCA half-site evident only on the 3' side (Fig. 2.13A), a finding supported by the recent crystallographic structure of the HNF4α DBD on DNA in which fewer hydrogen bonds were observed between the HNF4α protein and the 5' half site [28]. In the PWMs for the medium and weak binding motifs, the 3A's in the core appeared less frequently. Using ~1,400 strong HNF4α binding sequences obtained from PBM2, we determined the distribution of potential HNF4α binding sites in the human genome and found a broad distribution of sites with an enrichment within ~1 kb of the transcription start site (+1) (Fig. 2.13. This is in contrast to profiles of sites for some other TFs, such as Sp1 and ELK1, that are found more exclusively near +1 [41] but consistent with the fact that there are many well characterized HNF4α sites far from +1. We also found a small percentage (<1%) of sites that bound HNF4α well in PBM2 but did not contain the CAAAG core (see Supplemental Fig. 2.31 for the PWM and gel shift), but the biological relevance of these sequences remains to be verified.

2.4.5 Expression Profiling of an HNF4α RNAi Knockdown in Hepatic Cells.

To identify functional HNF4α target genes, we used RNAi to knock down HNF4α2 expression in HepG2 cells, a human hepatocellular carcinoma cell line that expresses endogenous HNF4α and many liver-specific genes (Fig. 2.14 top panels and Supplemental Fig. 2.29). Using the SVM2 model, we predicted several other potential HNF4α target genes and determined that they were also down regulated by RT-PCR (APOC4, RDH16, APOM, APOH, SPSB2, UBD, ZDHHC11) (Fig. 2.14 bottom panel). Whole genome expression profiling identified ~1500 additional genes that were down regulated (see Supplemental Table 6.5 for a complete list). Interestingly, the gene that was down regulated the most - Ninjurin 1 (NINJ1) (12.5-fold) - is not a gene typically associated with HNF4α function (i.e., intermediary metabolism); rather, it is involved in regulating the cell cycle. In order to determine whether NINJ1 is a direct target of HNF4α, we used SVM2 to identify a potential HNF4α binding site within the NINJ1 promoter region (Fig. 2.15) and subsequently verified that it was bound by HNF4α in vivo using a ChIP assay (Fig. 2.16) and in vitro using a gel shift assay (Fig. 2.17), suggesting that NINJ1 is indeed a direct target of HNF4α.

2.4.6 Gene Ontology analysis reveals complementary nature of PBM, expression profiling and ChIP analysis.

To compare the different methods of predicting target genes, we performed Gene Ontology (GO) on the HNF4α targets predicted by RNAi expression profiling and the PBM2 search (-2kb to +1 kb) as well as published HNF4α ChIP-chip results from primary human

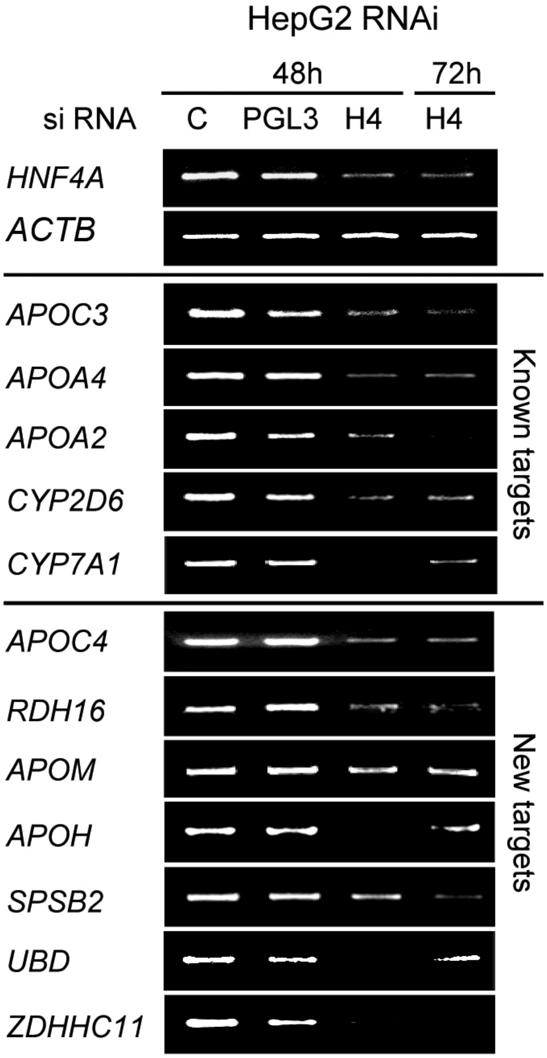


Figure 2.14. Verification of HNF4α1/2 knockdown. HepG2 cells treated with siRNA for the hours indicated. RT-PCR was performed on the indicated HNF4α targets. C, no siRNA. PGL3, control siRNA. H4, HNF4α siRNA (all splice variants from the P1 promoter are targeted).

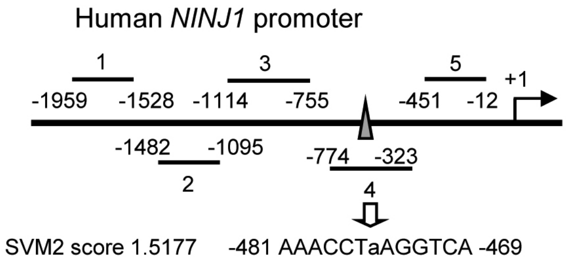


Figure 2.15. Human NINJ1 promoter showing regions amplified by PCR in ChIP in (C). Region 4 contains a predicted HNF4α binding site with an SVM2 score of ~1.5177 (moderate binding affinity). Region 4 contains a predicted HNF4α binding site with an SVM2 score of ~1.5177 (moderate binding affinity).

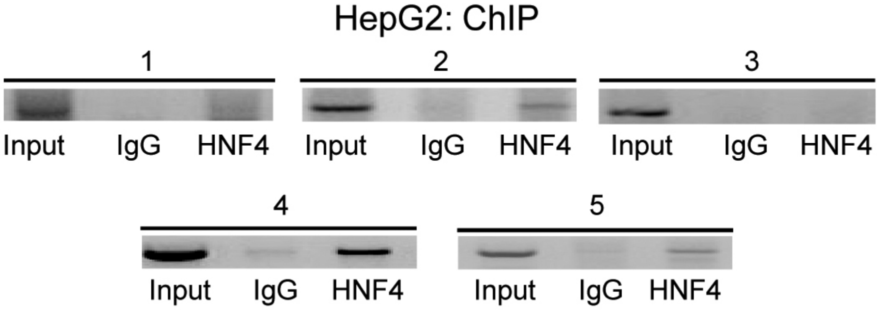


Figure 2.16. ChIP result of HNF4α in HepG2 cells on the human NINJ1 promoter using PCR primers that amplify regions 1-4 noted in (Fig. 2.15). IgG, normal rabbit IgG; HNF4, α445 antibody.

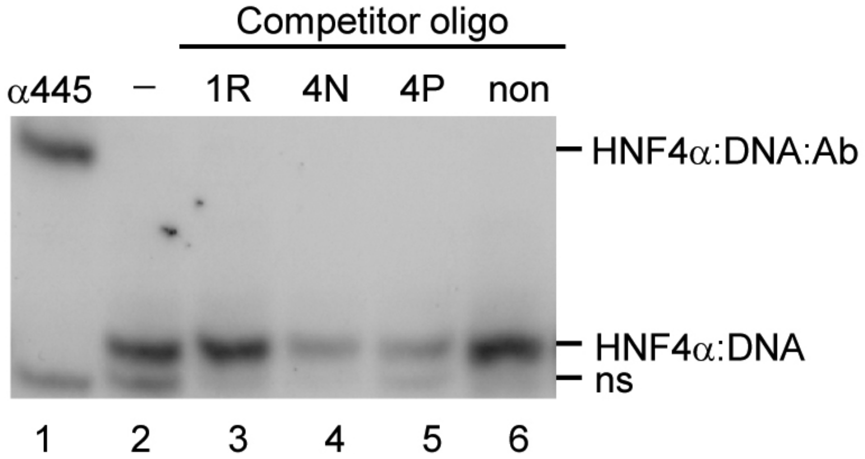


Figure 2.17. Gel shift assay using nuclear extracts from COS-7 cells transfected with rat HNF4α2, radiolabelled probe from the ApoA1 promoter and unlabelled competitors in 250-fold molar excess corresponding to the SVM site identified in region 4 with native flanking sequences (4N) or PBM flanking sequences (4P) as well as a known non binder (non, 175 TTR) and a randomly chosen sequence from region 1 (1R). Shown are the HNF4α:DNA shift complex, a supershift complex with the α445 antibody (HNF4α:DNA:Ab) and nonspecific band from the COS-7 extracts (ns); free probe is not shown. See Supplemental Materials and Methods for details on gel shift conditions, Figure 2.29 for immunoblot of HNF4α protein in the RNAi, Table 6.5 for a complete list of genes down regulated, Table 6.6 and 6.7 for primer sequences and Table 6.15 for gel shift sequences.

hepatocytes [31] Fig. 2.18, 2.19,2.20. In general, six broad biological processes contained significant GO terms for all three assays - metabolism, transport, development, regulation of signal transduction, protein modification, and apoptosis - showing the overlapping nature of the three assays. There were three additional categories - inflammatory response, cell cycle and nucleic acid metabolism - in which genes from at least one but not all three assays were overrepresented. The most notable difference between the PBM2 search from the other assays was an enrichment of genes involved in developmental processes. This is consistent with the known role of HNF4α in early development [26], and could be explained by the fact that the cells used in the ChIP-chip and RNAi assays are from adult, not embryonic, stages. In general, the ChIP assay yielded more significant GO terms in all categories, which is most likely a reflection of the more specific nature of this assay and the stringent cut off values used.

2.4.7 Identification of New HNF4α Target Genes and New Functions.

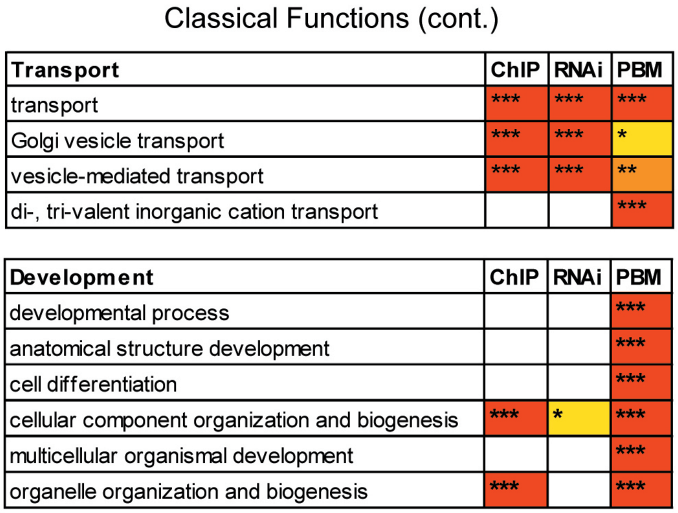
In order to more closely compare the three methods of identifying potential target genes, we cross referenced the PBM2 search results with the HNF4α RNAi and ChIP-chip results. We identified 198 genes that were positive in all three categories - i.e., bound by HNF4α in ChIP-chip, down regulated by HNF4α in HepG2 RNAi and containing one or more verified HNF4α binding sites in the -2kb to +1 kb region of the promoter (Fig. 2.21). A similar analysis with the SVM2 search yielded 135 genes (Fig. 2.22). Among these two categories, there were ~260 nonredundant genes, of which ~240 were not in the original list of HNF4α target genes from the literature (Supplemental Table 6.1). Several of these genes are new targets within known categories of HNF4α targets (e.g., homeostasis - solute

Classical Functions

Metabolism	ChIP	RNAi	PBM
metabolic process	***	***	**
alcohol biosynthetic process	***		
alcohol metabolic process	***	***	***
amine biosynthetic process	**	***	
amine catabolic process	***	*	*
amine metabolic process	***	***	**
amino acid and derivative metabolic process	***	***	
amino acid biosynthetic process	***	***	
amino acid catabolic process	***	**	*
amino acid metabolic process	***	***	*
aminoglycan metabolic process		***	*
aromatic compound metabolic process	***	**	
carbohydrate biosynthetic process	***	**	*
carbohydrate metabolic process	***	***	**
carboxylic acid metabolic process	***	***	**
catabolic process	***	***	*
cellular lipid metabolic process	***	***	***
cellular protein metabolic process	***	**	

Metabolism	ChIP	RNAi	PBM
cofactor metabolic process	***	*	
electron transport	***		
fatty acid metabolic process	***	*	*
fatty acid oxidation	***		
generation of precursor metabolites and energy	***		
glucose metabolic process	***		
glutamine family amino acid metabolic process		***	
hexose metabolic process	***	**	*
lipid biosynthetic process	***	***	**
lipid metabolic process	***	***	***
macromolecule biosynthetic process	***		
monocarboxylic acid metabolic process	***	***	*
monosaccharide biosynthetic process	***		
monosaccharide metabolic process	***	***	*
nitrogen compound metabolic process	***	***	**
organic acid metabolic process	***	***	**
pyruvate metabolic process	***		
steroid biosynthetic process	***	*	*

Figure 2.18. Comparative Gene Ontology for genes bound in vivo by HNF4α (ChIP-chip), down regulated in HNF4α RNAi and containing PBM or SVM HNF4α binding sites. Overrepresented categories from Gene Ontology analysis using DAVID [11] of HNF4α Chip-chip from primary human hepatocytes [31] (ChIP), expression profiling of HNF4α knocked down in HepG2 cells using RNAi (RNAi) and PBM2 search of -2 kb to +1 kb of all annotated human genes (UCSC hg18) (PBM). Shown are the Biological Processes for which at least one of the three methods had a p-value (EASE-score) of < 0.001 (***), < 0.01 (**), or <0.05 (*). Redundant categories were removed. Biological Processes related to classical HNF4α target genes well established in the literature (e.g., Table 6.1). (Figure continued on following page)



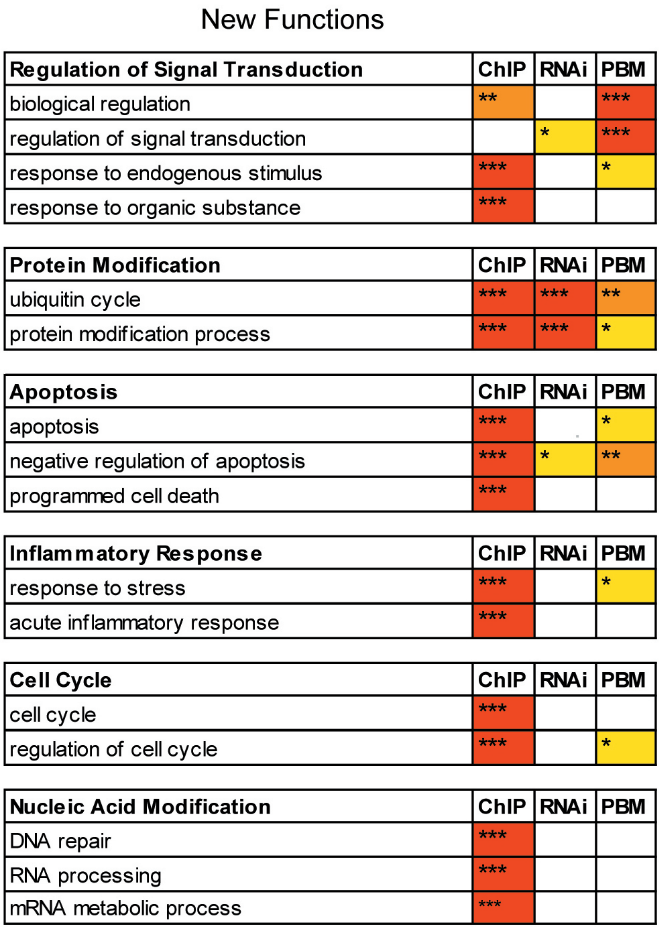
Classical Functions (cont.)

Transport	ChIP	RNAi	PBM
transport	***	***	***
Golgi vesicle transport	***	***	*
vesicle-mediated transport	***	***	**
di-, tri-valent inorganic cation transport			***

Development	ChIP	RNAi	PBM
developmental process			***
anatomical structure development			***
cell differentiation			***
cellular component organization and biogenesis	***	*	***
multicellular organismal development			***
organelle organization and biogenesis	***		***

Figure 2.19. Continued from previous page.

carrier proteins, SLC genes; lipid metabolism - e.g., ABCC2, DGAT2, HSD's), or more recently identified targets of HNF4α (e.g., CREB3L3, NR1I2, NR1H4, DO1) [24, 29, 33, 44]. There were also many genes that, like NINJ1, are in completely new categories of genes not typically associated with HNF4α (e.g., signal transduction, immune response, stress response, apoptosis, cancer related and cell structure) (Fig. 2.22), several of which are reminiscent of the new functional categories identified by GO (Fig. 2.20). In order to determine whether the ChIP signal overlapped with the PBM or SVM sites in these new targets, all three datasets were visualized using Integrated Genome Browser. While not all ChIP signals aligned exactly with the PBM or SVM sites, a very large number did; a sampling of these are shown in Figs. 2.23,2.24.



New Functions

Regulation of Signal Transduction	ChIP	RNAi	PBM
biological regulation	**		***
regulation of signal transduction		*	***
response to endogenous stimulus	***		*
response to organic substance	***		

Protein Modification	ChIP	RNAi	PBM
ubiquitin cycle	***	***	**
protein modification process	***	***	*

Apoptosis	ChIP	RNAi	PBM
apoptosis	***		*
negative regulation of apoptosis	***	*	**
programmed cell death	***		

Inflammatory Response	ChIP	RNAi	PBM
response to stress	***		*
acute inflammatory response	***		

Cell Cycle	ChIP	RNAi	PBM
cell cycle	***		
regulation of cell cycle	***		*

Nucleic Acid Modification	ChIP	RNAi	PBM
DNA repair	***		
RNA processing	***		
mRNA metabolic process	***		

Figure 2.20. Biological Processes not typically associated HNF4α. Analysis performed as in Fig. 2.18. See Supplemental Table 6.10 for a complete list of GO terms and p-values for the ChIP, PBM and RNAi as well as the SVM search (>4 sites in -2 kb to +1 kb).

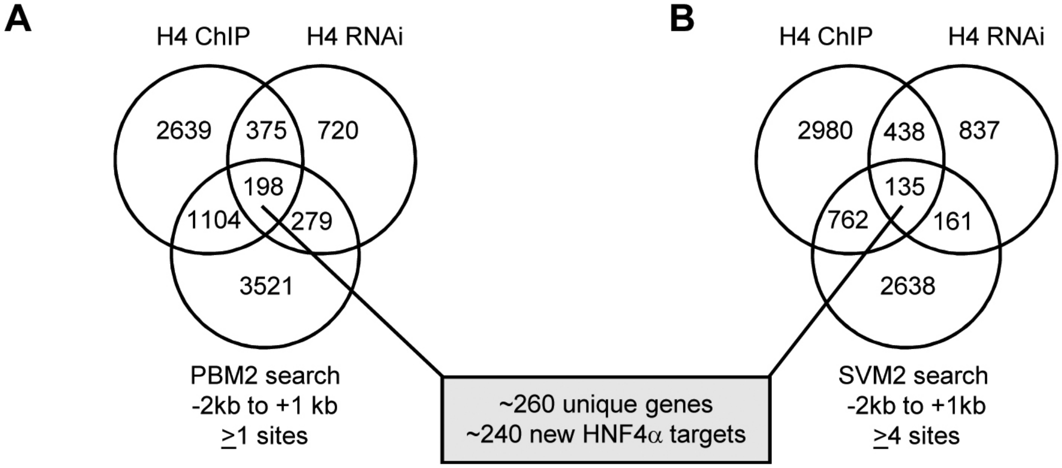
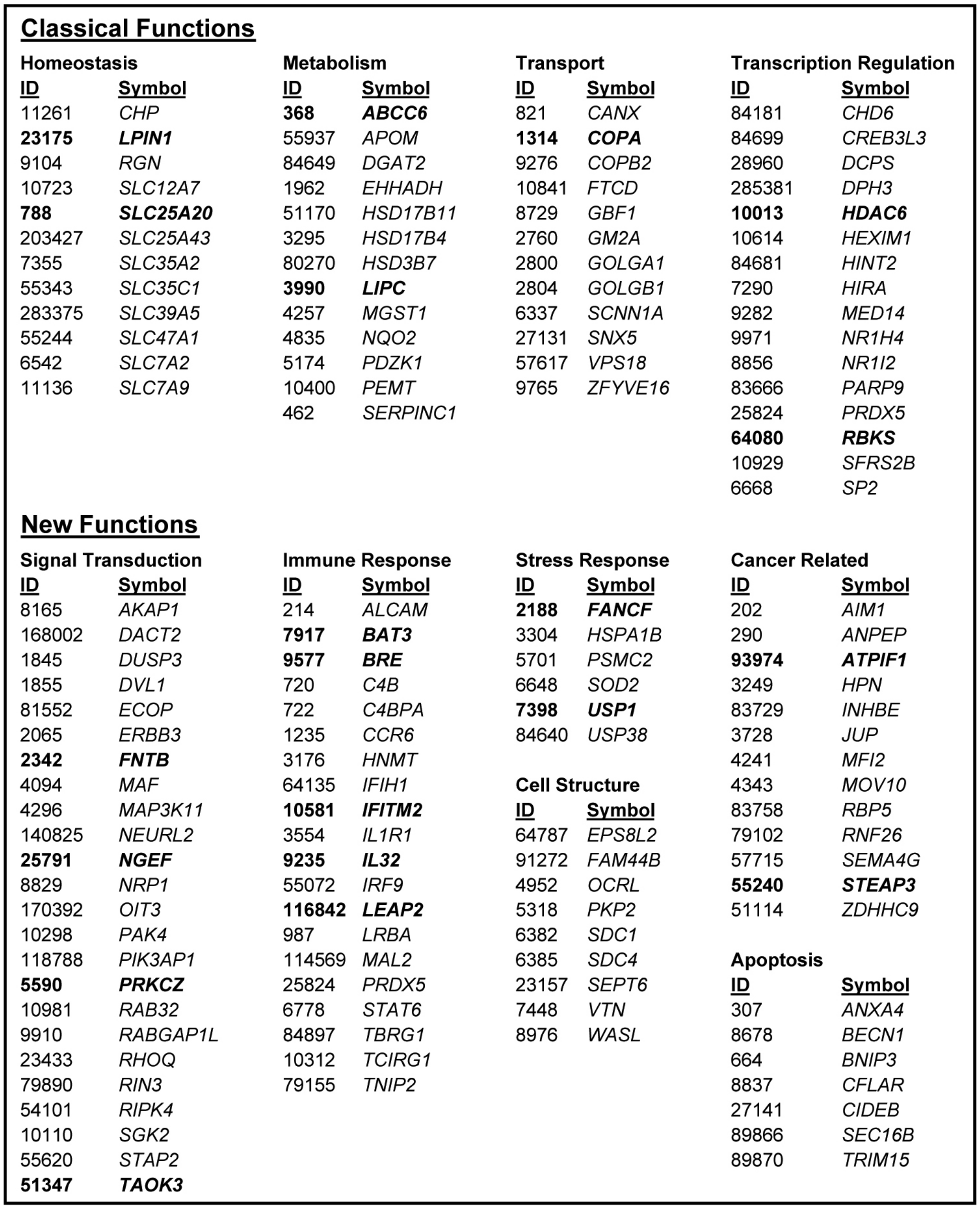


Figure 2.21. Cross reference of three methods used to identify potential human HNF4α target genes - ChIP-chip, RNAi expression profiling and PBM/SVM binding site search. (A) Venn analysis of genes: bound by HNF4α in primary human hepatocytes (H4 ChIP) [31]; down regulated in expression profiling by HNF4α siRNA in HepG2 cells (H4 RNAi) (Fig. 2.14); and containing a potential HNF4α binding site as determined by an exact match search using PBM2 results of annotated human genes (UCSC hg18) -2 kb to +1 kb relative to the TSS (PBM2 search). Shown are the number of genes; genes in the intersection are likely to be direct targets of HNF4α. (B) As in (A) except with SVM2 search of annotated human genes with 4 or more sites. (See Supplemental Tables 6.11 and 6.12 for a complete list of the 198 and 135 genes in the intersection of the Venn diagrams in A and B, respectively.)



Classical Functions

Homeostasis

ID	Symbol
11261	*CHP*
23175	***LPIN1***
9104	*RGN*
10723	*SLC12A7*
788	***SLC25A20***
203427	*SLC25A43*
7355	*SLC35A2*
55343	*SLC35C1*
283375	*SLC39A5*
55244	*SLC47A1*
6542	*SLC7A2*
11136	*SLC7A9*

Metabolism

ID	Symbol
368	***ABCC6***
55937	*APOM*
84649	*DGAT2*
1962	*EHHADH*
51170	*HSD17B11*
3295	*HSD17B4*
80270	*HSD3B7*
3990	***LIPC***
4257	*MGST1*
4835	*NQO2*
5174	*PDZK1*
10400	*PEMT*
462	*SERPINC1*

Transport

ID	Symbol
821	*CANX*
1314	***COPA***
9276	*COPB2*
10841	*FTCD*
8729	*GBF1*
2760	*GM2A*
2800	*GOLGA1*
2804	*GOLGB1*
6337	*SCNN1A*
27131	*SNX5*
57617	*VPS18*
9765	*ZFYVE16*

Transcription Regulation

ID	Symbol
84181	*CHD6*
84699	*CREB3L3*
28960	*DCPS*
285381	*DPH3*
10013	***HDAC6***
10614	*HEXIM1*
84681	*HINT2*
7290	*HIRA*
9282	*MED14*
9971	*NR1H4*
8856	*NR1I2*
83666	*PARP9*
25824	*PRDX5*
64080	***RBKS***
10929	*SFRS2B*
6668	*SP2*

New Functions

Signal Transduction

ID	Symbol
8165	*AKAP1*
168002	*DACT2*
1845	*DUSP3*
1855	*DVL1*
81552	*ECOP*
2065	*ERBB3*
2342	***FNTB***
4094	*MAF*
4296	*MAP3K11*
140825	*NEURL2*
25791	***NGEF***
8829	*NRP1*
170392	*OIT3*
10298	*PAK4*
118788	*PIK3AP1*
5590	***PRKCZ***
10981	*RAB32*
9910	*RABGAP1L*
23433	*RHOQ*
79890	*RIN3*
54101	*RIPK4*
10110	*SGK2*
55620	*STAP2*
51347	***TAOK3***

Immune Response

ID	Symbol
214	*ALCAM*
7917	***BAT3***
9577	***BRE***
720	*C4B*
722	*C4BPA*
1235	*CCR6*
3176	*HNMT*
64135	*IFIH1*
10581	***IFITM2***
3554	*IL1R1*
9235	***IL32***
55072	*IRF9*
116842	***LEAP2***
987	*LRBA*
114569	*MAL2*
25824	*PRDX5*
6778	*STAT6*
84897	*TBRG1*
10312	*TCIRG1*
79155	*TNIP2*

Stress Response

ID	Symbol
2188	***FANCF***
3304	*HSPA1B*
5701	*PSMC2*
6648	*SOD2*
7398	***USP1***
84640	*USP38*

Cell Structure

ID	Symbol
64787	*EPS8L2*
91272	*FAM44B*
4952	*OCRL*
5318	*PKP2*
6382	*SDC1*
6385	*SDC4*
23157	*SEPT6*
7448	*VTN*
8976	*WASL*

Cancer Related

ID	Symbol
202	*AIM1*
290	*ANPEP*
93974	***ATPIF1***
3249	*HPN*
83729	*INHBE*
3728	*JUP*
4241	*MFI2*
4343	*MOV10*
83758	*RBP5*
79102	*RNF26*
57715	*SEMA4G*
55240	***STEAP3***
51114	*ZDHHC9*

Apoptosis

ID	Symbol
307	*ANXA4*
8678	*BECN1*
664	*BNIP3*
8837	*CFLAR*
27141	*CIDEB*
89866	*SEC16B*
89870	*TRIM15*

Figure 2.22. Sampling of new HNF4α target genes that are bound *in vivo*, down regulated in HNF4α knockdown and containing >1 PBM or >4 SVM sites. Functions classically associated with HNF4α are shown as well as new functional categories. ID, Entrez Gene ID; Symbol, Official Gene Symbol. (See Supplemental Tables 6.13 and 6.14 for a complete listing of all human genes with 1 or more PBM sites and 4 or more SVM sites, respectively.)

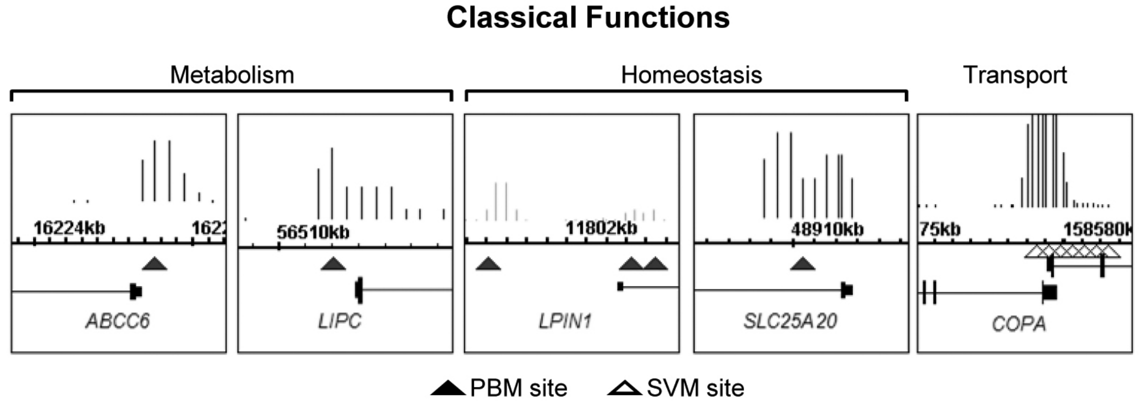


Figure 2.23. Illustration of select new HNF4α target genes down regulated in RNAi, bound in vivo and with PBM or SVM HNF4α binding sites. Screenshots from Integrated Genome Browser of HNF4α ChIP-chip signals from primary human hepatocytes in promoter regions [31] with PBM (closed triangle) sites indicated. SVM sites (open triangle) are indicated only for those genes lacking a PBM site in the region shown. ChIP signals are all statistically significant. Numbers are chromosome coordinates from UCSC hg18. Not all shots are on the same scale. Classical (A) and new functions (B,C,D) as defined in Fig. 7 are indicated.

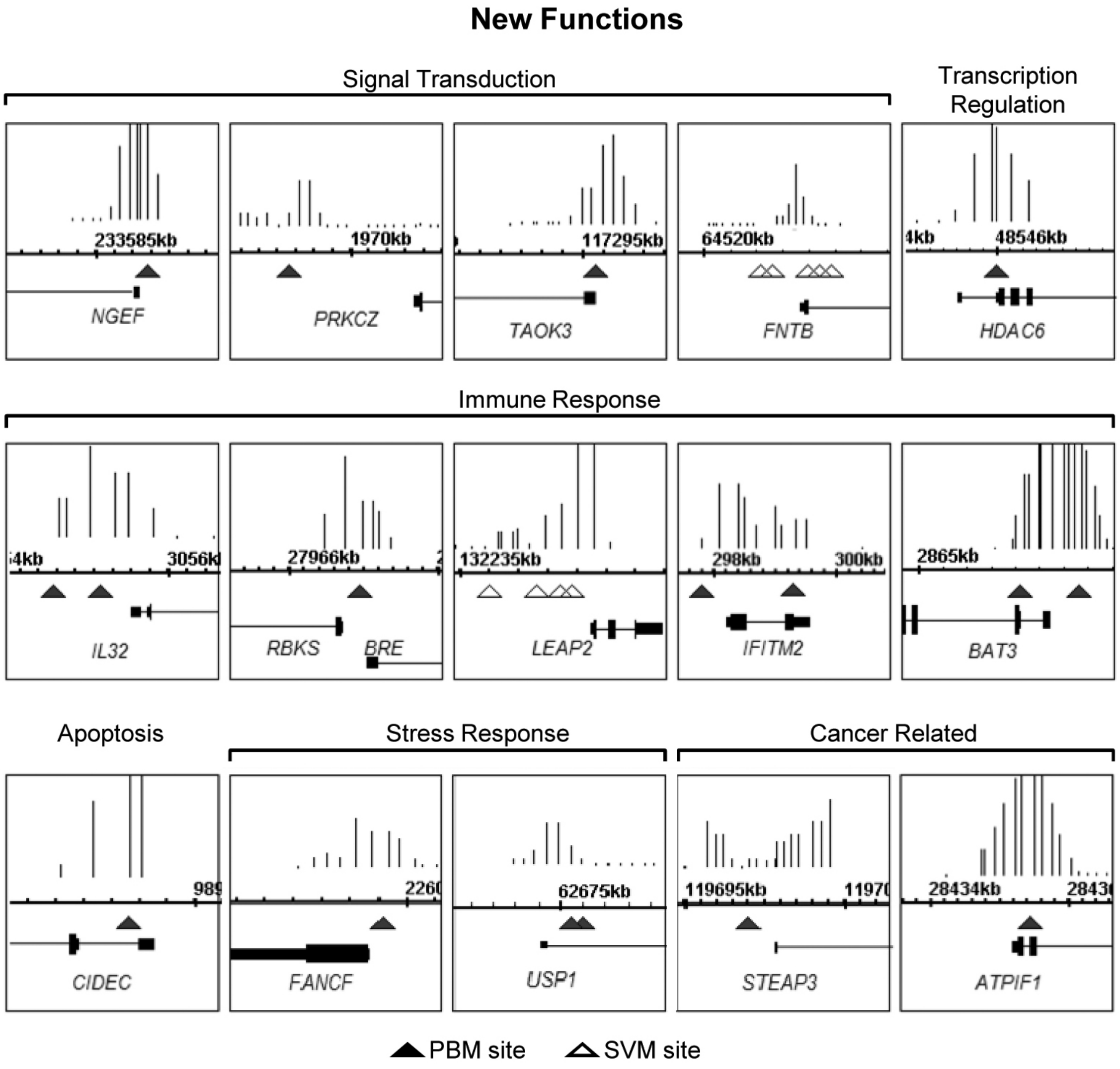


Figure 2.24. Illustration of select new HNF4α target genes down regulated in RNAi, bound in vivo and with PBM or SVM HNF4α binding sites. Screenshots from Integrated Genome Browser of HNF4α ChIP-chip signals from primary human hepatocytes in promoter regions [31] with PBM (closed triangle) sites indicated. SVM sites (open triangle) are indicated only for those genes lacking a PBM site in the region shown. ChIP signals are all statistically significant. Numbers are chromosome coordinates from UCSC hg18. Not all shots are on the same scale. Classical (A) and new functions (B,C,D) as defined in Fig. 7 are indicated.

2.5 Discussion

Identification of TF binding sites and target genes can be a laborious process. Recent genome-scale technologies such as expression profiling and genome-wide location analysis can greatly expand the repertoire of potential targets with relative ease, although the question remains as to which are direct targets that contain bona fide binding sites. Protein binding microarrays (PBMs) allow for a high throughput identification of DNA binding sequences that can then be integrated with the other techniques, and can also be used to predict potential new targets in additional tissues or developmental stages. Here, we successfully adapt the PBM technology to assess HNF4α DNA binding under conditions that more closely approximate physiological conditions (i.e., native full length receptor in a crude nuclear extract) (Fig. 2.2). We show that the PBM results are highly reproducible across different species (human and rat) and isoforms ($\alpha 2$ and $\alpha 8$) of HNF4α under a variety of conditions (Fig.s 2.7 , 2.8). We identify new rules for DNA binding and develop an SVM model to predict additional sites (Fig.s 2.10,2.13A). We compare the PBM and SVM results to RNAi expression profiling (Fig. 2.14) as well as to published ChIP-chip results in order to develop an integrated approach for the identification of human HNF4α target genes. We show that all three systems yield similar overrepresented categories of target genes (Fig.s 2.18, 2.19, 2.20), supporting the notion that specific TF binding sites in promoter regions are a major factor in driving gene expression. Using this integrated approach we identify ~240 new, direct targets of HNF4α, many of which are in new functional categories (Fig.s 2.21,2.22,2.23,2.24). To our knowledge, this is the first such integration of extensive PBM, CHIP-chip and expression profiling data for any transcription factor. Finally, to facilitate future HNF4α target gene research, we have developed a publicly available web-based tool (HNF4 Motif Finder) based on our PBM results that can be used to search any DNA sequence for potential HNF4α binding sites (http://nrmotif.ucr.edu). We define direct targets as genes that meet three criteria - contain a functional binding site in a regulatory region (PBM/SVM search), bind in vivo to the promoter (ChIP) and are down regulated when HNF4α expression is knocked down (RNAi). Applying these criteria, we expand upon the classical roles of HNF4α by identifying additional target genes involved in metabolism (e.g., *APOM*, *LIPC*, *LPIN1*), solute carrier transport (e.g., *SLC7A2*, *SLC12A7*, *SLC25A20*), protein transport and secretion (e.g., *COPA*, *GOLGB1*, *GOLGA1*), as well as transcription regulation (e.g., HDAC6, MED14, etc.). The integrated approach also identified new HNF4α targets in pathways not previously associated with HNF4α, such as regulation of signal transduction (e.g. *TAOK3*, *NGEF*, *PRKCZ*, *FNTB*), and inflammation and immune response (e.g. *IL32*, *BRE*, *LEAP2*, *IFITM2*, *BAT3*). Perhaps the most intriguing new categories of HNF4α target genes are those involved in apoptosis, DNA repair and cancer. HNF4α has long been considered a key factor in hepatocyte differentiation [18, 40] but there are an increasing number of reports indicating that HNF4α may act as a tumor suppressor [38, 42]. This view is supported by the new target genes identified here, such as *NINJ1* (Fig. 2.15), which may play a role in regulating cellular senescence by inducing the expression of p21, a cell cycle inhibitor gene [39], and is consistent with our previous findings that the p21 gene (CDKN1A) itself is a direct target of HNF4α [19]. Other new HNF4α target genes related to anti-growth effects are: *CIDEC*, which induces fragmentation of DNA upon apoptosis; *ATPIF1*, which inhibits an ATPase involved in angiogenesis; and *STEAP3*, which is induced by tumor suppressor p53 and whose down regulation is associated with a transition from cirrhosis to hepatocellular carcinoma [8]. There were also genes involved in stress responses such as the DNA repair gene *FANCF*, a Fanconi's anemia complementation group F, and *USP1*, a ubiquitin specific protease. In addition to the genes that meet the three criteria mentioned above, our analysis also revealed thousands of additional genes that met only one or two of the three criteria. While technical considerations (e.g., missing tiles in the ChIP-chip, malfunctioning probes in the expression arrays, false positives in the ChIP assay, etc.) are sure to account for some of those genes, other explanations are also possible. For example, the genes present only in the expression profiling could be indirect targets of HNF4α and hence yield no PBM/SVM or ChIP signal. Genes present in ChIP-chip alone could contain as yet unidentified HNF4α binding sites or recruit HNF4α in a nondirect fashion; it should also be noted that in Fig. 2.22 we imposed a fairly stringent requirement of four or more SVM sites for a gene to be included in that analysis. Genes identified only in the PBM/SVM searches could contain bona fide HNF4α binding sites but are simply not expressed in the hepatocellular carcinoma cell line (HepG2) used in the expression profiling nor in the particular set of primary human hepatocytes used in the ChIP-chip. It could also be that in adult hepatocytes the promoter regions of those genes are not available for binding (and hence activation) due to the structure of the chromatin. Genes found only in the PBM/SVM searches could also represent non hepatic targets that are expressed in other HNF4α-expressing tissues such as kidney, pancreas, intestine and colon. Finally, it is also possible that there may be potential HNF4α binding sites in the human genome that are never used by HNF4α. Whatever the reasons for the incomplete overlap between the three assays, the use of the PBM/SVM results presented here, as well as the web-based HNF4 Motif Finder, should greatly facilitate any future investigation of potential HNF4α target genes. Additionally, our approach of integrating data from multiple genome-wide assays, including PBMs, provides a powerful new framework for identifying direct targets of TFs.

2.6 Acknowledgements

This work was funded by grants to FMS (NIH DK053892), TJ (NSF IIS-0711129), FMS and TJ (UCR IIGB, NIH R21MH087397), EB (PhRMA Foundation pre-doctoral fellowship) and WH-V (UC Toxic Substance Training Grant). We would also like to thank the following for help: A. Karatzoglou (ksvm), S. Davis (ACME) and J. Schnabl (Table 6.1).

2.7 Figures

2.8 Supplemental Figures

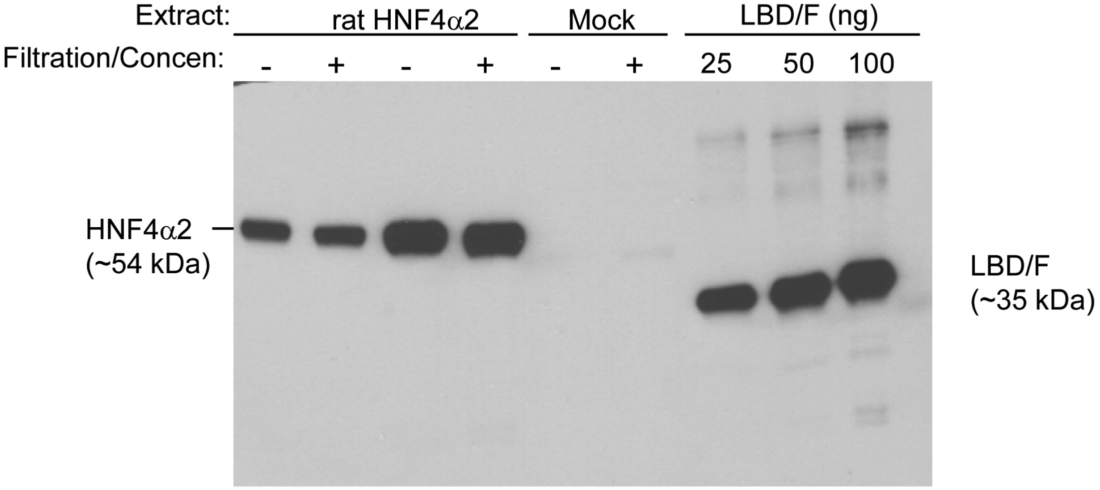


Figure 2.25. Specificity of HNF4α antibody used in PBMs. Immunoblot (IB) of crude nuclear extracts (~ 5μg per lane) from COS-7 cells transfected with pMT7.rHNF4α2 expressing full length rat HNF4α2 or mock transfected probed with the affinity purified C-terminal antibody (α445). Mock transfected COS-7 cells do not show a detectable level of HNF4α or any other cross reacting bands. Known quantities (25, 50, and 100 ng) of purified, recombinant LBD/F (ligand binding/F domain) allow for approximate quantification of HNF4α. Extracts applied to the PBMs were filtered and concentrated (see main text for details). IB analysis of the other antibodies (commercial mouse monoclonals) used to develop the PBMs (αCTD and αNTD) gave equally excellent, specific signals (blots not shown).

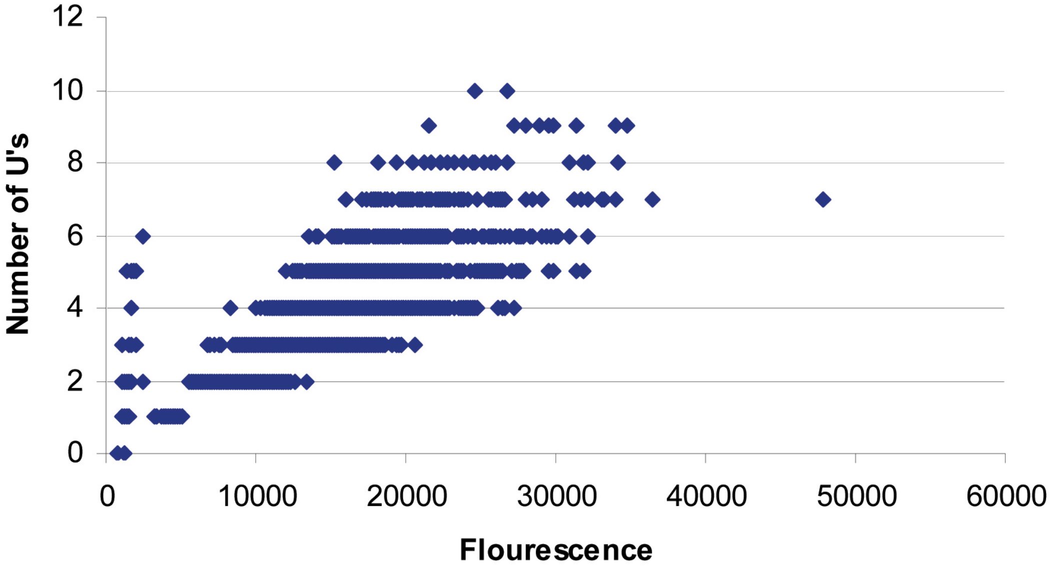


Figure 2.26. Linear relationship between Cy3 incorporation and number of As in the variable region. Scatter plot of Cy3-mediated fluorescence (arbitrary units) in DNA extended on PBM1 in the presence of dUTP-Cy3 and the number of uracils incorporated, based on the number of adenines in the variable region (see Fig. 1 for oligo design and Supplemental Table 6.3 for a complete list of sequences on PBM1).

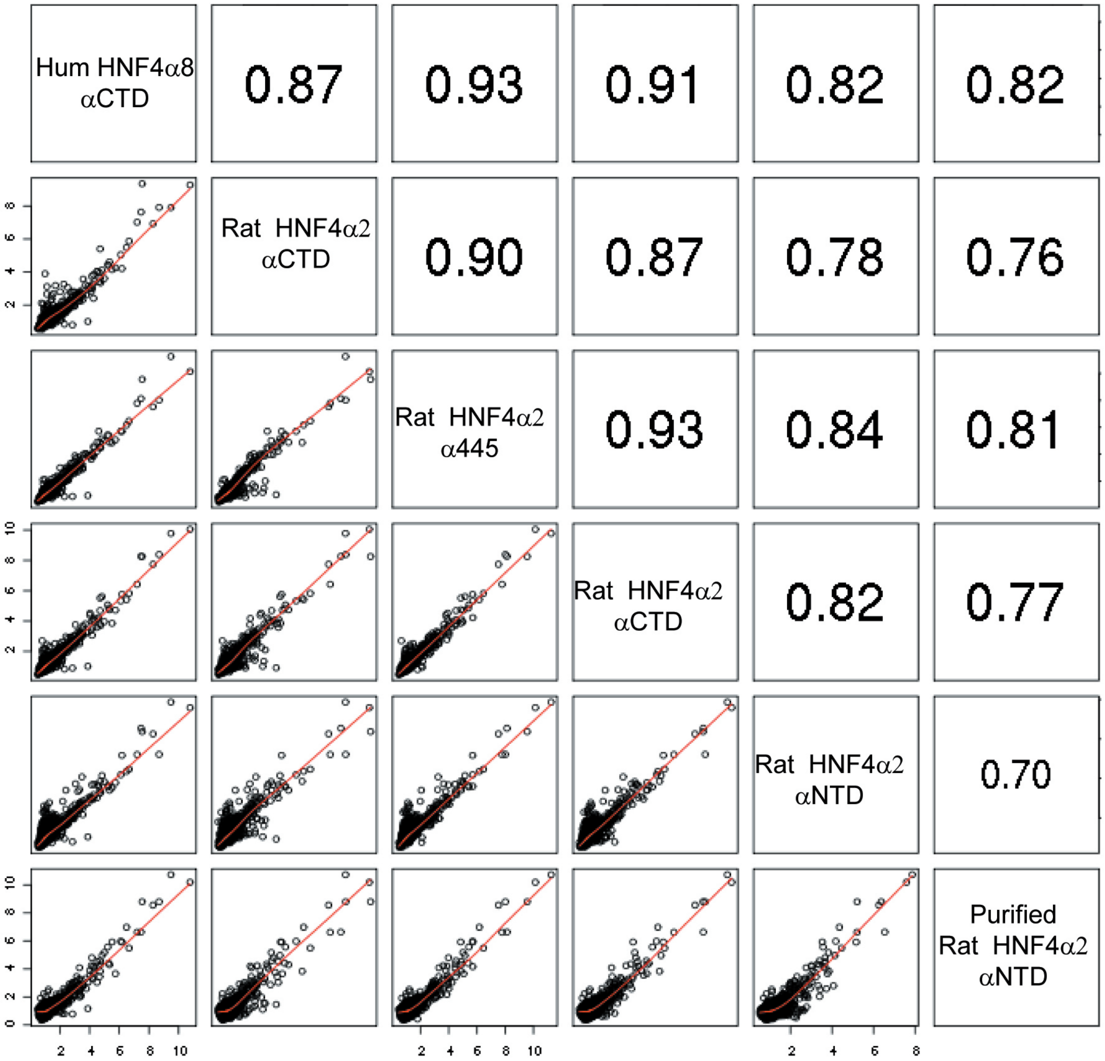


Figure 2.27. Reproducibility of PBM1. Scatter plot matrix of normalized fluorescence illustrating reproducibility of PBM1 across different HNF4α isoforms, species and antibodies. The plot shows that the intensities are highly reproducible and there is no significant difference between isoforms (HNF4α2 vs. HNF4α8), species (rat vs human) or antibodies (α445 vs. αCTD vs. αNTD) used in the arrays. All protein samples are from crude nuclear extracts from transfected COS-7 cells except for the purified material (see main text for details). Numbers, correlation coefficients, R squared.

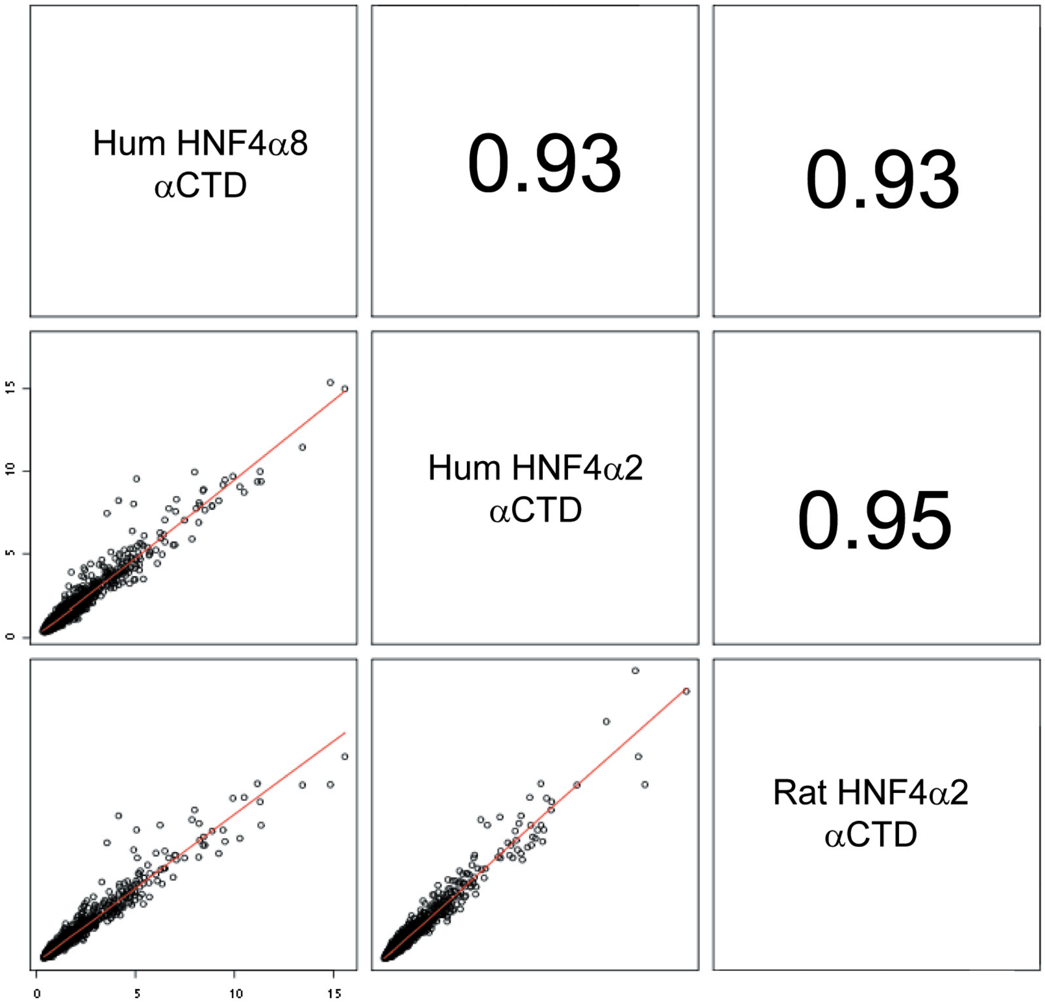


Figure 2.28. Reproducibility of PBM2. Scatter plot matrix of normalized fluorescence illustrating reproducibility of PBM1 across different HNF4α isoform and species. The plot shows that the intensities are highly reproducible and there is no significant difference between isoforms (HNF4α2 vs. HNF4α8) or species (rat vs. human) or between the arrays. Numbers, correlation coefficients, R squared. Increased correlation in PBM2 vs. PBM1 (Fig. 2.27) is attributed to an improvement in array treatment conditions.

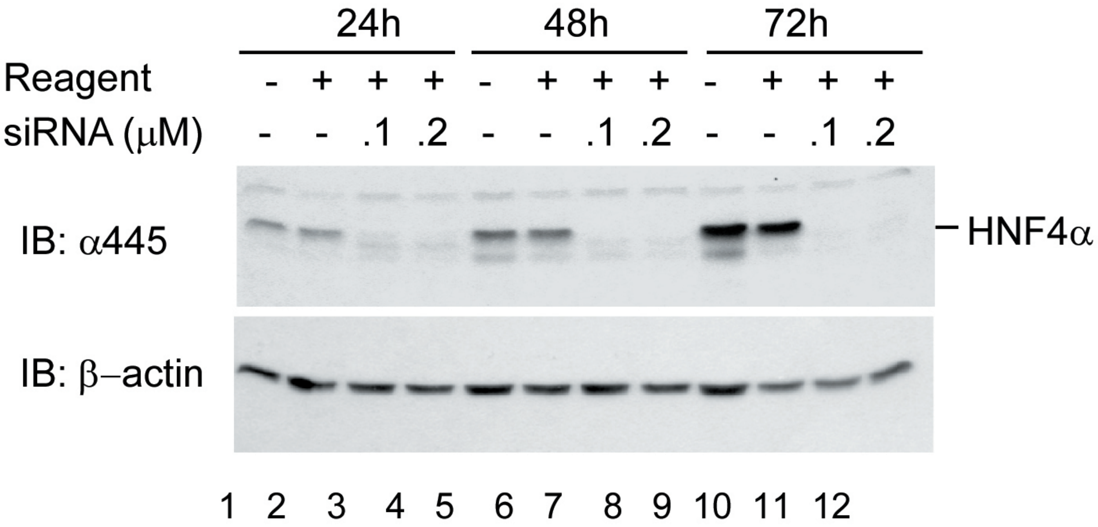

24h
48h
72h
Reagent - + + + - + + + - + + +
siRNA (μM) - - .1 .2 - - .1 .2 - - .1 .2
IB: α445
HNF4α
IB: β–actin
1 2 3 4 5 6 7 8 9 10 11 12

Figure 2.29. RNAα Knockdown of HNF4α in HepG2 cells. Immunoblot (IB) analysis with affinity purified antibody α445 to HNF4α showing a decrease in human HNF4α protein upon treatment of HepG2 cells with siRNAs directed against HNF4α1 as described in Materials and Methods in the main text. Reagent, transfection reagent. Cells were harvested at the indicated times after siRNAs were introduced. 20 μg total protein of whole cell extracts were loaded per lane.

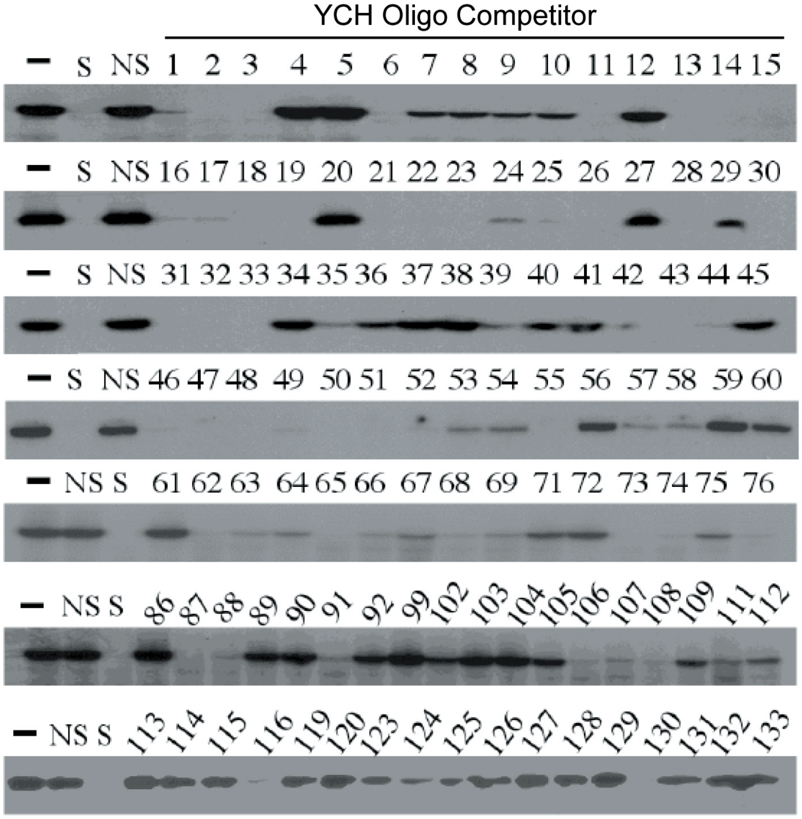
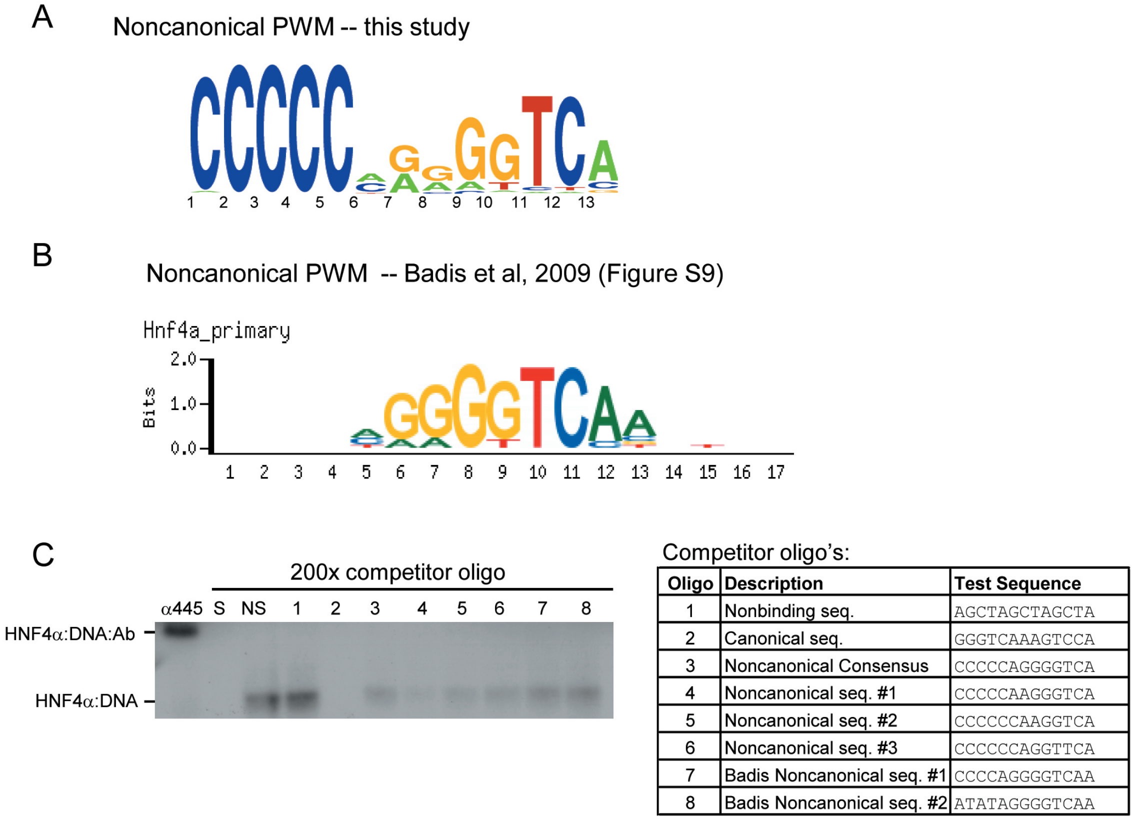

YCH Oligo Competitor
− S NS 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
− S NS 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
− S NS 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
− S NS 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
− NS S 61 62 63 64 65 66 67 68 69 71 72 73 74 75 76
− NS S 86 87 88 89 90 91 92 99 102 103 104 105 106 107 108 109 111 112
− NS S 113 114 115 116 119 120 123 124 125 126 127 128 129 130 131 132 133

Figure 2.30. Gel shift results used in Fig. 2.11 to compare to PBM results Gel shifts were performed as described in Materials and Methods using crude nuclear extracts from COS-7 cells transfected with pMT7.HNF4α1 (rat) and the ApoB.-85.-47 double-stranded oligonucleotide as a 32P-labelled probe. Unlabeled YCH oligonucleotide competitors (YCH1-133) as well as a specific competitor (S, ApoB.-85.-47) and a nonspecific competitor (NS, 175 TTR) were added to the shift reactions in 100-fold molar excess. Shown are the HNF4α:DNA complexes: binders are represented by a lack of a shift band; nonbinders contain a shift band analogous to the control lacking a competitor oligonucleotide (-). Not shown is the free probe which is in excess in all reactions. See Supplemental Table 6.15 for sequence of ApoB.-85.-47, 175TTR and the YCH oligos. Results from these competitions were compared to PBM2 results in Fig. 2.11 in the main text.



Competitor oligo's:

Oligo	Description	Test Sequence
1	Nonbinding seq.	AGCTAGCTAGCTA
2	Canonical seq.	GGGTCAAAGTCCA
3	Noncanonical Consensus	CCCCCAGGGGTCA
4	Noncanonical seq. #1	CCCCCAAGGGTCA
5	Noncanonical seq. #2	CCCCCCAAGGTCA
6	Noncanonical seq. #3	CCCCCCAGGTTCA
7	Badis Noncanonical seq. #1	CCCCAGGGGTCAA
8	Badis Noncanonical seq. #2	ATATAGGGGTCAA

Figure 2.31.
Binding of HNF4α to noncanonical sequences. (A,B) Position weight matrices (PWMs) of noncanonical binding sequences from PBM results from this paper (A) and Badis et al. 2009 (B). The PBMs used by Badis et al. resolve only up to 8-mers while the PBMs used in this study can resolve 13-mers. (C) Gel shift assay performed as described in Materials and Methods using crude nuclear extracts from COS-7 cells transfected with pMT7.HNF4α2 (human) and the ApoB.-85.-47 double-stranded oligonucleotide as a 32P-labelled probe. Unlabeled oligonucleotide competitors were added in 200-fold molar excess. Left, gel with HNF4α:DNA and supershift (HNF4α:DNA:Ab) complexes indicated. S, specific competitor (ApoB.-85.-47); NS, nonspecific competitor 175 TTR. α445, affinity purified antibody to HNF4α. Not shown is the free probe which is in excess in all reactions. Right, legend with descriptors and test sequence for Oligos 1-8 used in the competitions. See Supplemental Table 6.15 for complete sequences of all oligos.
Results: Gel shift analysis reveals that HNF4α2 binds noncanonical sequences identified by PBM although the binding is not as strong as to canonical sequences (e.g., oligo 2).

Bibliography

[1] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*, 2009.

[2] M. F. Berger, G. Badis, A. R. Gehrke, S. Talukder, A. A. Philippakis, L. Pena-Castillo, T. M. Alleyne, S. Mnaimneh, O. B. Botvinnik, E. T. Chan, F. Khalid, W. Zhang, D. Newburger, S. A. Jaeger, Q. D. Morris, M. L. Bulyk, and T. R. Hughes. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133:1266–76, 2008.

[3] M. F. Berger and M. L. Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc*, 4:393–411, 2009.

[4] A. A. Bogan, Q. Dallas-Yang, J. Ruse, M. D., Y. Maeda, G. Jiang, L. Nepomuceno, T. S. Scanlan, F. E. Cohen, and F. M. Sladek. Analysis of protein dimerization and ligand binding of orphan receptor HNF4alpha. *J Mol Biol*, 302:831–51, 2000.

[5] E. Bolotin, J. Schnabl, and F. Sladek. HNF4A (Homo sapiens). *Transcription Factor Encyclopedia http://www.cisreg.ca/tfe*, 2009, 2008.

[6] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–93, 2003.

[7] M. L. Bulyk. Analysis of sequence specificities of DNA-binding proteins with protein binding microarrays. *Methods Enzymol*, 410:279–299, 2006.

[8] F. Caillot, R. Daveau, M. Daveau, J. Lubrano, G. Saint-Auret, M. Hiron, O. Goria, M. Scotte, A. Francois, and J. P. Salier. Down-regulated expression of the TSAP6 protein in liver is associated with a transition from cirrhosis to hepatocellular carcinoma. *Histopathology*, 54:319–27, 2009.

[9] F. L. Chartier, J. P. Bossu, V. Laudet, J. C. Fruchart, and B. Laine. Cloning and sequencing of cDNAs encoding the human hepatocyte nuclear factor 4 indicate the presence of two isoforms in human liver. *Gene*, 147:269–72, 1994.

[10] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res*, 14:1188–90, 2004.

[11] J. Dennis, G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4:P3, 2003.

[12] J. Eeckhoute, P. Formstecher, and B. Laine. Maturity-onset diabetes of the young Type 1 (MODY1)-associated mutations R154X and E276Q in hepatocyte nuclear factor 4alpha (HNF4alpha) gene impair recruitment of p300, a key transcriptional coactivator. *Mol Endocrinol*, 15:1200–10, 2001.

[13] S. Ellard and K. Colclough. Mutations in the genes encoding the transcription factors hepatocyte nuclear factor 1 alpha (HNF1A) and 4 alpha (HNF4A) in maturity-onset diabetes of the young. *Hum Mutat*, 27:854–69, 2006.

[14] K. Ellrott, C. Yang, F. M. Sladek, and T. Jiang. Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics*, 18 Suppl 2:S100–9, 2002.

[15] R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors. *Limma: linear models for microarray data*. Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer, New York,, 2005.

[16] F. J. Gonzalez. Regulation of hepatocyte nuclear factor 4 alpha-mediated transcription. *Drug Metab Pharmacokinet*, 23:2–7, 2008.

[17] R. K. Gupta and K. H. Kaestner. HNF-4alpha: from MODY to late-onset type 2 diabetes. *Trends Mol Med*, 10:521–4, 2004.

[18] G. P. Hayhurst, Y. H. Lee, G. Lambert, J. M. Ward, and F. J. Gonzalez. Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. *Mol Cell Biol*, 21:1393–403, 2001.

[19] W. W. Hwang-Verslues and F. M. Sladek. Nuclear receptor hepatocyte nuclear factor 4alpha1 competes with oncoprotein c-Myc for control of the p21/WAF1 promoter. *Mol Endocrinol*, 22:78–90, 2008.

[20] G. Jiang, U. Lee, and F. M. Sladek. Proposed mechanism for the stabilization of nuclear receptor DNA binding via protein dimerization. *Mol Cell Biol*, 17:6546–54, 1997.

[21] G. Jiang, L. Nepomuceno, K. Hopkins, and F. M. Sladek. Exclusive homodimerization of the orphan receptor hepatocyte nuclear factor 4 defines a new subclass of nuclear receptors. *Mol Cell Biol*, 15:5131–43, 1995.

[22] G. Jiang, L. Nepomuceno, Q. Yang, and F. M. Sladek. Serine/threonine phosphorylation of orphan receptor hepatocyte nuclear factor 4. *Arch Biochem Biophys*, 340:1–9, 1997.

[23] G. Jiang and F. M. Sladek. The DNA binding domain of hepatocyte nuclear factor 4 mediates cooperative, specific binding to DNA and heterodimerization with the retinoid X receptor alpha. *J Biol Chem*, 272:1218–25, 1997.

[24] A. Kamiya, Y. Inoue, and F. J. Gonzalez. Role of the hepatocyte nuclear factor 4alpha in control of the pregnane X receptor during fetal liver development. *Hepatology*, 37:1375–84, 2003.

[25] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab:An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11, 2004.

[26] F. Lemaigre and K. S. Zaret. Liver development update: new embryo models, cell lineage control, and morphogenesis. *Curr Opin Genet Dev*, 14:582–90, 2004.

[27] J. Linnell, R. Mott, S. Field, D. P. Kwiatkowski, J. Ragoussis, and I. A. Udalova. Quantitative high-throughput analysis of transcription factor binding specificities. *Nucleic Acids Res*, 32:e44, 2004.

[28] P. Lu, G. B. Rha, M. Melikishvili, G. Wu, B. C. Adkins, M. G. Fried, and Y. I. Chi. Structural basis of natural promoter recognition by a unique nuclear receptor, HNF4alpha. Diabetes gene product. *J Biol Chem*, 283:33685–97, 2008.

[29] J. Luebke-Wheeler, K. Zhang, M. Battle, K. Si-Tayeb, W. Garrison, S. Chhinder, J. Li, R. J. Kaufman, and S. A. Duncan. Hepatocyte nuclear factor 4alpha is implicated in endoplasmic reticulum stress-induced acute phase response by regulating expression of cyclic adenosine monophosphate responsive element binding protein H. *Hepatology*, 48:1242–50, 2008.

[30] P. Neuvial, P. Hupe, I. Brito, S. Liva, E. Manie, C. Brennetot, F. Radvanyi, A. Aurias, and E. Barillot. Spatial normalization of array-CGH data. *BMC Bioinformatics*, 7:264, 2006.

[31] D. T. Odom, R. D. Dowell, E. S. Jacobsen, L. Nekludova, P. A. Rolfe, T. W. Danford, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. Core transcriptional regulatory circuitry in human hepatocytes. *Mol Syst Biol*, 2:2006 0017, 2006.

[32] D. T. Odom, N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303:1378–81, 2004.

[33] H. Ohguchi, T. Tanaka, A. Uchida, K. Magoori, H. Kudo, I. Kim, K. Daigo, I. Sakakibara, M. Okamura, H. Harigae, T. Sasaki, T. F. Osborne, F. J. Gonzalez, T. Hamakubo, T. Kodama, and J. Sakai. Hepatocyte nuclear factor 4alpha contributes to thyroid hormone homeostasis by cooperatively regulating the type 1 iodothyronine deiodinase gene with GATA4 and Kruppel-like transcription factor 9. *Mol Cell Biol*, 28:3917–31, 2008.

[34] A. Rada-Iglesias, O. Wallerman, C. Koch, A. Ameur, S. Enroth, G. Clelland, K. Wester, S. Wilcox, O. M. Dovey, P. D. Ellis, V. L. Wraight, K. James, R. Andrews, C. Langford, P. Dhami, N. Carter, D. Vetrie, F. Ponten, J. Komorowski, I. Dunham, and C. Wadelius. Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Hum Mol Genet*, 14:3435–47, 2005.

[35] M. J. Reijnen, F. M. Sladek, R. M. Bertina, and P. H. Reitsma. Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. *Proc Natl Acad Sci U S A*, 89:6300–3, 1992.

[36] F. Sladek and S. Seidel. Hepatocyte nuclear factor 4alpha. In T. Burris and E. McCabe, editors, *Nuclear Receptors and Genetic Diseases*, pages 309–361. Academic Press, London, 2001.

[37] F. M. Sladek, W. M. Zhong, E. Lai, and J. Darnell, J. E. Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. *Genes Dev*, 4:2353–65, 1990.

[38] T. Tanaka, S. Jiang, H. Hotta, K. Takano, H. Iwanari, K. Sumi, K. Daigo, R. Ohashi, M. Sugai, C. Ikegame, H. Umezu, Y. Hirayama, Y. Midorikawa, Y. Hippo, A. Watanabe, Y. Uchiyama, G. Hasegawa, P. Reid, H. Aburatani, T. Hamakubo, J. Sakai, M. Naito, and T. Kodama. Dysregulated expression of P1 and P2 promoter-driven

hepatocyte nuclear factor-4alpha in the pathogenesis of human cancer. *J Pathol*, 208:662–72, 2006.

[39] T. Toyama, Y. Sasaki, M. Horimoto, K. Iyoda, T. Yakushijin, K. Ohkawa, T. Takehara, A. Kasahara, T. Araki, M. Hori, and N. Hayashi. Ninjurin1 increases p21 expression and induces cellular senescence in human hepatoma cells. *J Hepatol*, 41:637–43, 2004.

[40] A. J. Watt, W. D. Garrison, and S. A. Duncan. HNF4: a central regulator of hepatocyte differentiation and function. *Hepatology*, 37:1249–53, 2003.

[41] C. Yang, E. Bolotin, T. Jiang, F. M. Sladek, and E. Martinez. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389:52–65, 2007.

[42] C. Yin, Y. Lin, X. Zhang, Y. X. Chen, X. Zeng, H. Y. Yue, J. L. Hou, X. Deng, J. P. Zhang, Z. G. Han, and W. F. Xie. Differentiation therapy of hepatocellular carcinoma in mice with recombinant adenovirus carrying hepatocyte nuclear factor-4alpha gene. *Hepatology*, 48:1528–39, 2008.

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

[44] Y. Zhang, F. Y. Lee, G. Barrera, H. Lee, C. Vales, F. J. Gonzalez, T. M. Willson, and P. A. Edwards. Activation of the nuclear receptor FXR improves hyperglycemia and hyperlipidemia in diabetic mice. *Proc Natl Acad Sci U S A*, 103:1006–11, 2006.

When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge of it is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced it to the stage of science.

Sir William Thompson, Lord Kelvin

Chapter 3

HNF4α Transcription Factor Binding Sequences are Widespread in Alu Repeats

3.1 Introduction

3.1.1 Alu Repeats

As much as 50% of the human genome is considered to be derived from repetitive DNA sequence[12]. This large percentile alone suggests that the repeats might have a significant impact on genome function and are worth investigating. However, for a long time repetitive DNA was dismissed as “junk.” Recently a new understanding of some of the repeat families has shown that repetitive DNA is important in evolution of the human genome, as well as gene regulation.

There are many types of repeats, but they can be grouped into two general categories, that come from different origins. Tandem repeats, a repeat of 2 or more nucleotides immediately adjacent to each other, usually result from DNA polymerase slippage during replication. The other category is interspersed repeats, sequences of $>$ 100 bp that are spread throughout the genome. These sequences typically originate from transposons, or "jumping genes" [3]. By far the largest sub category of repeats belongs to interspersed repeats, copies of inserted retroviruses and long terminal repeat (LTR) retrotransposons. Out of those, the most important category is long interspersed nuclear elements (LINEs) retrotransposons which constitute 20% of human genome. These retrotransposons encode all proteins nesessary for their moblization and thus are deemed "autonomous." Short interspersed nuclear elements (SINEs) do not encode reverse transcripase and thus are subsequently termed "non autonomous." They make up an additional ~10% of the human genome. SINEs are thought to be originated from LINEs, with missing or mutated reverse transcriptatase, and thus depend on the expression of reverse transcriptase from the LINEs to propagate.

First identified as ~300 nucleotide repetitive sequence in the 1970s and then characterized by the presence of an *Alu*I restriction enzyme site, from bacterium *Arthrobacter luteus*, with site (5'-AGCT-3'). The importance of Alu elements was hypothesized early on due to their large numbers in the human genome. [1, 6, 7, 19].

Alu elements constitute the majority of the SINEs in the human genome and are estimated be present in >1.2 million copies. They are still mobile in the human genome, but to move in the genome they require a L1 reverse transcriptase from LINE family [3]. Additionally, they are a relatively recent occurrence, being ~65 million years old, and have so far been found exclusively in primates, including humans [14]. The structure of the Alu element is shown in Fig. 3.2. The Alu insertions have been implicated with several diseases such as leukemia, hemophilia and breast cancer, and so their impact on human health is thought to be significant [5].

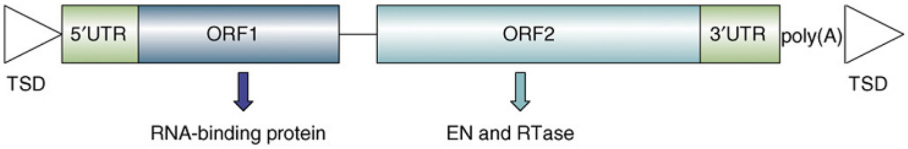


Figure 3.1. LINE (L1) element structure. L1 element is approximately ~6kb in length, containing two open reading frames ORF1 and ORF2. ORF1 contains nucleic acid chaperone activity, suspected to be involved in DNA melting in preparation for reverse transcription. ORF2 encodes endonuclease and reverse transcripatase essential for L1 movement [15, 16]. TSD is target site duplication sequence. Figure adapted from [16].

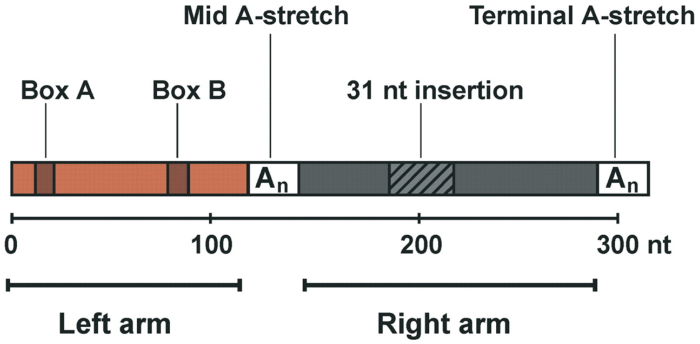


Figure 3.2. The structure of an Alu element. Alu sequence is ~300nt long and composed out of two related, but non identical monomers, the right and left arms. Box A and B are RNA pol III internal promoters, which are functional, but too weak to drive transcription. The right arm differs from the left arm by a 31 nucleotide insertion. Fig. adapted from of [8].

A variety of functions have been hypothesized for Alu repeats. Curiously, one of the first such hypothesis was by Davidson and Britton, in 1973 when they proposed that ~300bp repeats, including Alu elements, are actually transcription factor binding sites (TFBS). *"In summary, it can now be said that there is strong evidence for the existence of sensors in the genome. In addition, we regard it as reasonably likely that the interspersed repetitive and non repetitive sequence represent alternating receptors and structural genes. ... Our approach to gene regulation implies that the location of repetitive sequences provides the hereditary physical basis for the pattern of gene regulation* [4]." Since then a variety of TFBS have been characterized in Alu elements, including YY1 [9], Sp1 [18], as well as two nuclear receptors retinoid acid receptor (RAR) NR1B1 [13], and estrogen receptor (ESR1) NR3A1 [17]. Additionally, a recent study found numerous potential PWM in Alu elements for six classes of TFs such as zinc finger, homeo domain, and TATA binding proteins, from a search using the TRANSFAC database [20]. Finally, there have been reports in which Alu insertions changed expression of a target genes for at least six human genes: *CD8A*, keratin 18 *KRT18*, parathyroid hormone *PTH*, Wilms tumor 1 *WT1*, gamma chain of Fc receptor gene *FCER1G* and *BRCA-1* [2] (Fig. 3.3).

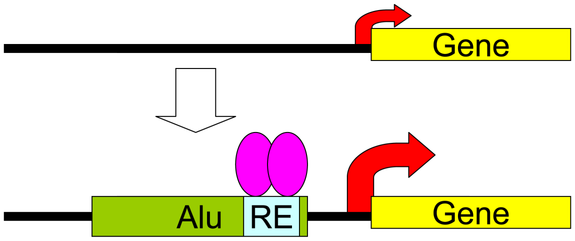


Figure 3.3. Diagram of an Alu insertion affecting gene expression. Shown is a hypothetical TF (purple oval) that binds to its cognate response element (RE) in an Alu sequence inserted into the promoter region of a gene.

During the examiation of HNF4α TFBS, it was noted that certain binding sites were extremely frequent in the human genome but not in the mouse genome. We hypothesized that these sequences might be in Alu repeats and designed a series of experiment to show that there are indeed HNF4α binding sequences in subset of Alu repeats. This leads to a hypothesis that Alu insertion influence gene regulation by HNF4α

3.2 Results

3.2.1 Bioinformatic Analysis

Frequency Profile

Frequency profile of the known HNF4α TFBS in the human genome with 217 sites 6.1, shows that motifs H4.141 (5'-AGGCTGaAGTGCA-3') and H4.109 (5'-AGGCTAaAGTGCA-3') were significantly (~>100x and ~>10x respectively) overrepresented compared to other motifs (Fig. 3.4). When the search was repeated in the mouse, these motifs were not overrepresented compared to other motifs. We hypothesized that this motif is related to repetitive elements, specifically Alu elements that are present in the human, but not in the mouse genome.

H4.141 is Found in Alu Sequences

To confirm Alu elements do indeed contain HNF4α TFBS, the RepBase database of all human repetitive elements was searched using an exact match search with H4.141 and confirmed that H4.141 matched several consensus Alu sequences, in particular M38064_HSAL002939

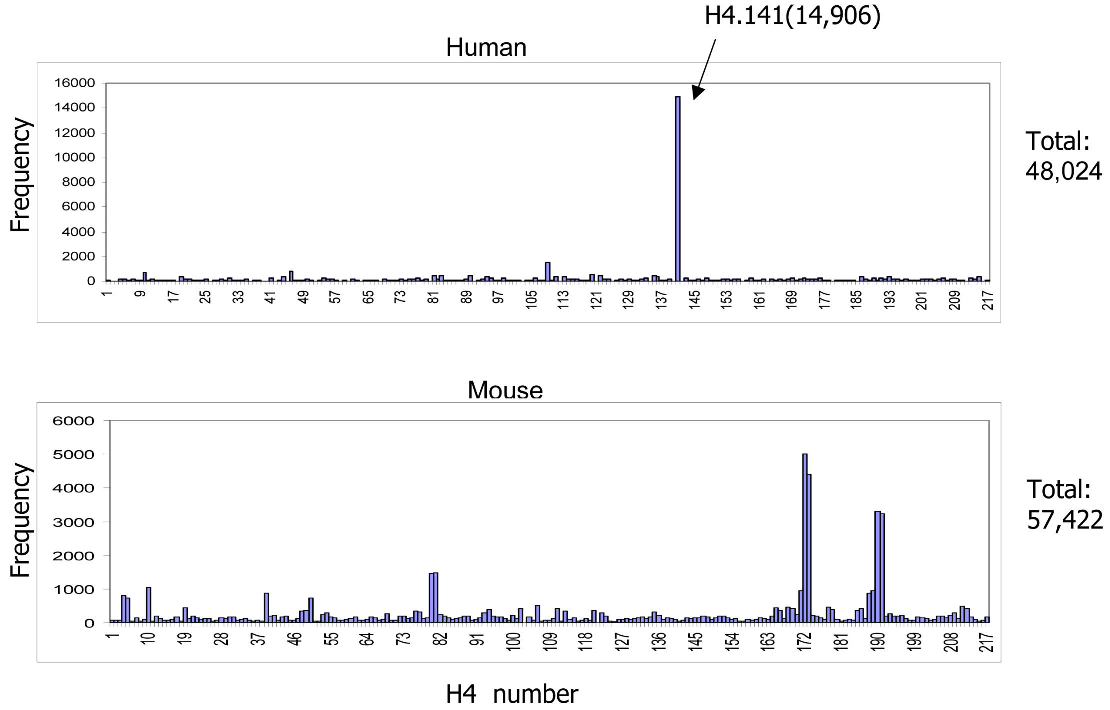


Figure 3.4. Frequency profile of 217 literature derived sequences in the mouse and human. Frequency (i.e., occurrence) of each of the 217 H4 elements from the literature found in the entire human and mouse genome. H4.141 motif is shown to be overrepresented in the human, but not mouse genome due to its association with the Alu repeats. H4.109 is also associated with Alu repeats and is over represented with frequency of 1557.

(AluSx), M90058_HSAL002952 (AluJ), M88006_HSAL001283 (AluJ), L05920_HSAL001628 (AluJ) [11]. We then searched the promoters of human genes for the instances of H4.141 binding to specific Alu elements, as opposed to their consensus. We identified 486 genes with Alu elements in the -10kb to +10kb relative to TSS (+1). Alignment of those Alu elements identified H4.141 to be located in the left arm of the Alu elements (Fig. 3.5). Since these elements were located in the promoter regions they could be involved in regulation of adjacent genes.

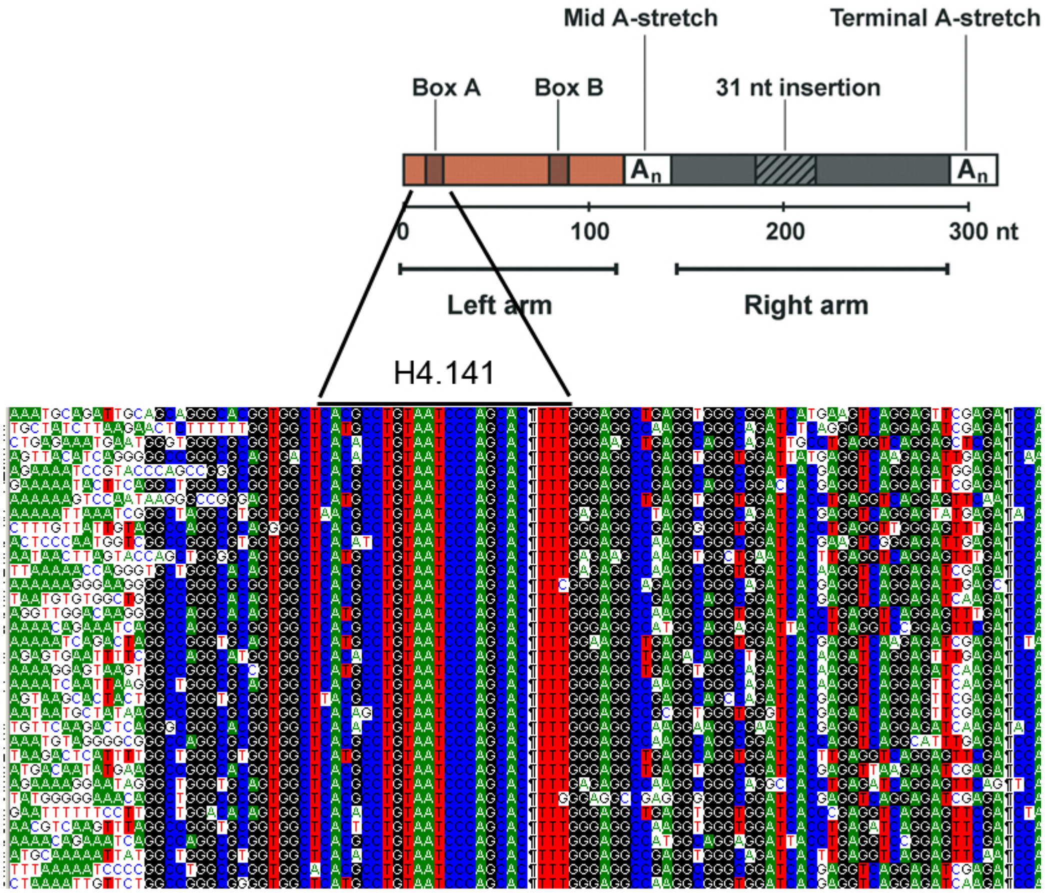


Figure 3.5. H4.141 found in the left arm of Alu sequences. To investigate location of H4.141, we aligned 486 Alu sequences, containing H4.141, from promoters of human genes. Each line is one Alu element from the promoter of a target gene. Shown is a small sample of total hits. The H4.141 consistently maps to the left Alu arm.

PBM3 design

The protein binding microarray (PBM) approach was used to identify Alu/HNF4α binding. A custom PBM version 3 (PBM3), a next generation of PBM for HNF4α, was designed specifically to include Alu sequences. Since PBMs cannot accommodate all possible 13 mers from all possible Alu sequences, an attempt was made to focus only on the most frequent or the most likely 13mers to bind HNF4α. The RepBase database was used to identify possible Alu 13mer subsequences. Every unique 13 mer from every Alu element consensus from RepBase database was extracted to make the Alu/13 mer library. Due to computational constraints, Alu/13mer library was used to search the human chromosome 21 instead of the entire genome. The frequency of each Alu/13mer from chromosome 21 was multiplied by 65.5 (since chromosome 21 size is 1/65.5 of the whole genome) to estimate genomic frequency; top 100 were included on the PBM3. The Alu/13mer database was further searched with support vector machine version 2 (SVM2) model. Every 13mer was assigned an SVM2 score to predict their likelihood of being bound by HNF4α; top 100 were included on PBM3. SVM2 was described in Section 2.3.7. Additional sequences on the PBM3 were top sequences predicted by SVM2 from human promoters, and top the 500 sequences bound to the PBM2.

HNF4α binds to Alu repeats *in vitro*

Overall in PBM3 27 out of 200 Alu derived 13-mers have been bound to human HNF4α2 significantly (>3 SD better than random controls, intensity score >0.74.). The binding did not show a significant difference between human isoforms HNF4α2 and HNF4α8, so average of the 4 sub grids 2xHNF4α2 and 2xHNF4α8 was used for PBM score. An exact match search, 0 mismatch, of the entire human genome shows that there are $\sim 820,000$ of these 27 sequences total in the genome. Although this is far less than the estimate total 3 million estimated total Alu sequences, this could be explained by slight variations in Alu sequence repeats. Intestingly, the consensus of those Alu sequences resembles non canonical binding sites from Fig. 2.31 and is shown in Fig. 3.6. For a complete list of Alu sequences significantly bound by HNF4α to the PBM3 and their estimated frequencies in the genome (hg18) see Table 3.1.

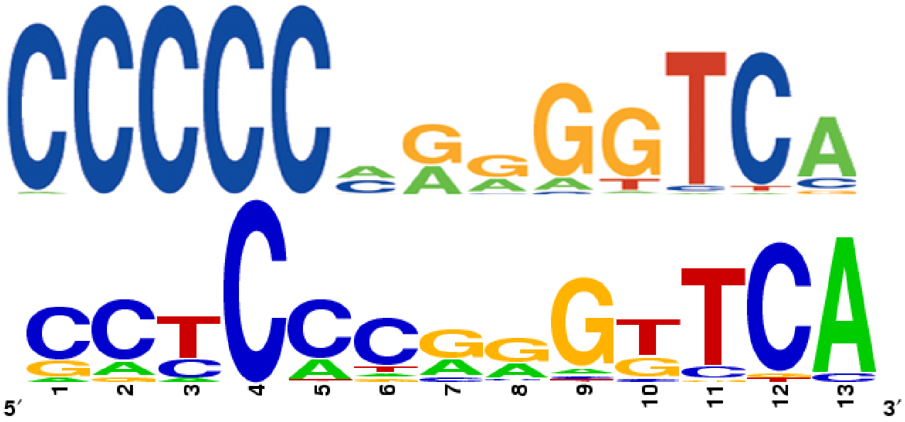


Figure 3.6. PWM for HNF4α non canonical sequences resembles PWM for sequences bound to Alu by HNF4α. Non canonical sequence logo, from Fig. 2.31, top. PWM motif for 27 Alu sequences bound by HNFα2 and HNF4α8, bottom.

Table 3.1. Alu subsequences significantly bound by human HNF4α2 and HNF4α8.

Alu sequence	**SVM score**	**Est Num**[1]	**PBM Score**[2]	**ProbeID**
CCCCCCAGGTTCA	1.789485	1,448	9.20016871	probe.11726_PBM3.320_aluscore8_rc
CCCCCCGGGTTCA	1.870117	1,888	9.02052817	probe.315_PBM3.315_aluscore2_rc
GCCCCCCGGGTTC	1.357512	1,424	6.75575924	probe.386_PBM3.386_aluscore92_rc
CCCCCAGGCTGGA	1.012541	9,824	4.597375	probe.7865_PBM3.261_alufreq46_rc
CCCCCTGGGTTCA	1.617804	2,256	4.51988568	probe.330_PBM3.330_aluscore21_rc
CCCCCGGGGTTCA	1.582771	984	4.2830797	probe.4135_PBM3.333_aluscore25_rc
CCTCCCCAGTTCA	1.47209	1,784	2.16820507	probe.11760_PBM3.354_aluscore52_rc
CCTCCACCTCCCA	1.084826	33,056	2.00582425	probe.4043_PBM3.241_alufreq25_rc
CCCCTGGGGTTCA	1.361815	920	1.36644783	probe.4185_PBM3.383_aluscore89_rc
CCTCCCACGTTCA	1.559414	1,072	0.9469763	probe.11741_PBM3.335_aluscore28_rc
CCTCCCGGGTCCA	1.539296	1,240	0.92074469	probe.11744_PBM3.338_aluscore32_rc
CCTCCCAAAGTCC	1.516302	4,120	0.91619789	probe.11746_PBM3.340_aluscore37_rc
CCTCCCGAAGTGC	1.115117	8,104	0.87336098	probe.4069_PBM3.267_alufreq52_rc
CCTCCCATGTTCA	1.500661	1,392	0.85694443	probe.345_PBM3.345_aluscore42_rc
GATCACGGGGTCA	1.498916	1,376	0.84839468	probe.4148_PBM3.346_aluscore43_rc
CCTCCCAAAGTGC	1.244758	526,272	0.84157283	probe.7823_PBM3.219_alufreq2_rc
CCTCCCGAGTTCA	1.631172	8,856	0.82439954	probe.7867_PBM3.263_alufreq48_rc
CCTCCCAAGTTCA	1.831154	5,144	0.81617825	probe.7915_PBM3.311_alufreq98_rc
CCTCCCAAGGTGC	1.433985	3,912	0.81181077	probe.11770_PBM3.364_aluscore64_rc
CCACCCAGGTTCA	1.519125	1,072	0.80220594	probe.4141_PBM3.339_aluscore36_rc
GATCGCGAGGTCA	1.341403	1,496	0.76963318	probe.4196_PBM3.394_aluscore100_rc
GATCAAGAGGTCA	1.629565	2,352	0.76383983	probe.7932_PBM3.328_aluscore19_rc
CCTCCCGGGTTCA	1.526696	189,992	0.75888256	probe.7825_PBM3.221_alufreq4_rc
CCTCCCAGGTCCA	1.624052	832	0.75693753	probe.7933_PBM3.329_aluscore20_rc
CCTCCCAAAGTTC	1.326812	7,832	0.75557577	probe.4073_PBM3.271_alufreq56_rc
CCACCCGGGTTCA	1.539869	1,032	0.75457148	probe.4139_PBM3.337_aluscore31_rc

[1] Est Num. Estimation of the frequency of these sequences in the genome, by searching chromosome 21 and multiplying it by the proportional size of the genome. The average number of a unique random 13mer in the hg18 genome is ∼50.

[2] PBM score. We considered only the sequences bound with >3 SD from the mean. (Mean = 0.46 , SD = 0.14)

HNF4α Binds to Promoter Alu repeats *in vivo*

To investigate HNF4α binding to Alu repeats in the promoters of HNF4α target genes *in vivo*, chromatin immuno-precipitation followed by PCR analysis (ChIP) for HNF4α was conducted. Several criteria were used for selecting a potential Alu sequences for PCR analysis. First, Alu repeats had to be contained in the promoter region of a gene, -5kb to +1 kb. Second, the gene containing the Alu repeat had to be down regulated > 1.4 fold in expression profiling of HNF4α RNAi in HepG2 cells as measured by Affymetrix microarray as described in chapter 2 (Section 2.3.8). Third, the Alu repeat must contain a probable HNF4α binding site from Table 3.1, or sequences H4.141 or H4.109. Fourth, the Alu repeats had to be amendable to primer design and PCR, a challenging task due to the nature of amplifying repeat sequences. Overall, 47 sets of primers for a total of 35 genes were designed. Out of the 47 primer sets, 15 sets gave a specific signal from Input control, indicating appropriate amplification of the Alu sequence. Out of 15 primer sets with positive input control, a total of 13 gave a significant signal in the ChIP sample and did not give a significant signal, in the corresponding negative control IgG (Fig. 3.7). The positive result in the ChIP analysis for 13 Alu containing genes, strongly suggests that HNF4α binds to Alu repeats *in vivo*. For a complete list of PCR primers giving a positive ChIP signal (see Table 3.2).

Table 3.2. Table of Alu primer sequences.

Gene Symbol	Primer Name	Forward primer (5' to 3')	Reverse primer (5' to 3')	Size	Dist
CANX	CANX	GCCCAGGGTTTTTCTAGACC	TAGCTGCTTCCCCAGGTAGA	551	- 1 kb
SOCS2	SOCS2	TGAGGGAGTAAACCTTGCAG	AGAATGCTCCAACCCTGATG	487	- 1 kb
SOD2	SOD2	GTGTTGGGGTGAAAAAGGAA	ACTAGCCTGCACTCCCTTCA	530	- 1 kb
TTR	TTR	ATGCCCAATGCAGAAGAGTC	AAGGAAAAACCCTTGGCAGT	509	- 1 kb
PRODH2	PRODH2	CCCAAATGTCCATCAAAAGG	CACGCATGTATTCCCAACAT	549	- 0.5 kb
PRLR	PRLR	TTGCTGGTGTCATTTGATGC	CAGCTAACAGAACCAGGTGGA	671	- 2.5 kb
GSTM4	GSTM4	GGAATGACCAAATGGGTGAA	GACGATAGCACCATGCACAC	554	- 0.5 kb
IL32	IL32	GAATTCCTAAGCCCCAGGAC	AGACGTCTCTTCCCTCACGA	616	- 0.5 kb
ATPIF1	ATPIF1	TGACCATAGCTTGGGGAAAC	CAGCCCACGATTTCAATTCT	591	- 0.5 kb
APOM	APOM	ATGGGGTCTTGCTATGTTGC	GCTGAGGCTTGCGATTTAGT	499	- 2.5 kb
FEM1A	FEM1A	TGATCCAGGAATCCCACTTC	AAATGACGCGCTCACTTCTT	509	- 1 kb
APOA4	APOA4.1	GCACAGCCTCCCACATACTT	ACATAGCGACACCCCATCTC	490	- 2.5 kb
ABCC3	ABCC3	GGCTGAAGTGCAGTAGCACA	GGAGCCCCTGACTAGAAACC	300	+ 5 kb
IP6K2	IP6K2	CCCACTTCAGGATTTGGAGA	CCTGACCTCATGATCCAACC	395	+ 3 kb
P21	P21.2(+)	GCCTGTTTTCAGGTGAGGAA	AGTTTGCAACCATGCACTTG	273	
P21	P21.4(-)	GACAGCAGTGGGGCTTAGAG	TCTACCTCACACCCCTGACC	417	
NINJ1	NINJ1.1(-)	TGGGTAAACAGCATTGAGCA	AGCTGGGACTACAGGTGTGC	400	
NINJ1	NINJ1.3(-)	TGTGTGAATGGTGCTGGATT	TATTTCCAGAAGGGCAGTGG	360	
NINJ1	NINJ1.4(+)	CCACTGCCCTTCTGGAAATA	GCCCCTAGTAACAGCGTCAG	452	

3.2.2 Description of Alu Sequences Containing HNF4α Binding Sequence

The Alu elements giving a positive ChIP signal were derived from several Alu families (Table. 3.3). While most of the Alu elements were common among primate genomes and thus fairly ancient (>25 million years old), one Alu element must have inserted more recently into *IL32* gene, due to its presence in human, chimp and orangutan, but not in rhesus (Fig. 3.8). *SOCS2* and *CANX*, contain Alu elements that are present in human, but not in chimp and rhesus respectively. Since AluS families are >25 million years old and thus should be conserved across all four species it is unknown if the lack of Alu elements in *SOCS2* and *CANX* in rhesus genome are due mistakes in the genome assembly, mistakes in Alu classification or a deletion from those genomes.

Table 3.3. Family classificaiton of Alu repeats from the promoters of corresponding human genes. Presence or absence of an element in the genome of a chimp, orangutan and rhesus from UCSC genome browser are indicated.

Gene name	**Alu family**	**Chimp**	**Orangutan**	**Rhesus**	**Marmoset**	**Human**
CANX	AluSx1	+	+	-	-	+
SOCS2	AluSz	-	+	+	+	+
SOD2	AluSz	+	+	+	+	+
TTR	AluJb	+	+	+	+	+
PRODH2	AluSp	+	+	+	-	+
PRLR	AluSc8	+	+	+	+	+
GSTM4	AluSz	+	+	+	-	+
IL32	AluSq2	+	+	-	-	+
ATPIF1	AluSx1	+	+	+	+	+
APOM	AluJr	+	+	-	-	+
FEM1A	AluSq2	+	+	+	-	+
APOA4.1	AluSx1	-	+	+	+	+
ABCC3	AluSz	+	+	+	+	+
IP6K2	AluSg4	+	+	+	-	+

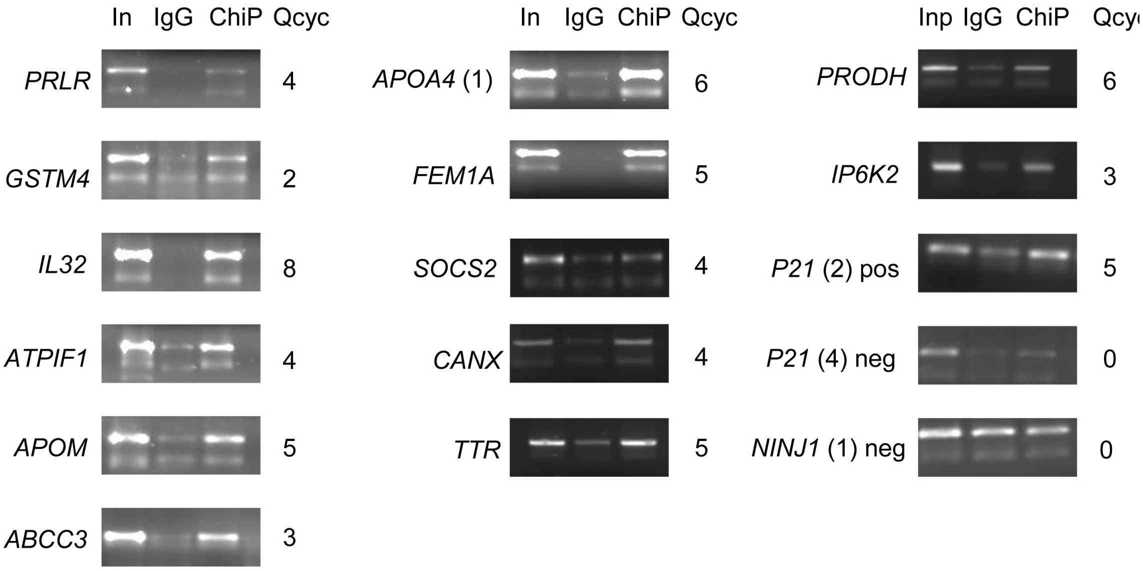


Figure 3.7. ChIP/PCR results for select genes and primer set designations. In, input control of genomic DNA. IgG, control IP with normal IgG. ChIP, Chromatin Immuno precipitation PCR. Qcyc is the difference in cycles, within linear range, between IgG and ChIP signal as determined by quantitative real time PCR (QRTPCR). X number of cycles difference signifies $\sim 2^X$ fold difference in starting material.

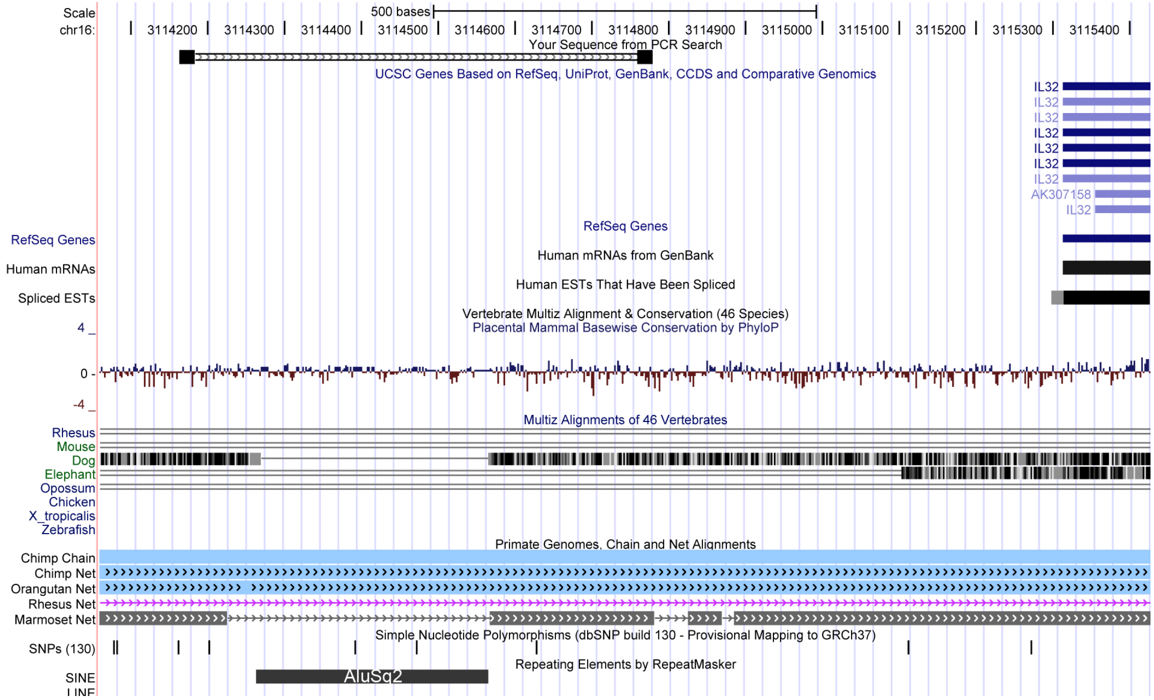

Scale
500 bases
chr16:
3114200
3114300
3114400
3114500
3114600
3114700
3114800
3114900
3115000
3115100
3115200
3115300
3115400
Your Sequence from PCR Search
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
IL32
IL32
IL32
IL32
IL32
IL32
IL32
AK307158
IL32
RefSeq Genes
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & Conservation (46 Species)
Placental Mammal Basewise Conservation by PhyloP
4 _
0 -
-4 _
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Primate Genomes, Chain and Net Alignments
Chimp Chain
Chimp Net
Orangutan Net
Rhesus Net
Marmoset Net
Simple Nucleotide Polymorphisms (dbSNP build 130 - Provisional Mapping to GRCh37)
SNPs (130)
Repeating Elements by RepeatMasker
SINE
AluSq2
LINE

Figure 3.8. AluSq2 insertion in the promoter of the human *IL32* gene is visualized using UCSC genome browser. The RepMasker track shows AluSq2 sequence. There is a clear gap in the promoters of Rhesus and Marmoset, but not in Orangutan and Chimp, illustrating that Rhesus, and Marmoset lacks AluSq2, and suggesting that Alu insertion occurred before Orangutan and Chimp diverged from a common ancestor.

3.3 Materials and Methods

3.3.1 PBM Design and Reagents

PBM3 included several distinct groups of sequences on the array. Two hundred sequences were designed to bind specifically to Alu elements. RepBase database was used to identify possible Alu 13mer subsequences. Every unique 13 mer from every Alu element consensus from RepBase database was extracted, to make Alu/13 mer library. Due to computational constraints, Alu/13mer library was used to search the human chromosome 21 instead of the entire genome. The frequency of each Alu/13mer from chromosome 21 was multiplied by 65.5 to estimate genomic frequency, top 100 were include on the PBM3. The Alu/13mer database was further searched with SVM2 model. Every 13mer was assigned a SVM2 score to predict their likelihood of being bound by HNF4α, top 100 were included on PBM3. SVM2 was described in Section 2.3.7. One hundred 13mer and 50 14mer random sequences were included as a negative control. 704 sequences were included from permutation of DR2 consensus (5'-AGGTCAAAGGTCA-3'), by permuting three adjacent positions in every combination. 768 sequences were also generating by permuting three adjacent position in every combination from a DR2 consensus (5'-AGGTCAAAAGGTCA-3'). 2,061 sequences were generated from SVM2 search of every gene with threshold >1.937, 2SD greater than SVM2 assigned score to the control sequences. Several miscellaneous sequences were included as described in Section 2.3.4. The total amount of sequences after duplicate removal was 3,802, each was replicated 4 times on the PBM for a total of 15,208 sequences. Flanking sequence was identical to the PBM2 design. 8x15k array design containing these sequences was ordered from Agilent. Human HNF4α2 and

HNF4α8 expressed in COS-7 cells were were hybridized to PBM3 as described in Section 2.3.4. Secondary antibody CTD monoclonal, (DD-H145-00 from R&D) were used for the array. The hybridization protocol was performed identical fashion to that described in Section 2.3.4.

3.3.2 ChIP and RNAi Expression Profiling

ChIP was performed exactly as described in [10]. One 150-mm plate of HepG2 cells (~80% confluent) was treated with 1% formaldehyde for 10 min at room temperature. Cross-linking was stopped by the addition of 0.125 M glycine (final concentration). Cells were harvested in cold PBS and lysed in ChIP sonication buffer (1% Triton X-100, 0.1% deoxycholate, 50 mM Tris-Cl (pH 8.0), 150 mM NaCl, 5 mM EDTA, 2 g/ml aprotinin, 2 g/ml leupeptin, 0.2 mM phenylmethylsulfonyl fluoride). The DNA fragments were sonicated to an average size of 500 bp. Immuno precipitations (IPs) were performed with anti-HNF4 (-445) and corresponding control (IgG) antibodies, and DNA-protein complexes were eluted in 1% SDS elution buffer (1% SDS, 0.1 M NaHCO3, 0.01 mg/ml herring sperm DNA). The cross-links were reversed by heating at 65 C overnight, proteins were digested by proteinase K (0.17 g/l; New England Biolabs, Ipswich, MA), and the DNA was extracted with phenol-chloroform, precipitated with ethanol, and dissolved in 100 l Tris-EDTA buffer [10 mM Tris-Cl (pH 8.0), 1 mM EDTA].

Quantitative-PCR (qPCR) was performed using BioRad IQ SYBR Green Supermix. Each reaction of 23.5 μl included 12.5 μl of Supermix, 0.25 μl of 100 nmol of each primer, 0.5 μl of template and 10 μl of ddH_2O. The qPCR was performed as follows: 95°C for 5 min (hot start), followed by 40 cycles 95 °C for 30 sec (melt) , 30 sec at Tm (anneal and extend), followed by a melt curve. Tm was determined experimentally for each primer by using a temperature gradient qPCR which was then visualized on a an agarose gel using ethidium bromide to control for product size. All qPCR was performed using the BioRad iQ5 and myQ5 thermocyclers. HepG2, RNAi array was performed as described in Section 2.4.5.

3.4 Discussion

Here we have shown that HNF4α binds to Alu-derived 13mers *in vitro* and to Alu repeats in the promoters of HNF4α target genes *in vivo*. More experiments are required to determine if HNF4α regulates transcription of these genes directly through the Alu element. One such experiment could be investigation of the Alu element with a luciferase driven promoter assay, in which the promoter of a target genes with and without the Alu repeat drives an expression of the luciferase gene. If the Alu repeat influences transcription, the luminescence would be correlated with the presence of the Alu element. However, functions other than regulation of transcription of particular genes are also possible for HNF4α binding Alu elements, for example HNF4α could be sequestered to the Alu sites as a "sink" or a storage of superfluous HNF4α molecules, by Alu elements located far from any regulatory regions.

It is unknown if HNF4α regulates these genes exclusively through these TFBS, or they are used as a complimentary sites. Nevertheless, these experiments do suggest a potential biological role for these TFBS and Alu elements in gene regulation *in vivo*. Additionally, the number of human exclusive Alu repeats is only ~7,000, suggesting that if the Alu elements did play a role in gene regulation, the vast majority of the elements have played a role in early primate evolution [12].

Bibliography

[1] S. W. V. Arsdell, R. A. Denison, L. B. Bernstein, A. M. Weiner, T. Manser, and R. F. Gesteland. Direct repeats flank three small nuclear RNA pseudogenes in the human genome. *Cell*, 26(1 Pt 1):11–17, Oct 1981.

[2] R. J. Britten. DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci U S A*, 93(18):9374–9377, Sep 1996.

[3] R. Cordaux and M. A. Batzer. The impact of retrotransposons on human genome evolution. *Nat Rev Genet*, 10(10):691–703, Oct 2009.

[4] E. H. Davidson and R. J. Britten. Organization, transcription, and regulation in the animal genome. *Q Rev Biol*, 48(4):565–613, Dec 1973.

[5] P. L. Deininger and M. A. Batzer. Alu repeats and human disease. *Mol Genet Metab*, 67(3):183–193, Jul 1999.

[6] C. M. Houck, F. P. Rinehart, and C. W. Schmid. Fractionation of renatured repetitive human DNA according to thermal stability, sequence length, and renaturation rate. *Biochim Biophys Acta*, 518(1):37–52, Mar 1978.

[7] C. M. Houck, F. P. Rinehart, and C. W. Schmid. A ubiquitous family of repeated DNA sequences in the human genome. *J Mol Biol*, 132(3):289–306, Aug 1979.

[8] J. Hsler and K. Strub. Alu elements as regulators of gene expression. *Nucleic Acids Res*, 34(19):5491–5497, 2006.

[9] G. W. Humphrey, E. W. Englander, and B. H. Howard. Specific binding sites for a pol III transcriptional repressor and pol II transcription factor YY1 within the internucleosomal spacer region in primate Alu repetitive elements. *Gene Expr*, 6(3):151–168, 1996.

[10] W. W. Hwang-Verslues and F. M. Sladek. Nuclear receptor hepatocyte nuclear factor 4alpha1 competes with oncoprotein c-Myc for control of the p21/WAF1 promoter. *Mol Endocrinol*, 22:78–90, 2008.

[11] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110(1-4):462–467, 2005.

[12] E. S. Lander, L. M. Linton, B. Birren, and C. N. et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[13] D. Laperriere, T.-T. Wang, J. H. White, and S. Mader. Widespread Alu repeat-driven expansion of consensus DR2 retinoic acid response elements during primate evolution. *BMC Genomics*, 8:23, 2007.

[14] G. E. Liu, C. Alkan, L. Jiang, S. Zhao, and E. E. Eichler. Comparative analysis of Alu repeats in primate genomes. *Genome Res*, 19(5):876–885, May 2009.

[15] S. L. Martin, D. Bushman, F. Wang, P. W.-L. Li, A. Walker, J. Cummiskey, D. Branciforte, and M. C. Williams. A single amino acid substitution in ORF1 dramatically decreases L1 retrotransposition and provides insight into nucleic acid chaperone activity. *Nucleic Acids Res*, 36(18):5845–5854, Oct 2008.

[16] R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine. Which transposable elements are active in the human genome? *Trends Genet*, 23(4):183–191, Apr 2007.

[17] J. Norris, D. Fan, C. Aleman, J. R. Marks, P. A. Futreal, R. W. Wiseman, J. D. Iglehart, P. L. Deininger, and D. P. McDonnell. Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *J Biol Chem*, 270(39):22777–22782, Sep 1995.

[18] S.-L. Oei, V. S. Babich, V. I. Kazakov, N. M. Usmanova, A. V. Kropotov, and N. V. Tomilin. Clusters of regulatory signals for RNA polymerase II transcription associated with Alu family repeats and CpG islands in human promoters. *Genomics*, 83(5):873–882, May 2004.

[19] C. M. Rubin, C. M. Houck, P. L. Deininger, T. Friedmann, and C. W. Schmid. Partial nucleotide sequence of the 300-nucleotide interspersed repeated human DNA sequences. *Nature*, 284(5754):372–374, Mar 1980.

[20] B. G. Thornburg, V. Gotea, and W. Makaowski. Transposable elements as a significant source of transcription regulating signals. *Gene*, 365:104–110, Jan 2006.

When you have a hammer, everything looks like a nail.

Attributed to Mark Twain

Chapter 4

Investigation of Binding Sequences of Lef/TCF and COUP-TF2, and Effects of Linoleic Acid and PGC1α on HNF4α DNA Binding

4.1 Introduction

Protein Binding Microarrays (PBMs) are a powerful technology. While the PBM in Chapters 2 and 3 were developed for HNF4α, the protocol and techniques are amendable to any other transcription factors (TFs). Indeed, there are TFs that have consensus sequences similar to that of HNF4α, and should therefore potentially bind to a significant number of sequences on the PBM developed for HNF4α. Those TFs are NR's superfamily members including COUP-TF2 (NR2F2), RXR (NR2R1), PPARs (NR1C) [18], and another TF family of genes belonging to lymphoid enhancer-binding factor 1/T-cell specific factors (LEF/TCFs) such as *LEF1*, *TCF7*, *TCF7L1*, *TCF7l2* [8]. Therefore, we leveraged the PBM technology to investigate the binding specificity of two important transcription factors TCF-1 and COUP-TF2 and compared the specificity to that of HNF4α. Potential overlap in their DNA sequence preferences could be an indication that these TFs compete for DNA binding and regulation of target gene promoters. Additionally, we leveraged the PBM technology to investigate the allostery, or differential binding affinity, of HNF4α on DNA under the influence of its endogenous ligand linoleic acid (LA) [35], and potential allostery under the influence of the coactivator PGC1α.

4.2 LEF/TCF PBM

4.2.1 Introduction

The LEF/TCF family of proteins are an essential component of the canonical Wnt signaling pathway. The Wnt signaling pathway is a well studied pathway that is critical in a multitude of cellular processes, most notably, regulation of cell fate, morphogenisis, apoptosis, development and cancer. Additionally, it is well known to participate in other biological processes necessary for the function of an organism, such as regulation of the immune system [6, 23]. The importance of the Wnt pathway is underscored by its conservation in metazoan animals; the Wnt pathway is remarkably similar even between mammals and the nematode *C. elegans* [8]. In brief, Wnt is a signaling peptide which binds to receptor frizzled (Fz), triggering a series of steps that lead to stabilisation of β-catenin via phosphorylation, which is degraded in cytoplasm in the abscence of Wnt signaling. Phospho-β-catenin in turn translocates to the nucleus and binds to the LEF/TCF family of proteins. LEF/TCF family proteins without β-catenin typically repress transcription, but LEF/TCF bound in a β-catenin complex is able to bind to a wide variety of genes and activate a variety of cellular events (Fig. 4.1).

HNF4α and LEF/TCF have previously been shown to compete for DNA response elements on the promoters of genes [3, 7]. However, they have not been shown to compete for binding sequences. Since, LEF/TCF and HNF4α consensus binding sequences are similar it is possible that HNF4α and LEF/TCF compete for binding sites directly on the DNA.

4.2.2 Results

Crude nuclear extract from COS-7 cells containing FLAG-tagged TCF-1 protein transcribed by human *TCF7* gene, (one of the members of LEF/TCF family), was hybridized to PBM3, described in Section 3.2.1, and has showed strong immunoflourescence when interrogated with the anti-FLAG antibody. Overall, 358 out of 3,000 unique DNA sequences bound TCF-1 significantly better than random controls (>2 st. dev. from mean binding affinity of random sequences). While the number of sequences that bound is significantly smaller than the number of sequences for HNF4α (358 vs $\sim$ 1,758), that was expected since PBM3 was developed for HNF4α. These sequences when analyzed for position frequencies represent LEF/TCF consensus sequence (5'C/T-CTTTG-A/T-A/T3') and PWM (Fig. 4.2) [1, 8]. To identify potential TCF-1 target genes the human genome was searched from -2kb to transcription start (+1) site for every human gene. We excluded 59 of the

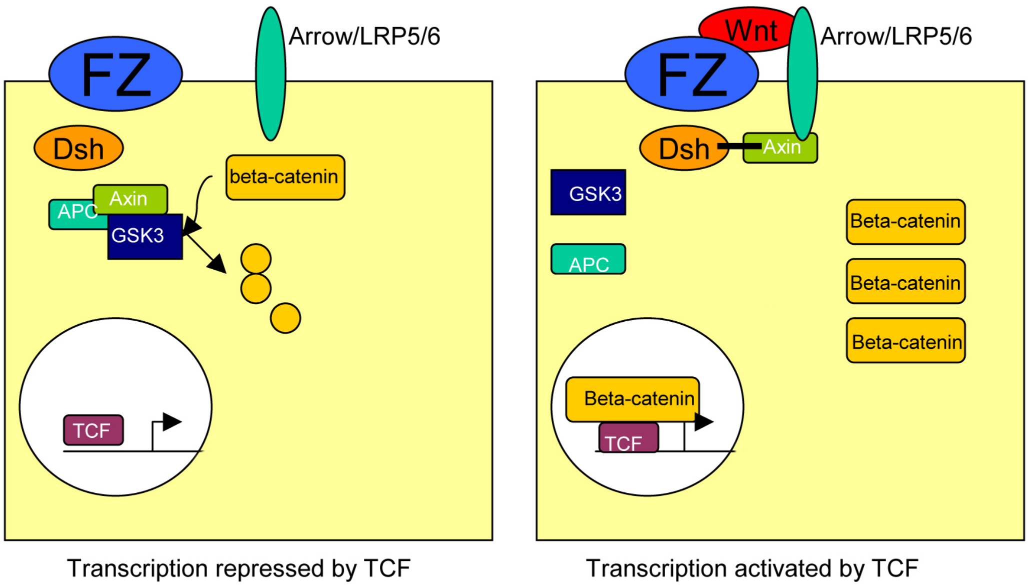


Figure 4.1. Overview of Wnt signaling. In cells not exposed to Wnt (left panel), β-catenin is degraded through the interactions with axin, APC, and GCK3 kinase. LEF/TCF is a nuclear protein that acts as a repressor in the absence of β-catenin. When Wnt activates Fz/LRP receptor, the receptor interacts with Actin and Dsh preventing degradation of β-catenin. β-catenin then accumulates in cytoplasm and subsequently translocates to the nucleus, where it binds to the LEF/TCF. β-catening/LEF/TCF complex then acts as a transcriptional activator and activates the expression of target genes. Figure adapted from [23].

sequences associated with Alu repeats, sine they will generate a large number of false positive ; reducing the number that bound TCF-1 was 299 sequences. Overall, 698 genes with an exact binding sequence for TCF-1 in their promoter regions were identified (-2kb to +1). When subjected to gene ontologogy (GO) analysis, those genes were overrepresented in 28 statistically significant categories listed in Table 4.1. One of those categories was ventral spinal cord development, a well known Wnt pathway function [34], hence validating our approach. Additionally, the Wnt pathway is known to be involved in cell adhesion, also an overrepresented GO term [23]. However, a fair number of categories implicate Wnt pathway in the regulation of transport and metabolism in the cell, a role not typically attributed to Wnt. We next compared the DNA sequences bound by TCF-1 to those bound by HNF4α. The majority of TCF-1 binders (235 sequences) also bound HNF4α, a significant overlap (Fig. 4.3). This result suggests that HNF4α and LEF/TCF might compete for binding sites in the promoters of target genes. Additionally, the intersection PWM appears to resemble the HNF4α PWM than TCF-1 PWM. This suggest that competition occurs primarily on HNF4α target genes.

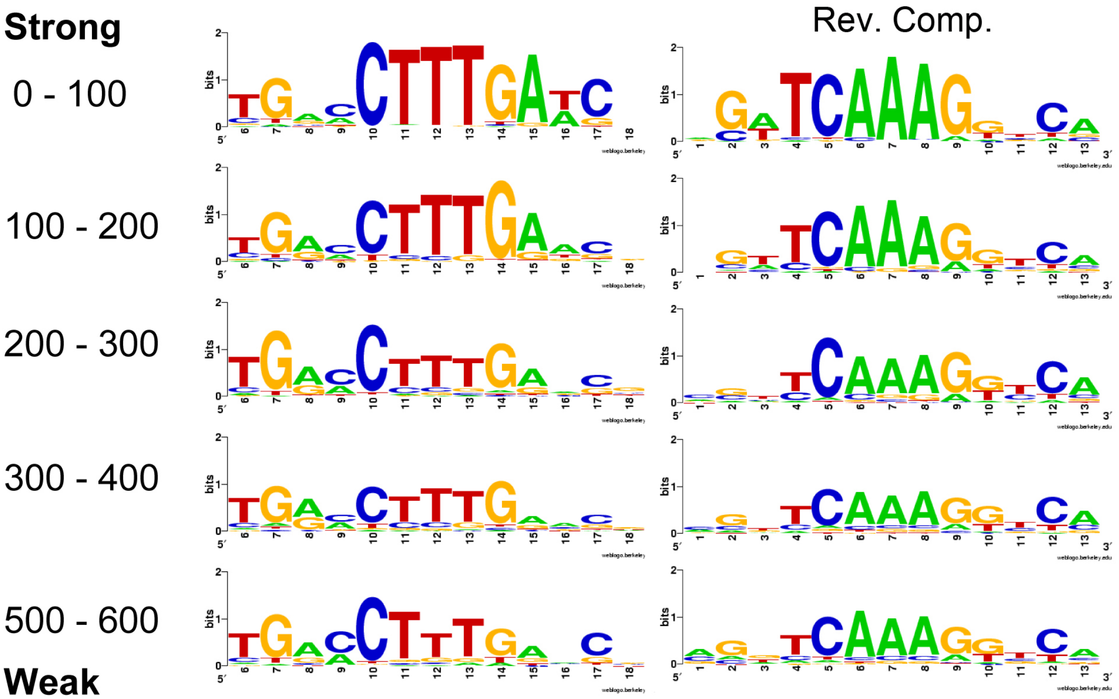


Figure 4.2. PWM for human TCF-1 from the analysis of PBM3. PWMs were created from 100 sequence groups sorted in order of affinity.

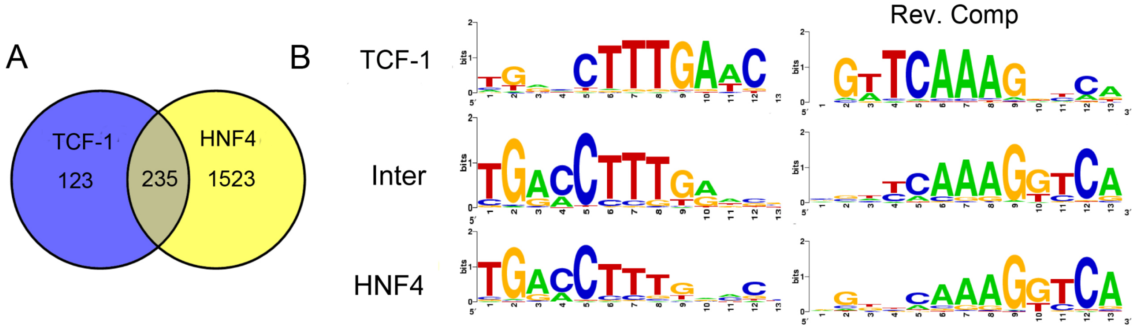


Figure 4.3. Comparison of human TCF-1 and human HNF4α2 and HNF4α8 binding sequences. **A**. Venn diagram showing the number of binding sequences from HNF4α and LEF/TCF from PBM3. All sequences bound >2SD higher intensity compared to random controls. **B**. TCF-1, PWM for TCF-1 only; Inter, PWM for Intersection only; HNF4α; PWM for HNF4α only. Left and right PWMs are reverse complements of TCF-1 and HNF4α.

Table 4.1. Statistically significant GO terms from search with TCF-1 binding sequences from PBM3 of annotated human genes (hg18) -2kb to +1.

Go Id	Go Biological Process Term	P-Value
GO:0006865	amino acid transport	0.002335
GO:0016485	protein processing	0.003324
GO:0046058	cAMP metabolic process	0.003873
GO:0015674	di-, tri-valent inorganic cation transport	0.004627
GO:0015837	amine transport	0.004797
GO:0051604	protein maturation	0.006034
GO:0046942	carboxylic acid transport	0.009483
GO:0015849	organic acid transport	0.009932
GO:0021517	**ventral spinal cord development**	0.012999
GO:0006816	calcium ion transport	0.017052
GO:0051605	protein maturation by peptide bond cleavage	0.019175
GO:0060255	regulation of macromolecule metabolic process	0.0264
GO:0043255	regulation of carbohydrate biosynthetic process	0.031088
GO:0010676	positive regulation of cellular carbohydrate metabolic process	0.031088
GO:0045913	positive regulation of carbohydrate metabolic process	0.031088
GO:0010468	**regulation of gene expression**	0.032713
GO:0009187	cyclic nucleotide metabolic process	0.034106
GO:0010565	regulation of cellular ketone metabolic process	0.036206
GO:0016337	**cell-cell adhesion**	0.037208
GO:0006325	chromatin organization	0.040445
GO:0080090	regulation of primary metabolic process	0.044482
GO:0019222	regulation of metabolic process	0.045705
GO:0031647	regulation of protein stability	0.046417
GO:0008285	negative regulation of cell proliferation	0.046679
GO:0007155	**cell adhesion**	0.04686
GO:0021515	**cell differentiation in spinal cord**	0.04786
GO:0022610	biological adhesion	0.04799

Bold categories are known LEF/TCF family protein functions.

4.2.3 Materials and Methods

Human flag tagged TCF-1 protein was expressed from *TCF7* gene cloned into PC-DNA4TO vector and transfected into COS-7 cells (designation Flag-dnTCF1Emut). Flag-dnTCF1Emut differs from wtTCF1 protein by replacement of CRARF amino acids (critical for activation of transcription) with VAVAL in the E-tail (activation of transcription) [1]. Cell culture, overexpression and nuclear extracts experiments were performed as described in Sections 2.3.3 and 2.3.1 by M. Waterman group at UC Irvine. PBM of PBM3 design was performed essentially as described in Section 2.3.4 and Section 3.2.1 with the following exceptions: primary antibody anti-flag M2 monoclonal (Sigma-Aldrich), secondary antibody was fluorescent anti-mouse NL635 (R&D Systems). Exact match searches were performed on human genome (hg18) using seqmap [11]. GO analysis was performed using DAVID [5] for biological processes only.

4.2.4 Discussion

We have successfully applied the PBM technology to flag tagged human TCF1 protein, identifying 358 binding sequences for TCF-1. Furthermore, we identified a significant overlap between HNF4α and TCF-1 binding specificity that warrants further investigation. While there are some reports of HNF4α and LEF/TCFs competing on the same pathway, the interaction on the same binding sequence has not been shown [3, 7]. Wnt pathway and HNF4α are both involved in development, but the interactions between the two pathways have not been reported previously. These experiments suggest that there could be binding sequence competition, between HNF4α and LEF/TCF, although additional experiments are required to prove the competition in *in vivo* and in vitro. Since HNF4α and Wnt pathway have been implicated in development and cancer such competition could be involved in the biology of those processes [6, 23]. Additionally, the GO analysis suggests that TCF-

1 regulate genes involved in metabolism and transport, a promising novel function for LEF/TCFs and Wnt pathway that awaits experimental verification.

4.3 COUP-TF2

4.3.1 Introduction

NR2F2, commonly known as chicken ovalbumin upstream promoter transcription factor 2 (COUP-TF2), is a nuclear receptor closely related to HNF4α. Like HNF4α COUP-TF2 knockout is an embryonic lethal in mice; the embryos display gross abnormalities in head, heart, and vasculature formation [26]. These phenotypes suggest a COUP-TF2 function in development, specifically head, heart, and spine. Additionally COUP-TF2 is known to be involved in cancer, cell fate and cell differentiation [26]. COUP-TF2 binds as a homodimer to the same consensus sequence as HNF4α (5'-AGGTCAAAGGTCA-3') or as monomer to the half site (5'-AGGTCA-3'), presumably due to their close amino acid and evolutionary similarity [18] (Fig. 4.4). However, the full spectrum of binding sequences for COUP-TF2 is not known; therefore we probed PBM with COUP-TF2 protein.

4.3.2 Results

PBM

Crude nuclear extract from COS-7 cells transfected with COUP-TF2 expression vector was applied to PBM2. The bound protein was detected by monoclonal antibody against COUP-TF2 followed by conjugated goat anti mouse fluorescent secondary antibody. Overall 1044

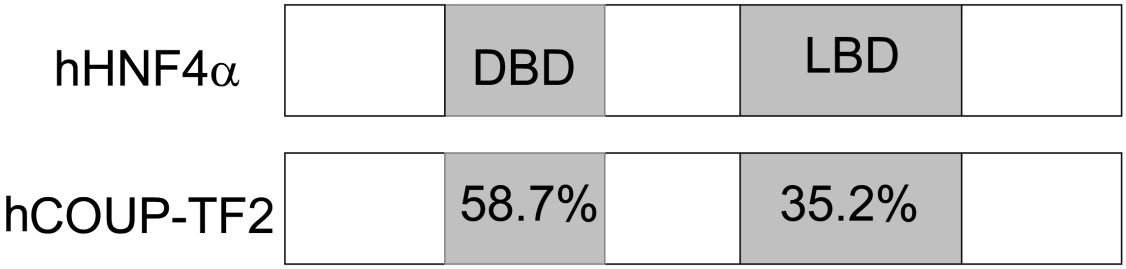


Figure 4.4. Human COUP-TF2 shares a large degree of amino acid conservation with HNF4α. Illustrated are the % amino acid identity in the DNA binding domain (DBD) and ligand binding domain (LBD) between human COUP-2 and human HNF4α.

out of 3000 sequences bound TCF-1 significantly better than random controls (>2 st. dev. from mean binding affinity of random sequences). This number only somewhat smaller than the number of sequences for HNF4α (1044 vs ~ 1400), which was not too surprising considering COUP-TF2 was hybridized to the PBM developed specifically for HNF4α. Interestingly, the overall PWM for COUP-TF2 is slightly different from that of HNF4α. Furthermore, there is a bimodal peak in the histogram for the sequence intensity values of COUP-TF2, something never observed for HNF4α and requiring further investigation (Fig. 4.5). Interestingly, the two peaks correspond to two different subgroups of sequences falling into two different PWM motifs, primary and secondary. In the primary motif COUP-TF2 prefers the right half site (peak 3), but as the binding affinity decreases, but before it drops to below that of a random control, the COUP-TF2 preference for the consensus switches from right half site to the left half site (Fig. 4.6, peak2). It is not known whether if this effect is caused by COUP-TF2 acting as a monomer or due to the interaction between two subunits in a homodimer. Peak 1 mainly consists of non or weak binding sequences.

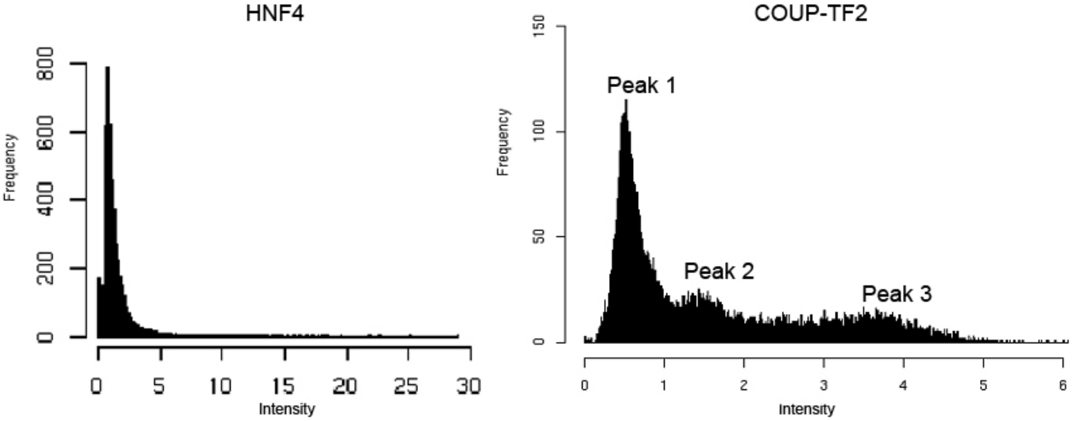


Figure 4.5. Histogram of peak intensities for HNF4α and COUP-TF2 PBMs. Peak 1 consists mainly of nonbinding sequences and weak binding sequences, peak 2 and peak 3 are present in COUP-TF2, but were never observed in HNF4α. Statistically significant binding sequences are $>\sim 1$ for both HNF4α and COUP-TF2.

We next determined the overlap between the DNA binding specificity of COUP-TF2 and HNF4α. 686 sequences bound significantly(>2 st. dev. from random) to both factors, while 358 sequences were bound only by COUP-TF2 and 718 sequences only by HNF4α2. Hence, there was considerable overlap in binding between these two highly related NRs, as well as many unique sequences that were distinct for each receptor (Fig. 4.7).

Genomic Analysis and GO

A search of the -2kb to +1 of all annotated human genes (hg18) with sequences bound by COUP-TF2 in PBM2 revealed significant patterns. Overall 2606 genes were found to contain 1044 sequences significantly bound by COUP-TF2, giving 256 statistically significant biological process GO categories containing overrepresented genes (p-value < 0.05). While there are too many categories to include in a separate table, they group into several distinct classes: Metabolism, development, transport, and apoptosis. Metabolism is further subdivided into: lipid and cholesterol metabolism, both predicted functions of COUP-TF2 [19, 29, 33]. Development was represented primarily by spinal and neuronal morphogensis, a well known and well studied COUP-TF2 function [25, 29]. While evidence for COUP-TF2 being involved in biological process "transport" is sparse in the literature, there is some evidence that it regulates apolipoprotein family genes,indeed COUP-TF2 originally was named (apolipoprotein regulatory protein 1) Arp1 because it was cloned by virtue of its ability to bind a DNA response element in the human *APOA1* promoter [2, 15, 21, 24]. There is a convincing report of COUP-TF2 involvement in apoptosis as a repressor in the brain using a knockout mouse [12]. A new category previously not observed for COUP-TF2 is Rho/Ras signal transduction, this novel function is in need of additional experimen-

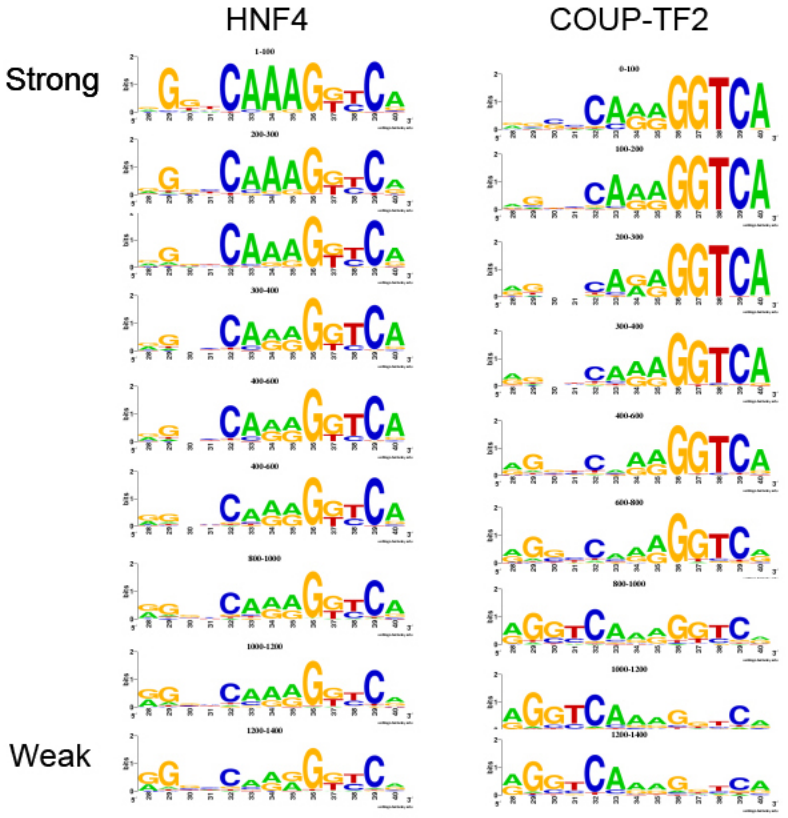


Figure 4.6. Comparison of PWM between COUP-TF2 and HNF4α. Sequences from HNF4α and COUP-TF2 from PBM2 are sorted, grouped into 200 sequence groups and used to generate PWM. "Strong" are sequences with the highest binding intensity on the PBM, while "weak" are the sequences HNF4α and COUP-TF2 bind least well, but better than random controls. For HNF4α PWM "degrades" (i.e. keeps the same profile while getting less and less specific), while COUP-TF2 PWM "shifts" over to the left, radically changing the profile as the binding affinity decreases.

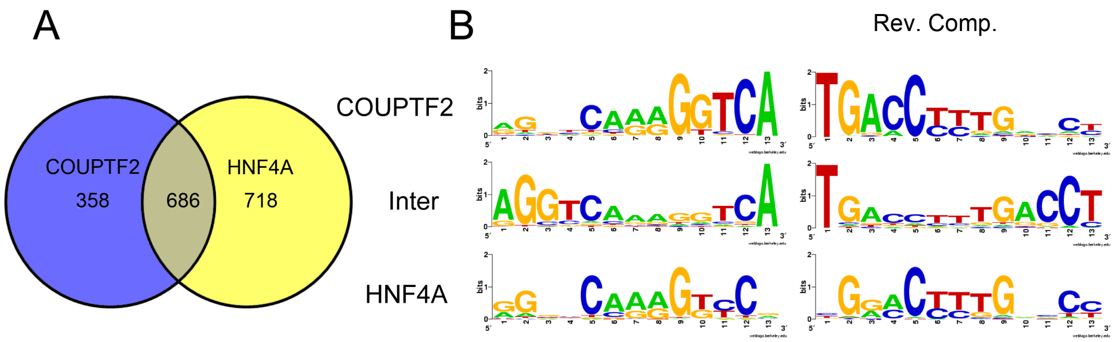


Figure 4.7. Comparison of overlap in binding sequences between human HNF4α and human COUP-TF2. **A**. Venn diagram showing the number of binding sequences by HNF4α and COUP-TF2. All the sequences have better than >2SD in binding intensity than random controls. **B**. COUPTF2: PWM of sequences that are specific to COUP-TF2 only, Inter: PWM of sequences that bind to both HNF4α and COUP-TF2. HNF4A:PWM of sequences that bind only HNF4α. Left and right PWMs are revere complements of each other.

tal confirmation. These results are summarized in Table 4.2, the number of “sub” categories does not make them more or less significant in the analysis. All categories are significant with a P-value < 0.05.

4.3.3 Materials and Methods

Crude nuclear extract containing human COUP-TF2 from a pMT2 vector overexpressed in COS-7 cells was applied to PBM (Fig. 4.8) [21]. Cell culture, overexpression and nuclear extracts were performed as described in Sections 2.3.3 and 2.3.1. While COUP-TF2 gave a great signal on a western blot (Fig. 4.8), the true concentration for COUP-TF2 is unknown and needs to be determined in future experiments. PBM experiments was performed essentially as described in Section 2.3.4 on PBM2 and PBM3 with the following exceptions: the primary mouse monoclonal antibody anti-COUP-TF2 was purchased from

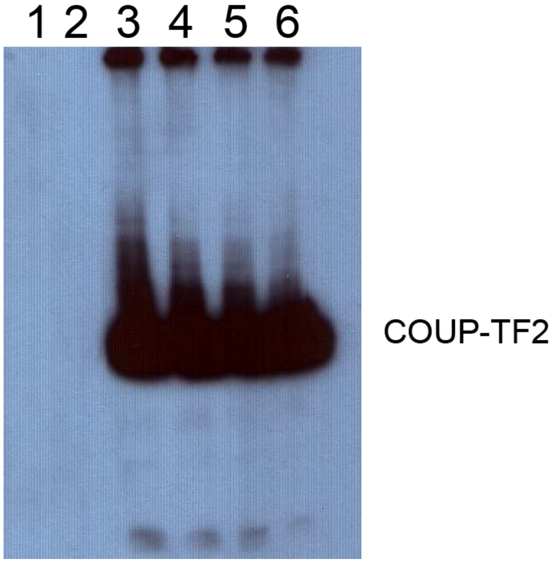


Figure 4.8. Immunoblot of COUP-TF2. Immunoblot analysis indicating specificity of antibody. ~20 μg of COS-7 nuclear extract loaded in each lane. Mock transfected COS-7 controls lane1 and lane 2, did not show a specific band indicating that COUP-TF2 monoclonal antibody is specific. Lanes 3,4,5,6 show a very strong band indicating >30 ng/μl of COUP-TF2.

Table 4.2. GO categories from COUP-TF2 binding sequences exact match promoter search of annotated humang genes (hg18) -2kb to +1.

GO general categories	Number of sub categories*
Development (neuronal, muscle)	22
Morph (neuronal)	9
Apoptosis	3
Differentiation (immune system)	10
Transport	14
Metabolism (lipid, sterol)	52
Biosynthetic process (lipid, sterol)	18
Signal transduciton (Rho, Ras)	4
Misc	124

*The number of "sub" categories does not make them more or less significant. All categories are significant with P-value < 0.05.

R&D systems (PP-H7147-00), secondary fluorescent donkey anti-mouse antibody NL635 was also purchased from (R&D Systems). Exact match search was conducted using seqmap [11]. Gene ontology analysis was conducted using David [5].

4.3.4 Discussion

These experiments describe for the first time a PBM experiment using full length human receptor COUP-TF2 in a crude nuclear extract in which over 4,000 unique sequences were assayed for COUP-TF2 binding; 1044 new binding sequences were identified as bound. These 1044 sequences reveal an ability of COUP-TF to change preference from right to left half of the DR1. Since COUP-TF2 binds a large number of sequences that are also bound by HNF4α on the PBM, it is possible that like TCF-1, COUP-TF2 competes for DNA binding with HNF4α. In fact there are reports of such interactions [14, 21]. Since

COUP-TF2 tends to repress transcription [26] and HNF4α activates transcription [30] this competition could present additional layer of regulation of transcription in development and essential for adult liver function. The high degree of overlap between HNF4α TFBS and COUP-TF2 TFBS suggests that these interactions are widespread. While COUP-TF2 is not always expressed in the same tissues as HNF4α, but we anticipate that since COUP-TF1 NR2F1 is more widely expressed and is highly related to COUP-TF2 these interactions are commonplace in many cell types. Furthermore, the sequences specific to COUP-TF2 binding, could attribute for differences between HNF4α and COUP-TF2 target genes.

4.4 Linoleic Acid (LA) and HNF4α interaction

4.4.1 Introduction

Recently our group identified the endogenous ligand for HNF4α [35]. This ligand is the essential fatty acid LA (C18:2ω6). While the ligand "de-orphaned" HNF4α, its function remains elusive because it does not appear to affect transcriptional activation activity of HNF4α, as do other nuclear receptor ligands. However, it is possible that the ligand would act in a non traditional manner and either affect the binding affinity of HNF4α to DNA and/or affect the specificity with which HNF4α binds DNA response elements. To test these hypothesis, we conducted PBM on HNF4α expressed in the presence and absence of LA.

4.4.2 Results for HNF4α in Presence or Absence of LA

HNF4α expressed in COS-7 cells in the presence or absence of LA bound successfully to PBM3. When comparing the HNF4α signal with and without LA, no significant difference in individual sequence signal intensity was detected as shown in Fig.4.9. If there were a significant difference between HNF4α with LA vs HNF4α without LA, then the correlation between arrays would have been significantly lower. Additionally, if HNF4α with LA were to prefer a different subset of sequences, then we would have seen a group of sequences with significantly different binding affinities from HNF4α without LA (i.e., points distant from the line); a result we did not observe. However, more detailed Kd studies are required to eliminate the possibility of minor differences in binding affinities due to the effect of the ligand.

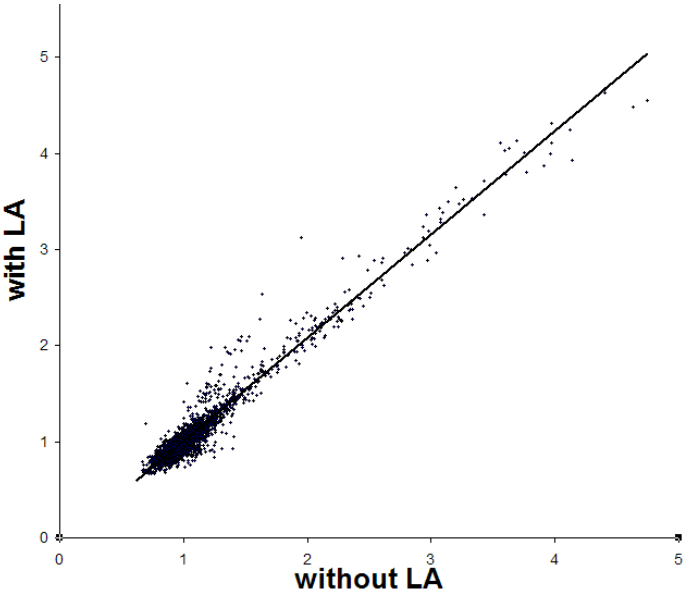


Figure 4.9. Results of PBM experiments for HNF4α in the presence and absence of LA. Since correlation between HNF4α with and without LA is R^2=~94 it is highly unlikely that LA affects HNF4α binding. This experiment was replicated one more time with similar result (data not shown).

4.4.3 Materials and Methods

Cell Culture and Nuclear Extracts

COS-7 cells (ATCC CRL-1651), maintained at 37C and 5% CO2 in Dulbecco's modified Eagle's medium (DMEM) (CellGro) supplemented with 10% bovine calf serum (Hyclone) lipid-depleted serum ("Stripped Serum") [Controlled Process Serum Replacement (CPSR3), Sigma] and penicillin/streptomycin (CellGro), were transiently transfected with pMT7.rHNF4a2 via calcium phosphate precipitation as previously described [6]. Where applicable, 30 μM of exogenous fatty acids were added 12-24 hr after transfection. Cells were harvested after additional 24-34 hr incubation. Nuclear extracts were prepared as previously described [10]. "Mock" transfected samples contained either the pMT7 empty vector or no DNA.

PBM

HNF4α in the presence and absence of LA were incubated on the PBM version 3 essentially as described in Section 2.3.4. The culture and extraction was performed essentially as described in Sections 2.3.3 and 2.3.1. The differences were as follows: rat HNF4α2 nuclear extract from COS-7 cells grown with and without added LA. Primary antibody was rabbit N1.14 1:100 against N terminus of HNF4α. Secondary antibody was cy3 anti rabbit northern lights from R&D(NL008) diluted 1:50.

4.4.4 Discussion

The hypothesis that LA would affect the HNF4α/DNA interaction was not confirmed in these experiments, thus we cannot reject the null hypothesis. Recently, reports described DNA allostery for TFs such as NFκB and Glucocorticoid Receptor (GR) [16, 20] (see below for a more detailed description). While these reports describe the effect of DNA sequence on cofactor recruitment, it has recently been shown that the ligand binding domain (LBD) interacts with the DNA binding domain (DBD) in the full length NR [22]. Thus, it was possible that LA would have an effect of HNF4α DNA binding through LBD/DBD interaction. This did not turn out to be the case under the conditions we investigated. It remains to be determined whether if DNA allostery is possible for other NRs or for HNF4α under conditions not yet investigated. It is also possible that PBM is unable to detect DNA allostery and a more sensitive method needs to be used.

4.5 HNF4α and coactivator PGC1α on the PBM

4.5.1 Introduction

DNA allostery occurs when the DBD of a TF acts as an allosteric site in response to binding a specific DNA sequence, subsequently causing an effect at a different portion of the TF. While ligand binding did not induce an allosteric response on the ability of HNF4α to bind DNA, it is possible that the DNA response element could act as a ligand and induce and allosteric reaction the HNF4α LBD, and therefore alter its ability to recruit coactivators. DNA allostery has been hypothesized for quite some time, but only recently examples have been found for GR and NFκB [16, 20]. Those studies show that GR and NFκB are able to change their activation activity depending on the DNA sequence they bind, potentially by recruiting different coregulators. HNF4α is known to recruit a number of coregulators such as: CBP, Trip3, SMILE, PGC1α, SRC-1 and GRIP1 [4, 9, 17, 31, 32]. Since GR is also a NR and shares amino acid conservation with HNF4α, it is highly possible that HNF4α can also differentially recruit a coactivator depending on the DNA sequence that HNF4α binds. To test this hypothesis, HNF4α was hybridized together with the coactivator PGC1α on the PBM. PGC1α was chosen for the cofactor investigation due to its strong binding to HNF4α in pull down assay as well as supershifts in EMSA (Sladek lab unpublished data). Additionally PGC1α is well known to be involved in metabolism and is an important cofactor for certain HNF4α target genes [27, 28]. PBM is ideal type of experiment to determine the NR/coactivator allostery due to its ability to investigate of thousands of binding sequences for protein interaction simultaneously. To investigate PGC1α/HNF4α interaction, two different primary antibodies were then used, one for HNF4α and one for PGC1α and imaged simultaneously on the array. A lack of correlation between HNF4α and PGC1α signal intensities would indicate a differential recruitment and thus DNA allostery.

4.5.2 Experimental Design

To determine whether the coactivator PGC1α affects HNF4α DNA binding activity, an experiment with a combinatoric design was conducted (Fig. 4.10). Experimental Grid 1 containing rat HNF4α2 expressed in COS-7 cells and a peptide fragment of PGC1α fused to a GST-HA tag expressed in bacteria, was probed with anti-HNF4α antibody as well as fluorescent secondary antibody. If Grid 1 gave a significantly different signal from Grid 2, we would conclude that PGC1α had a significant effect on HNF4α binding.

In Grid 2, the HNF4α protein was incubated with a similar amount of non fused GST protein; That grid was also probed with anti-HNF4α antibody. Grid 3 and Grid 4 had a similar set up, except they were probed with anti-GST antibody. Grid 3 allows us to ascertain whether PGC1α is binding to HNF4α on the DNA. Grid 4, serves as a control to verify that any signal observed with the anti-GST antibody is due to the presence of HNF4α and not PGC1α binding DNA directly, or via some other protein in the COS-7 nuclear extract. Since PGC1α is not know to bind DNA, any signal in Grid 4 was not expected.

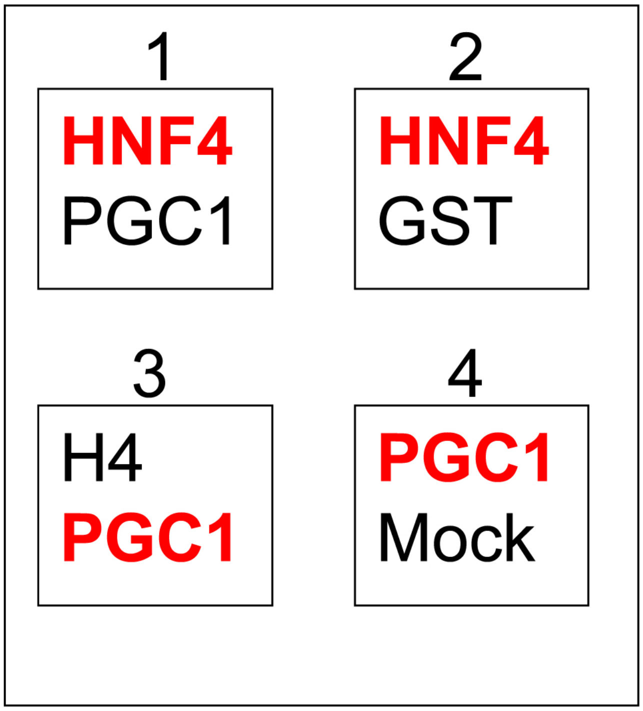


Figure 4.10. HNF4α and PGC1α PBM experimental design. Each grid represents a separate PBM experiment, conducted on the same slide. While the wash conditions were identical, the proteins and antibodies were varied according to the indicated scheme. Grid 1 contained HNF4α and GST-PGC1α, in which HNF4α was detected by anti-HNF4α antibody. Grid 2 contained HNF4α and GST alone, in which HNF4α was being detected by anti-HNF4α antibody. Grid 3 contained HNF4α and GST-PGC1α, in which PGST-GC1α was detected by anti-GST antibody. Grid4 contained GST-PGC1α and COS7 mock (lacking HNF4α), in which GST-PGC1α was detected by anti-GST antibody.

4.5.3 Results

Regression analysis of the array results is summarized in Fig. 4.11. The controls worked as expected. The mock array (Grid 4) did not produce any signal except for a few spots; none of the other grids had a significant correlation with Grid 4. Grid 1 and Grid 2 had a $R^2 = 0.56$, a high correlation, but not the usual >0.90 for HNF4α alone, suggesting that GST-PGC1α is affecting HNF4α binding activity.

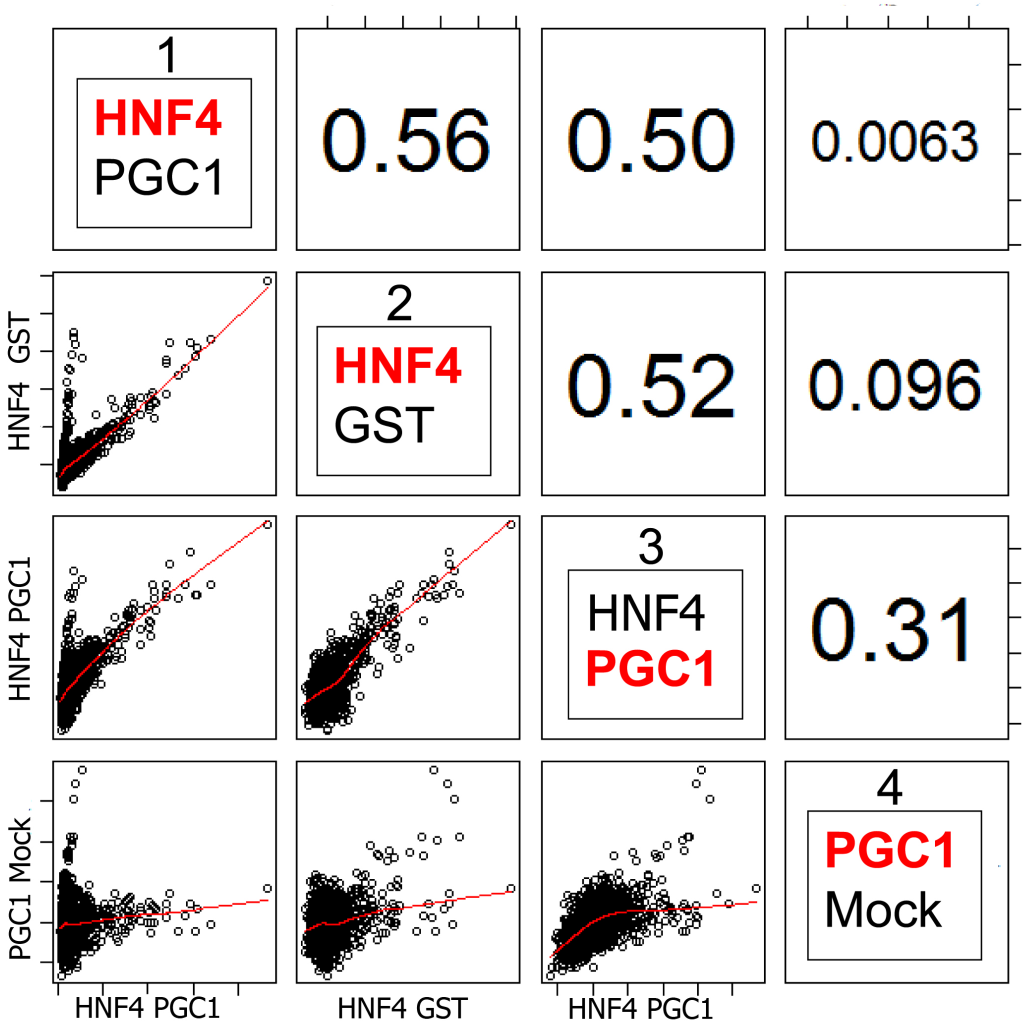


Figure 4.11. Scatterplot matrix of correlations for grids described in Fig. 4.10. Correlation coefficient squared (R^2) values for the corresponding grids are displayed in the upper right hand. The red labeling indicates that the protein is being detected using primary antisera.

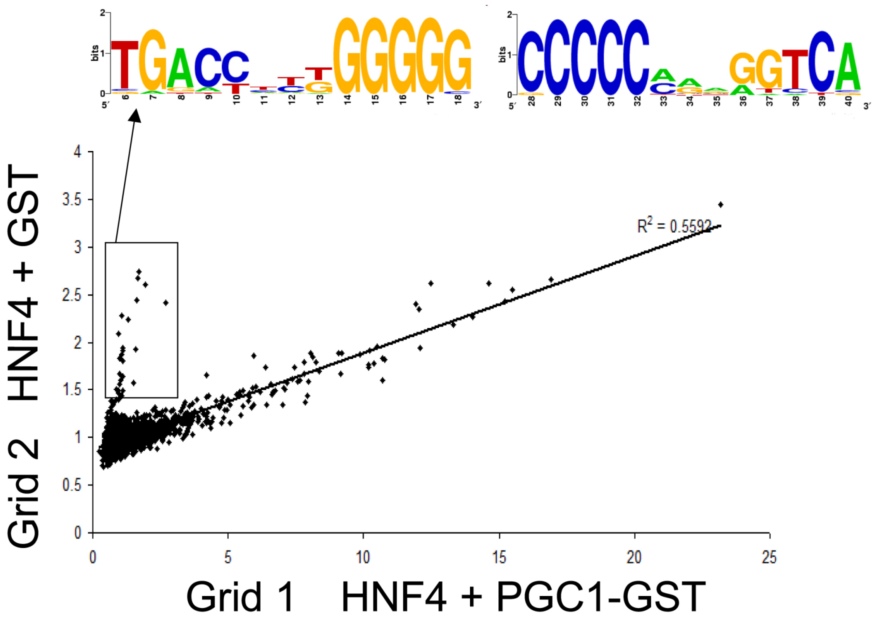


Figure 4.12. Scatter plot for grids Grid 1 vs Grid 2, from figure 4.11. R^2 = 0.55. The sequences that do not correlate (detected in Grid 2 but not in Grid 1) are highlighted (box). The PWM for these sequences and a reverse complement resemble non canonical sequences from 2.31.

Further investigation revealed non canonical sequences bound in Grid 2 (HNF4α with GST), but not Grid 1 (HNF4α with GST-PGC1α), from 2.31 are the primary reason for reduction of correlation (Fig. 4.12). The reason for this difference is unknown. Correlation between Grid 1 vs Grid 3 is almost identical to Grid 1 vs Grid 2. Both Grids 1 and 3 had HNF4α and GST-PGC1α, but they were detected by two different antibodies (anti-HNF4α and anti-GST). The low correlation is once again is due to sequences with non canonical motifs. It is possible that non canonical sequences are binding in non specific fashion since these sequences also significantly bound to Grid 4 (mock). This result is particularly troubling because Grid 4 should not have HNF4α protein, and thus there should not be any significant binding on that grid. One possible explanation for this result is an antibody/DNA interaction. The ability of the non canonical sequences to be bound in the mock, suggests that the non canonical sequences do not represent actual HNF4α binding. It is possible that PGC1α binds to the non canonical sequences, additionally its possible that protein from other grids is acting as a contaminant, or trace amounts of HNF4α from COS-7 cells binding is amplified by PGC1α binding.

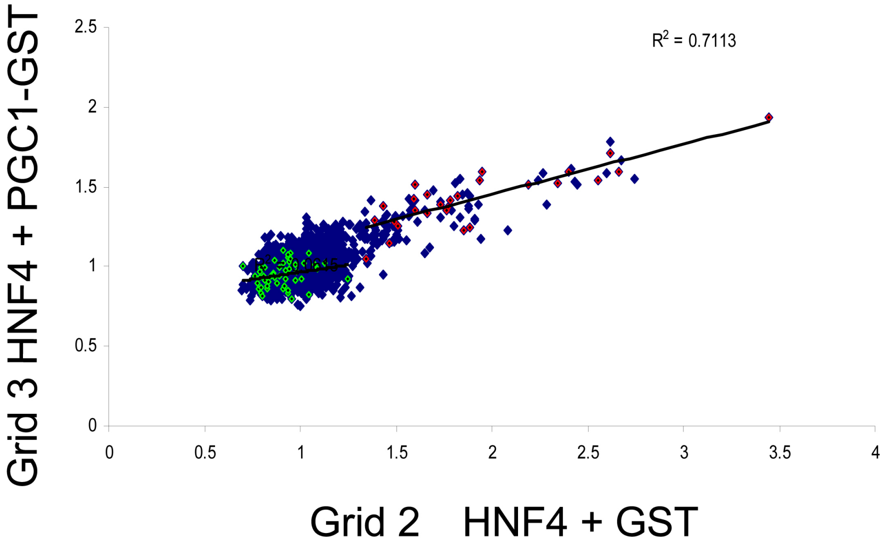


Figure 4.13. Scatter plot for grids Grid 2 vs Grid 3, from figure 4.11. R^2 = 0.52 Green are the sequences that are random controls (negative controls), red are significantly bound sequences from PBM1 (positive controls). The correlation for the positive controls alone is R^2 = 0.73

Finally, experimental result Grid 2 vs Grid 3, shows medium correlation $R^2 = 0.52$ (Fig. 4.13). However, this result is important because it shows for the first time an interaction of the NR and coactivator on the PBM. While, we cannot detect DNA allostery on this PBM, it could be because of the high noise that is created by requiring the multiple interactions between antibody, HNF4α and PGC1α (Fig. 4.14).

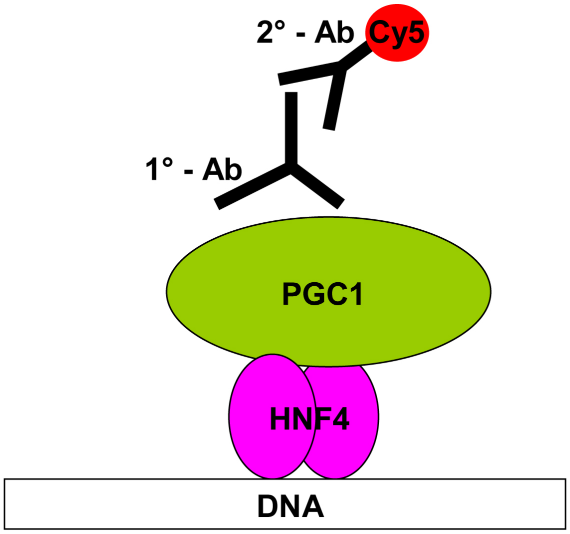


Figure 4.14. Interactions of HNF4α with PGC1α on the PBM. Multiple interactions required to happen in order to detect PGC1α binding to HNF4α and binding to DNA. The number of these interactions each with individual kinetics could explain the low correlation in the PBM experiments.

4.5.4 Materials and Methods

Nuclear extracts containing rat HNF4α2 expressed in COS-7 cells as described in Sections 2.3.3 and 2.3.1 and bacterially expressed human GST-HA-PGC1α fragment aa 91 to 408 in pGEX plasmid as described in [13] at concentration of 0.5 μg/ul (Fig 4.15). N1.14 rabbit anti HNF4α was used to detect HNF4α.

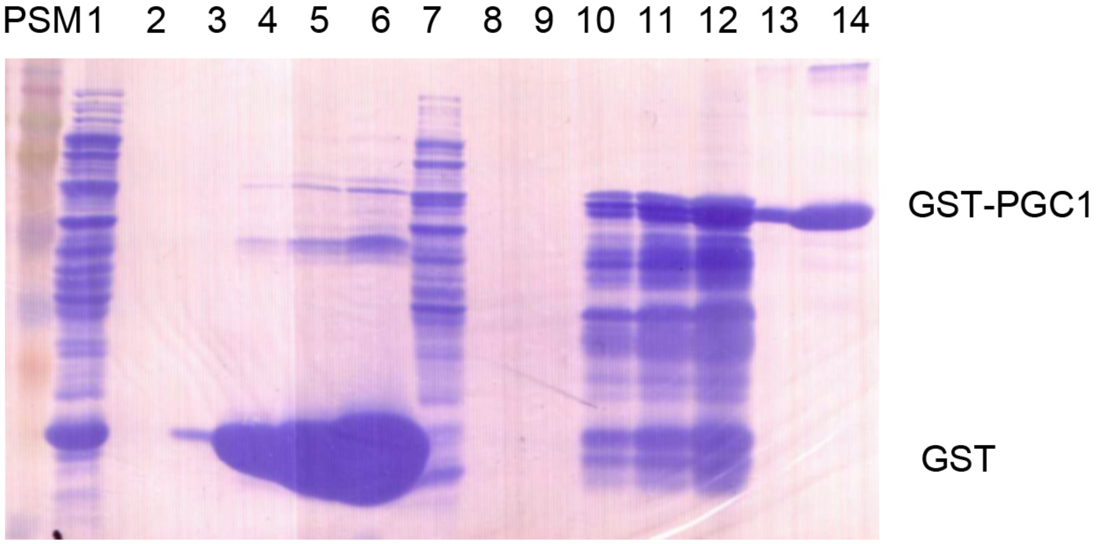


Figure 4.15. Comassie blot of bacterially expressed immuno precipitated and eluted PGC1α. Lanes 1-6 contain various fractions of GST transducted bacterial extract. Lane 1 contained bacterial supernatant, lane 2 and 3 contained column flow through, lanes 4,5,6 contained 1, 2.5 and 5μl of eluted GST respectively. Lanes 7-12 contained various fractions of bacterially expressed GST-HA-PGC1α protein. Lane 7 contained bacterial supernatant, lane 8 and 9 contain column flow though, lanes 10, 11 and 12 contain 1, 2.5 and 5μl of eluted GST respectively. Lanes 13 and 14, contain 1 and 5μl of bovine serum albumin. The comassie blot shows successful elution of GST-HA-PGC1α protein.

PBM3 were performed as previously described in Section 2.3.4 with the design described in 3.2.1, with the following exceptions. Anti-GST monoclonal from Sigma was used to detect GST-HA-PGC1α (G1160). Secondary donkey anti rabbit Cy5 analog from R&D (NL005), and donkey anti mouse Cy5 analog from R&D (NL008) were used to detect N1.14 and Anti-GST, respectively. HNF4α and PGC1 were preincubated for 30 minutes at room temperature before being applied to the PBM.

4.5.5 Discussion

DNA allostery is an important concept in gene regulation. While DNA allostery was not observed in our experiments, the interaction between PGC1α, HNF4α and DNA was observed. Clearly, optimization is required to improve the signal to noise ratio, and to prove DNA allostery. In these experiments HNF4α and PGC1α were preincubated together before applying to the array, a sequential binding of HNF4α followed by the coactivator was not explored and could prove fruitful. However, an alternative explanation is that HNF4α/PCG1α interaction does not exhibit the DNA allostery. In this case investigation of different coactivator/NR pair such as GRIP1/GR might demonstrate allostery. Additional, appropriately designed experiments should be able to answer these questions. Netherthe-less, there is a case for optimism in the ability of PBM technology to resolve these interactions, and we plan to perform these experiments shortly.

Overall, we successfully adapted the PBM to the two other full length transcription factors in mammalian cell types. We have identified lots of new binding sites as well as considerable overlap between HNF4α and other TFs. We have investigated the effect of ligand binding on HNF4α and investigated the effect of coactivator PGC1α on HNF4α binding, and observed for the first time a coactivator/NR interaction on the PBM. However, additional properly controlled experiments are required to observe allostery on the PBM.

Bibliography

[1] F. A. Atcha, A. Syed, B. Wu, N. P. Hoverter, N. N. Yokoyama, J.-H. T. Ting, J. E. Munguia, H. J. Mangalam, J. L. Marsh, and M. L. Waterman. A unique DNA binding domain converts T-cell factors into strong Wnt effectors. *Mol Cell Biol*, 27(23):8352–8363, Dec 2007.

[2] P. Cardot, J. Chambaz, D. Kardassis, C. Cladaras, and V. I. Zannis. Factors participating in the liver-specific expression of the human apolipoprotein A-II gene and their significance for transcription. *Biochemistry*, 32(35):9080–9093, Sep 1993.

[3] M. Colletti, C. Cicchini, A. Conigliaro, L. Santangelo, T. Alonzi, E. Pasquini, M. Tripodi, and L. Amicone. Convergence of Wnt signaling on the HNF4alpha-driven transcription in controlling liver zonation. *Gastroenterology*, 137(2):660–672, Aug 2009.

[4] H. Dell and M. Hadzopoulou-Cladaras. CREB-binding protein is a transcriptional coactivator for hepatocyte nuclear factor-4 and enhances apolipoprotein gene expression. *J Biol Chem*, 274(13):9013–9021, Mar 1999.

[5] J. Dennis, G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4:P3, 2003.

[6] C. Fuerer, R. Nusse, and D. T. Berge. Wnt signalling in development and disease. Max Delbrck Center for Molecular Medicine meeting on Wnt signaling in Development and Disease. *EMBO Rep*, 9(2):134–138, Feb 2008.

[7] P. Hatzis, L. G. van der Flier, M. A. van Driel, V. Guryev, F. Nielsen, S. Denissov, I. J. Nijman, J. Koster, E. E. Santo, W. Welboren, R. Versteeg, E. Cuppen, M. van de Wetering, H. Clevers, and H. G. Stunnenberg. Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol*, 28(8):2732–2744, Apr 2008.

[8] S. Hoppler and C. L. Kavanagh. Wnt signalling: variety at the core. *J Cell Sci*, 120(Pt 3):385–393, Feb 2007.

[9] H. Iwahashi, K. Yamagata, I. Yoshiuchi, J. Terasaki, Q. Yang, K. Fukui, A. Ihara, Q. Zhu, T. Asakura, Y. Cao, A. Imagawa, M. Namba, T. Hanafusa, J. ichiro Miyagawa, and Y. Matsuzawa. Thyroid hormone receptor interacting protein 3 (trip3) is a

novel coactivator of hepatocyte nuclear factor-4alpha. *Diabetes*, 51(4):910–914, Apr 2002.

[10] G. Jiang, L. Nepomuceno, K. Hopkins, and F. M. Sladek. Exclusive homodimerization of the orphan receptor hepatocyte nuclear factor 4 defines a new subclass of nuclear receptors. *Mol Cell Biol*, 15:5131–43, 1995.

[11] H. Jiang and W. H. Wong. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, 24(20):2395–2396, Oct 2008.

[12] B. J. Kim, N. Takamoto, J. Yan, S. Y. Tsai, and M.-J. Tsai. Chicken Ovalbumin Upstream Promoter-Transcription Factor II (COUP-TFII) regulates growth and patterning of the postnatal mouse cerebellum. *Dev Biol*, 326(2):378–391, Feb 2009.

[13] D. Knutti, A. Kaul, and A. Kralli. A tissue-specific coactivator of steroid receptors, identified in a functional genetic screen. *Mol Cell Biol*, 20(7):2411–2422, Apr 2000.

[14] J. A. Ladias, M. Hadzopoulou-Cladaras, D. Kardassis, P. Cardot, J. Cheng, V. Zannis, and C. Cladaras. Transcriptional regulation of human apolipoprotein genes ApoB, ApoCIII, and ApoAII by members of the steroid hormone receptor superfamily HNF-4, ARP-1, EAR-2, and EAR-3. *J Biol Chem*, 267(22):15849–15860, Aug 1992.

[15] J. A. Ladias and S. K. Karathanasis. Regulation of the apolipoprotein AI gene by ARP-1, a novel member of the steroid receptor superfamily. *Science*, 251(4993):561–565, Feb 1991.

[16] T. H. Leung, A. Hoffmann, and D. Baltimore. One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers. *Cell*, 118(4):453–464, Aug 2004.

[17] J.-F. Louet, G. Hayhurst, F. J. Gonzalez, J. Girard, and J.-F. Decaux. The coactivator PGC-1 is involved in the regulation of the liver carnitine palmitoyltransferase I gene expression by cAMP in combination with HNF4 alpha and cAMP-response element-binding protein (CREB). *J Biol Chem*, 277(41):37991–38000, Oct 2002.

[18] D. J. Mangelsdorf, C. Thummel, M. Beato, P. Herrlich, G. Schtz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon, and R. M. Evans. The nuclear receptor superfamily: the second decade. *Cell*, 83(6):835–839, Dec 1995.

[19] M. U. D. Martino, S. Alesci, G. P. Chrousos, and T. Kino. Interaction of the glucocorticoid receptor and the chicken ovalbumin upstream promoter-transcription factor II

(COUP-TFII): implications for the actions of glucocorticoids on glucose, lipoprotein, and xenobiotic metabolism. *Ann N Y Acad Sci*, 1024:72–85, Jun 2004.

[20] S. H. Meijsing, M. A. Pufall, A. Y. So, D. L. Bates, L. Chen, and K. R. Yamamoto. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science*, 324(5925):407–410, Apr 2009.

[21] M. Mietus-Snyder, F. M. Sladek, G. S. Ginsburg, C. F. Kuo, J. A. Ladias, J. E. Darnell, and S. K. Karathanasis. Antagonism between apolipoprotein AI regulatory protein 1, Ear3/COUP-TF, and hepatocyte nuclear factor 4 modulates apolipoprotein CIII gene expression in liver and intestinal cells. *Mol Cell Biol*, 12(4):1708–1718, Apr 1992.

[22] D. Moras. Structure of full-length PPARgamma-RXRalpha: a snapshot of a functional complex? *Cell Metab*, 9(1):8–10, Jan 2009.

[23] R. Nusse. Wnt signaling in disease and in development. *Cell Res*, 15(1):28–32, Jan 2005.

[24] A. Ochoa, S. Bovard-Houppermans, and M. M. Zakin. Human apolipoprotein A-IV gene expression is modulated by members of the nuclear hormone receptor superfamily. *Biochim Biophys Acta*, 1210(1):41–47, Dec 1993.

[25] F. A. Pereira, Y. Qiu, M. J. Tsai, and S. Y. Tsai. Chicken ovalbumin upstream promoter transcription factor (COUP-TF): expression during mouse embryogenesis. *J Steroid Biochem Mol Biol*, 53(1-6):503–508, Jun 1995.

[26] F. A. Pereira, M. J. Tsai, and S. Y. Tsai. COUP-TF orphan nuclear receptors in development and differentiation. *Cell Mol Life Sci*, 57(10):1388–1398, Sep 2000.

[27] P. Puigserver. Tissue-specific regulation of metabolic pathways through the transcriptional coactivator PGC1-alpha. *Int J Obes (Lond)*, 29 Suppl 1:S5–S9, Mar 2005.

[28] P. Puigserver and B. M. Spiegelman. Peroxisome proliferator-activated receptor-gamma coactivator 1 alpha (PGC-1 alpha): transcriptional coactivator and metabolic regulator. *Endocr Rev*, 24(1):78–90, Feb 2003.

[29] Y. Qiu, V. Krishnan, F. A. Pereira, S. Y. Tsai, and M. J. Tsai. Chicken ovalbumin upstream promoter-transcription factors and their regulation. *J Steroid Biochem Mol Biol*, 56(1-6 Spec No):81–85, Jan 1996.

[30] F. Sladek and S. Seidel. Hepatocyte nuclear factor 4alpha. In T. Burris and E. McCabe, editors, *Nuclear Receptors and Genetic Diseases*, pages 309–361. Academic Press, London, 2001.

[31] J. C. Wang, J. M. Stafford, and D. K. Granner. SRC-1 and GRIP1 coactivate transcription with hepatocyte nuclear factor 4. *J Biol Chem*, 273(47):30847–30850, Nov 1998.

[32] Y.-B. Xie, B. Nedumaran, and H.-S. Choi. Molecular characterization of SMILE as a novel corepressor of nuclear receptors. *Nucleic Acids Res*, 37(12):4100–4115, Jul 2009.

[33] Z. Xu, S. Yu, C.-H. Hsu, J. Eguchi, and E. D. Rosen. The orphan nuclear receptor chicken ovalbumin upstream promoter-transcription factor II is a critical regulator of adipogenesis. *Proc Natl Acad Sci U S A*, 105(7):2421–2426, Feb 2008.

[34] W. Yu, K. McDonnell, M. M. Taketo, and C. B. Bai. Wnt signaling determines ventral spinal cord cell fates in a time-dependent manner. *Development*, 135(22):3687–3696, Nov 2008.

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

What we call the beginning is often the end. And to make an end is to make a beginning. The end is where we start from.

T.S. ELIOTT FROM FOUR QUARTETS

Chapter 5

Discussion

5.1 Future Directions

Transcriptional regulation is a complicated process which we are just beginning to understand. However, in the last 30 years tremendous progress was made, starting from hypothesis that "regulatory elements exist", we have identified locations of promoter regions and enhancers throughout in the human genome [4]. We have identified general consensus sequences and PWM for many transcription factors, and discovered the mechanisms for transcription factor (TF) to DNA specificity from crystallographic studies [5, 6, 16]. However, some mysteries still remain. Not even 10 years ago a completely novel mechanism for gene regulation through miRNA was identified, and it is possible that several additional mechanisms remain to be discovered. Although multiple motifs for a single TF have been reported prior [9], the landmark paper by Bulyk et al. has identified that they are widespread [1]. In this thesis we have observed a TFs binding to several disparate PWMs result for HNF4α, with “non canonical” sequences and COUP-TF2 with “left shifting” motif (Fig. 4.6).

Ideally, we would like to have a database of every single biologically relevant binding sequence for every TFs. This database would go a long way towards complete understanding of transcriptional regulation in a mammalian cells. Alternately, a model that could predict TFBS with perfect accuracy would serve just as well. Unfortunately, that kind of model remains to be discovered. In view of that, this dissertation is just a small step in the direction of complete understanding of transcriptional regulation.

5.1.1 PBM/SVM SNP Applications

Single Nucleotide Polymorphisms (SNPs) are randomly occurring DNA sequence variations in in the human and other genomes of one nucleotide. They differ from mutations, in that they have occurred some time ago and had a chance to spread with relatively high (>1%) frequency through the population [17, 18]. The SNPs is a DNA variation, and similar to a mutation can disrupt protein function. Additionally, many examples of SNPs in the promoters of human genes have been known to alter transcriptional regulation. One of the first such examples identified was for HNF4α regulation of Factor IX leading to hemophilia [11]. Another example of SNP disrupting a binding site for HNF4α associated with a disease is the *VKORC1* gene [15]. *VKORC1* has additionally been associated with warfarin sensitivity, further emphasizing its importance [13, 20].

Previously, one had few methods for identifying transcription disrupting SNPs. Genome wide disease association (GWAS) studies can link a SNP with disease, but cannot elucidate how a particular SNP is associated with disease. One can get “lucky” and identify a TFBS

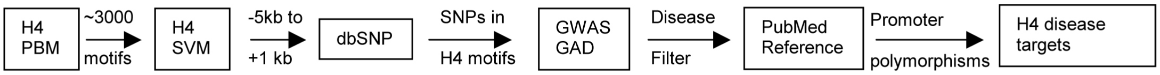


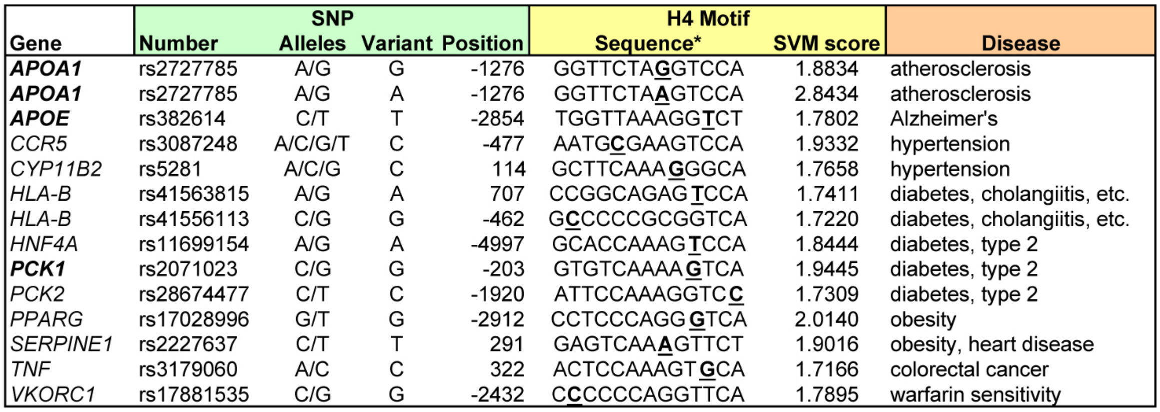
Figure 5.1. Work flow for identification of TFBS disrupting SNPs. Motifs from the HNF4α PBM2 were used to train an SVM2 model to identify potential HNF4α binding sites in SNPs from dbSNP (build 130). Hits (~6700) were cross referenced with GWAS data in the Genetic Association Database (GAD) and those genes (~340) linked to Metabolic or Cardiovascular Disease, or Pharmacogenetics were further analyzed for polymorphisms in their promoters that lead to disease.

that is present in the same location as a SNP, but this method is not amendable to high throughput analysis, and it still requires validation.

The PBM data links sequences with relative binding affinities. This database allows us to use simple exact searches to uncover SNP/TFBS locations. It also allows us to tell if a SNP allele would significantly alter the TFBS binding affinity. Since the SVM model also allows the estimation of binding affinity, it can be used to expand our number of potential TFBS disrupting SNPs (SNP/TFBS).

We have conducted a preliminary search to survey the possible number of SNP/TFBS with the SVM model. Using custom Perl scripts we have searched SNP129 database with HNF4α SVM2 in the -5kb to +1kb region that could be potentially associated with disease. The work flow is illustrated in Fig. 5.1.

Genomewide search has identified ~6,700 SNPs that are located in the promoters (-5kb to +1 kb) of genes that contain HNF4α TFBS, and can be potentially disruptive. These genes were then cross referenced to the GWAS database (GAD) to identify their relation to disease, specifically with metabolic, cardiovascular diseases, and phrmacogenetics due to HNF4α relation to those categories. Analysis has shown that ~340 of those genes were linked to one or more of these categories. Further, about a third, or ~100 of those genes



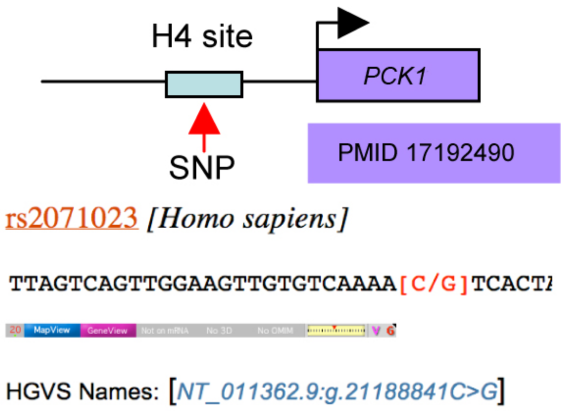
	SNP				H4 Motif		
Gene	Number	Alleles	Variant	Position	Sequence*	SVM score	Disease
APOA1	rs2727785	A/G	G	-1276	GGTTCTA**G**GTCCA	1.8834	atherosclerosis
APOA1	rs2727785	A/G	A	-1276	GGTTCTA**A**GTCCA	2.8434	atherosclerosis
APOE	rs382614	C/T	T	-2854	TGGTTAAAGG**T**CT	1.7802	Alzheimer's
CCR5	rs3087248	A/C/G/T	C	-477	AATG**C**GAAGTCCA	1.9332	hypertension
CYP11B2	rs5281	A/C/G	C	114	GCTTCAAA**G**GGCA	1.7658	hypertension
HLA-B	rs41563815	A/G	A	707	CCGGCAGAG**T**CCA	1.7411	diabetes, cholangiitis, etc.
HLA-B	rs41556113	C/G	G	-462	G**C**CCCCGCGGTCA	1.7220	diabetes, cholangiitis, etc.
HNF4A	rs11699154	A/G	A	-4997	GCACCAAAG**T**CCA	1.8444	diabetes, type 2
PCK1	rs2071023	C/G	G	-203	GTGTCAAAA**G**TCA	1.9445	diabetes, type 2
PCK2	rs28674477	C/T	C	-1920	ATTCCAAAGGTC**C**	1.7309	diabetes, type 2
PPARG	rs17028996	G/T	G	-2912	CCTCCCAGG**G**TCA	2.0140	obesity
SERPINE1	rs2227637	C/T	T	291	GAGTCAA**A**GTTCT	1.9016	obesity, heart disease
TNF	rs3179060	A/C	C	322	ACTCCAAAGT**G**CA	1.7166	colorectal cancer
VKORC1	rs17881535	C/G	G	-2432	C**C**CCCCAGGTTCA	1.7895	warfarin sensitivity

*sequence of motif may be given in reverse complement, SNP is underlined and in bold

Figure 5.2. Examples of genes identified with a SNP in a potential HNF4α binding site, within -5kb to +1 kb of a gene associated with a disease. Given are the sequences predicted to be bound by HNF4α and the corresponding SVM score (HNF4α is not predicted to bind the alternate alleles). Genes in bold are known HNF4α targets from the literature (see Table 6.1)

were reported as having polymorphisms in their promoter regions associated with the disease.

Fig. 5.2 gives examples of the genes with promoter polymorphisms in their SVM predicted HNF4α TFBS along with the SNP ID number, H4 binding sequence, position of the SNP relative to +1 and the SVM score. Several known HNF4α target genes were identified (e.g., *APOA1, APOE, PCK1*), and at least one of those genes, *PCK1* (*PEPCK*, a key glcuoneogenesis gene), had a SNP in the predicted HNF4α binding site that was directly associated with diabetes [21] (Fig. 5.3).



rs2071023 *[Homo sapiens]*

TTAGTCAGTTGGAAGTTGTGTCAAAA[C/G]TCACT

HGVS Names: [*NT_011362.9:g.21188841C>G*]

Figure 5.3. *PCK1* promoter contains a TFBS/SNP. Example of one gene, *PCK1*, in which the SNP in the HNF4α binding site has been directly linked to a disease in the literature [21].

To determine whether any of these potential HNF4α binding sites are bound *in vivo*, we cross referenced them with published HNF4α ChIP-chip data from primary human hepatocytes [8]. Shown in Fig. 5.4 are snapshots of those ChIP signals visualized in Integrated Genome Browser with the HNF4α binding site containing the SNP indicated (red arrowhead). The results show that the SNP (rs2071023) in the HNF4α binding site in *PCK1* that is associated with type 2 diabetes is indeed bound by HNF4α *in vivo*; this is the first report that this particular SNP is contained in a HNF4α binding site.

Furthermore, the PBM/SVM results indicate that HNF4α binds just one of the SNP alleles, suggesting that a loss or gain of HNF4α binding in this region of the PCK1 gene is one potential mechanism by which this SNP is results in diabetes. It is also evident from the ChIP-chip results that the HNF4α binding sites in the *HLA-B*, *PCK2* and *SERPINE1* genes that were predicted by the SVM are also bound by HNF4α *in vivo*. Whereas these SNPs have not yet been associated with a disease in the human population, all of these

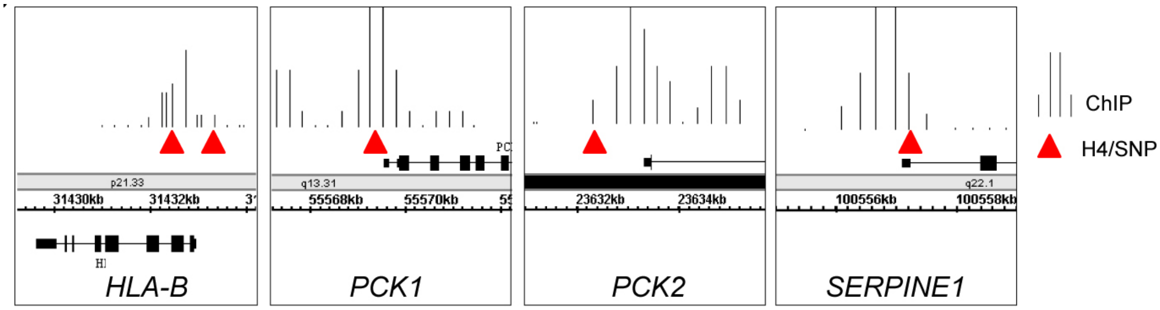


Figure 5.4. H4 ChIP-chip results from the literature [8] visualized in Integrated Genome Browser showing that the HNF4α/SNP sites in (B) are bound by HNF4α in human hepatocytes.

genes, listed in Fig. 5.2, have been shown to have polymorphisms in their promoter regions that are associated with disease. Hence, it is possible that, like the other SNPs, a SNP in the HNF4α binding site could contribute to the risk of disease, although that remains to be proven.

Even though a somewhat limited SVM model was used, capable of predicting only ~10,000 HNF4α binding sites (based on PBM data from a total of 3,000 motifs, both positive and negative binders), with a relatively small region (-5 kb to +1 kb promoter), it nonetheless shows the power of this method. Future experiments will be conducted with 1 million SNP array, so we would be able to verify thousands of SNPs not just *in silico*, but also *in vitro*. This preliminary study shows that TFBS/SNP interactions are widespread and require more study and immediate attention, making it a priority for future PBM studies.

5.1.2 Other Nuclear Receptors and Networks

We plan to conduct the PBM experiments for other TFs and NRs, to further elucidate gene regulation. Indeed, we are already working on PBMs for *RXR* and *GR*. This would allow us to identify novel TFBS for many NRs, a much needed resource. It is important to use this newly acquired regulatory information to interpret the relationships between the NRs in light of regulatory networks. For example, HNF4α is known to regulate *PPARa* as well as other NRs [2, 19], and in Fig. 5.2, we identify a new potential HNF4α binding site in the regulatory region of *PPAR*γ. The HNF4α promoter is also known to have binding sites for RXR and COUP-TFs, as well as itself [3]. These interactions could be represented as networks of transcriptional regulation. The nodes of these networks are typically TFs and the edges connect the TFs they regulate. The networks could be used for simple visualization of TF interactions, but could also be used to form hypothesis and make predictions about changes in TF regulation. Networks are the next logical step for visualizing TFBS/TF interactions which would be supremely useful to the TF community.

These networks in addition to identifying TFBS/TF interactions could contain other information such as tissue specificity, and could further incorporate known regulatory information. The tissue specificity could be obtained from the NR tissue profiles available on the NURSA website. The RNA levels of all 48 human NRs have been quantified by qRT-PCR in nearly every tissue/cell type. Another potential piece of useful information is the expression data from Gene Expression Atlas established by the Novartis Institute (http://expression.gnf.org/cgi-bin/index.cgi). This database gives extensive tissue distribution information on all known human genes based on Affymetrix gene chip results as well as other data as it becomes available.

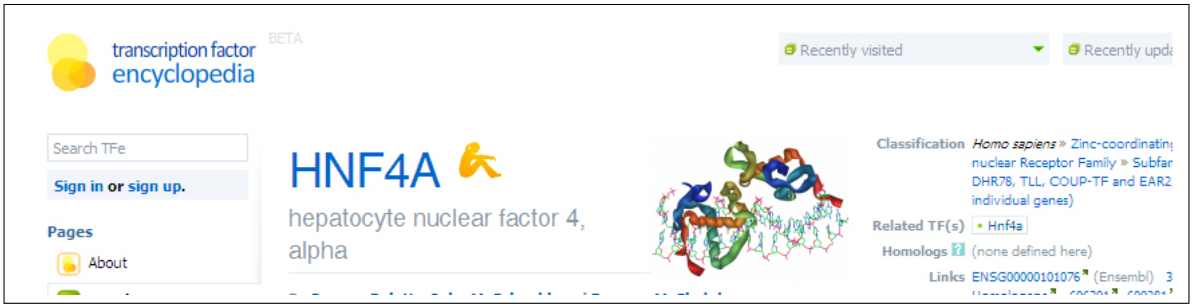
5.2 Data Sharing

5.2.1 Overview of Existing Databases

In addition to acquiring large amounts of accurate data on transcriptional factor binding positions and sequences, a vital challenge is dissemination of this data. It is not enough to attach the long list of regulatory elements and genes in a supplement to a manuscript. It is important to make the results of the genomewide study interactive and easily available to specialists and nonspecialists alike. To this end, several databases have been developed and are rapidly growing. Bulyk lab has been sharing their raw and analyzed data at UniPROBE database (http://the_brain.bwh.harvard.edu/uniprobe/), as of now it hosts data on >200 TF binding motifs derived from PBM data [7]. TRANSFAC is a venerable commercial database of PWMs, that has been around since the 1990's and is slowly growing by incorporating new motif information (http://www.biobase-international.com/) [22]. It has information on >12,000 TFs, but the coverage has been generally spotty. Additionally, UCSC genome browser(http://genome.ucsc.edu/) [12] started incorporating ChIP-chip and ChIP-seq data on their tracks generated by ENCODE project (http://www.genome.gov/10005107) [14], and it is only a matter of time until Encode starts incorporating comprehensive TFBS information.

5.2.2 Data Sharing by Sladek Lab

Sladek group has been disseminating our extensive knowledge of HNF4α TFBSs by sharing it with public on the PAZAR database (http://www.pazar.info/) [10]. PAZAR is a public database aimed at competing with commercial TRANSFAC and has been slowly increasing its TFBS coverage. Additionally, we have been sharing our HNF4α knowledge of interacting proteins, binding sequences and splicing data on the Transcription Factor encyclopedia (TFe), also developed by Wasserman group (http://www.cisreg.ca). See Fig. 5.5 for snapshot of HNF4α page.



transcription factor encyclopedia BETA

Recently visited

Recently upd

Search TFe

Sign in or sign up.

Pages

About

HNF4A

hepatocyte nuclear factor 4, alpha

Classification *Homo sapiens* » Zinc-coordinatin nuclear Receptor Family » Subfar DHR78, TLL, COUP-TF and EAR2 individual genes)

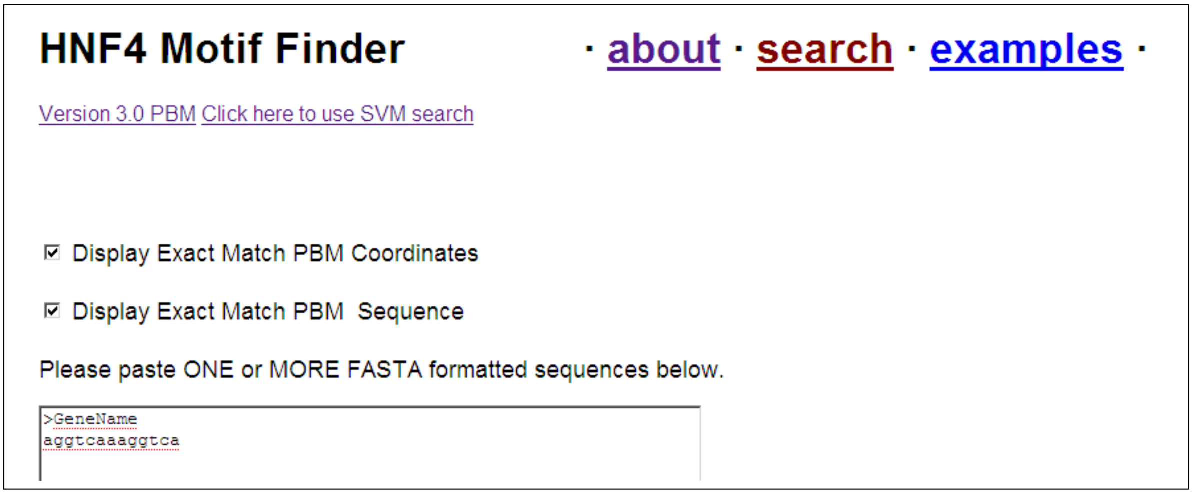
Related TF(s) • Hnf4a

Homologs (none defined here)

Links ENSG00000101076 (Ensembl) 3

Figure 5.5. Screenshot of HNF4α TFe mini-review (www.cisreg.ca)

We have published our prediction algorithms on the web with a easy interface (nrmotif.ucr.edu). One can paste in any DNA sequence to identify PBM experimentally confirmed or SVM predicted HNF4α binding sequence. See Fig. 5.6 for a screenshot of the interface.



HNF4 Motif Finder

· about · search · examples ·

Version 3.0 PBM Click here to use SVM search

☑ Display Exact Match PBM Coordinates

☑ Display Exact Match PBM Sequence

Please paste ONE or MORE FASTA formatted sequences below.

```
>GeneName
aggtcaaaggtca
```

Figure 5.6. Screenshot of HNF4α motif finder. (nrmotif.ucr.edu)

For the future, we plan to make all of the PBM data, network data and prediction models available in an easily accessible comprehensive format on the internet.

5.3 Novel Framework

The experimental framework utilized for TCF-1 and COUP-TF2 was similar to one used for HNF4α. First, a PBM experiment was conducted to determine exact binding sequences for a given TF. Following the PBM experiment, the promoters of all human genes are searched by exact match or SVM constructed from the PBM data. The genes containing the binding sequences are analyzed using GO over representation analysis in order to identify new functions for the TF. Our results show that this approach is applicable to any TF that can be expressed in sufficient quantities. The use of the flag epitope for TCF-1 also shows that an antibody to the native TF need not be available. Furthermore, since purification is not required and only a minimal processing via a simple centrifugation step is needed, these experiments are not labor intensive, and amendable to high throughput. The flexibility of the PBM is illustrated by the fact that HNFα designed PBM gave bound different sets of sequences for HNF4α, TCF-1, and COUP-TF2 confirming that the binding was specific to these TFs. Furthermore, these experiments revealed many known and novel categories of genes regulated by TCF-1, COUP-TF2 and HNF4α. Since there are still hundreds of TFs, and dozens of NRs for which the complete set of binding sequences is unknown, and the target genes have not been exhaustively explored, our approach could prove fruitful for those TFs.

5.4 Conclusion

Regulation of transcription is an amazingly complex process. However, in the last 20 years huge strides were made made to advance its understanding. Complete understanding of transcriptional regulation would mean not only complete understanding of every transcriptional event in every state of every cell type, it would also mean the same understanding for every cell in every disease state. Additionally, this understanding would enable accurate simulations of the transcriptional activity of any cell *in silico* over a period of time starting from any genomics, transcriptomic, or proteomic state of the cell. While this is certainly not possible with current technology, small steps are being taken in that direction by elucidation of pathways and networks for specific cell types. Molecular biology community has sketched out the details for many pathways, and we now know most of the components in some of the very well studied pathways such as Wnt signaling. With modern proteomic, microarray and sequencing technologies new pathways are being rapidly identified and roles for the hundreds of transcription factors in the genome are being discovered. While these technologies give a broad overview of transcriptional regulation and generate massive number of experimental leads, there will always be room for a traditional hypothesis driven molecular biology approach to answer difficult mechanistic questions. I believe that we will ultimately achieve the supreme goal of understanding all the transcriptional regulation in mammalian cells, and I hope that it will happen within my lifetime.

Bibliography

[1] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*, 2009.

[2] N. Brianon and M. C. Weiss. In vivo role of the HNF4alpha AF-1 activation domain revealed by exon swapping. *EMBO J*, 25(6):1253–1262, Mar 2006.

[3] P. Hatzis and I. Talianidis. Regulatory mechanisms controlling human hepatocyte nuclear factor 4alpha gene expression. *Mol Cell Biol*, 21(21):7320–7330, Nov 2001.

[4] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, May 2009.

[5] P. Lu, G. B. Rha, M. Melikishvili, G. Wu, B. C. Adkins, M. G. Fried, and Y. I. Chi. Structural basis of natural promoter recognition by a unique nuclear receptor, HNF4alpha. Diabetes gene product. *J Biol Chem*, 283:33685–97, 2008.

[6] D. Moras. Structure of full-length PPARgamma-RXRalpha: a snapshot of a functional complex? *Cell Metab*, 9(1):8–10, Jan 2009.

[7] D. E. Newburger and M. L. Bulyk. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res*, 37(Database issue):D77–D82, Jan 2009.

[8] D. T. Odom, R. D. Dowell, E. S. Jacobsen, L. Nekludova, P. A. Rolfe, T. W. Danford, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. Core transcriptional regulatory circuitry in human hepatocytes. *Mol Syst Biol*, 2:2006 0017, 2006.

[9] K. Pfeifer, T. Prezant, and L. Guarente. Yeast HAP1 activator binds to two upstream activation sites of different sequence. *Cell*, 49(1):19–27, Apr 1987.

[10] E. Portales-Casamar, D. Arenillas, J. Lim, M. I. Swanson, S. Jiang, A. McCallum, S. Kirov, and W. W. Wasserman. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res*, 37(Database issue):D54–D60, Jan 2009.

[11] M. J. Reijnen, F. M. Sladek, R. M. Bertina, and P. H. Reitsma. Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. *Proc Natl Acad Sci U S A*, 89:6300–3, 1992.

[12] B. Rhead, D. Karolchik, R. M. Kuhn, A. S. Hinrichs, A. S. Zweig, P. A. Fujita, M. Diekhans, K. E. Smith, K. R. Rosenbloom, B. J. Raney, A. Pohl, M. Pheasant, L. R. Meyer, K. Learned, F. Hsu, J. Hillman-Jackson, R. A. Harte, B. Giardine, T. R. Dreszer, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–D619, Jan 2010.

[13] M. J. Rieder, A. P. Reiner, B. F. Gage, D. A. Nickerson, C. S. Eby, H. L. McLeod, D. K. Blough, K. E. Thummel, D. L. Veenstra, and A. E. Rettie. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N Engl J Med*, 352(22):2285–2293, Jun 2005.

[14] K. R. Rosenbloom, T. R. Dreszer, M. Pheasant, G. P. Barber, L. R. Meyer, A. Pohl, B. J. Raney, T. Wang, A. S. Hinrichs, A. S. Zweig, P. A. Fujita, K. Learned, B. Rhead, K. E. Smith, R. M. Kuhn, D. Karolchik, D. Haussler, and W. J. Kent. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res*, 38(Database issue):D620–D625, Jan 2010.

[15] E. E. Schadt, C. Molony, E. Chudin, K. Hao, X. Yang, P. Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, J. Zhu, J. Millstein, S. Sieberts, J. Lamb, D. GuhaThakurta, J. Derry, J. D. Storey, I. Avila-Campillo, M. J. Kruger, J. M. Johnson, C. A. Rohl, A. van Nas, M. Mehrabian, T. A. Drake, A. J. Lusis, R. C. Smith, F. P. Guengerich, S. C. Strom, E. Schuetz, T. H. Rushmore, and R. Ulrich. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*, 6(5):e107, May 2008.

[16] P. L. Shaffer and D. T. Gewirth. Structural basis of VDR-DNA interactions on direct repeat response elements. *EMBO J*, 21(9):2242–2252, May 2002.

[17] B. S. Shastry. Snp alleles in human disease and evolution. *J Hum Genet*, 47(11):561–566, 2002.

[18] B. S. Shastry. Snps: impact on gene function and phenotype. *Methods Mol Biol*, 578:3–22, 2009.

[19] I. P. Torra, Y. Jamshidi, D. M. Flavell, J.-C. Fruchart, and B. Staels. Characterization of the human PPARalpha promoter: identification of a functional nuclear receptor response element. *Mol Endocrinol*, 16(5):1013–1028, May 2002.

[20] D. L. Veenstra, J. H. S. You, M. J. Rieder, F. M. Farin, H.-W. Wilkerson, D. K. Blough, G. Cheng, and A. E. Rettie. Association of Vitamin K epoxide reductase complex 1 (VKORC1) variants with warfarin dose in a Hong Kong Chinese patient population. *Pharmacogenet Genomics*, 15(10):687–691, Oct 2005.

[21] C. J. Willer, L. L. Bonnycastle, K. N. Conneely, W. L. Duren, A. U. Jackson, L. J. Scott, N. Narisu, P. S. Chines, A. Skol, H. M. Stringham, J. Petrie, M. R. Erdos, A. J. Swift, S. T. Enloe, A. G. Sprau, E. Smith, M. Tong, K. F. Doheny, E. W. Pugh, R. M. Watanabe, T. A. Buchanan, T. T. Valle, R. N. Bergman, J. Tuomilehto, K. L. Mohlke, F. S. Collins, and M. Boehnke. Screening of 134 single nucleotide polymorphisms (SNPs) previously associated with type 2 diabetes replicates association with 12 SNPs in nine genes. *Diabetes*, 56(1):256–264, Jan 2007.

[22] E. Wingender, P. Dietze, H. Karas, and R. Knuppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 24:238–41, 1996.

Chapter 6

Appendix

Table 6.1. Literature derived HNF4α binding sequences.

H4 No.	**Site Sequence**	**Symbol**	**Reference**	**PMID**
H4.1	GGGCCAaAGGTCT	*Hmgcs2*	Rodriguez et al. 1998	9464279
H4.2	GGTTCAaAGGTCT	*Acaa1*	Nicolas-Frances et al. 2000	10708554
H4.3	GGGGCTaAGTCCA	*SERPINA1*	Hardon et al. 1988	2844524
H4.4	AGTCAAaAGTCCA	*AMBP*	Rouet et al. 1995	7533900
H4.5	CTGCCAaGGGCCA	*AMBP*	Rouet et al. 1998	9729465
H4.6	GTCTAAgAGTCCA	*AMBP*	Rouet et al. 1998	9729465
H4.7	AGGACAaAGGTCA	*Acox*	Tugwood et al. 1992	1537328
H4.8	AGGTCAgGGTCCC	*ALDH2*	Pinaire et al. 1999	10352676
H4.9	GGGTCAaAGGCAC	*ALDH2*	Stewart et al. 1998	9765594
H4.10	AGGGCAgAGGGCA	*AGT*	Yanai et al. 1999	10574924
H4.11	GGGGCCaAGGTTC	*AGT*	Yanai et al. 1999	10574924
H4.12	AGGTCAaAGGCTG	*SERPINC1*	Tremp et al. 1995	7758957
H4.13	AGTGTAgAGCCCA	*SERPINC1*	Fernandez-Rachubinski et al. 1996	8910619

Table 6.1. Literature derived HNF4α binding sequences cont.

H4 No.	**Site Sequence**	**Symbol**	**Reference**	**PMID**
H4.14	AGTTCAaGGATCA	*APOA1*	Chan et al. 1993	8464705
H4.15	GGGGTCaAGGGTT	*APOA1*	Hardon et al. 1988	2844524
H4.16	AGGGTAaAGGTTG	*APOA2*	Ladias et al. 1992	1639815
H4.17	GTCACAaAAGTCC	*APOA4*	Ktistaki et al. 1994	7984419
H4.18	GGTCCAaAGGGCG	*APOB*	Metzger et al. 1993	8344962
H4.19	AGAACAaAGAGCA	*APOB*	Antes et al. 2000	10859308
H4.20	AGGCCAaAGTCCT	*APOC2*	Kardassis et al. 1998	9651383
H4.21	TGGGCAaAGGTCA	*APOC3*	Sladek et al. 1990	2279702
H4.22	AGTCCAgAGGTCA	*APOC3*	Vergnes et al. 1997	9366246
H4.23	GGTCCAgAGGGCA	*APOC3*	Kardassis et al. 1997	9012660
H4.24	TGATCAgACAAAG	*CEACAM1*	Hauck et al. 1994	8055923
H4.25	GAGTCAaAGGTCA	*Rbp2*	Nakshatri and Chambon 1994	8288643
H4.26	AGACCAaAGTCCG	*Cyp2A4*	Yokomori et al. 1997	9408084
H4.27	GGACCAaAGTCCA	*Cyp2C1*	Chen et al. 1994a	7772258
H4.28	GGTCCAaAGTCCA	*Cyp2C2*	Chen et al. 1994b	8106524
H4.29	AGACCAaAGTGCA	*Cyp2C3*	Chen et al. 1994a	7772258
H4.30	GGGTCAaAGTCCT	*Cyp2C9*	Ibeanu and Goldstein 1995	7794915
H4.31	AGGGCAaAGGCCA	*CYP2D6*	Cairns et al. 1996	8810289
H4.32	GTACCAaAGTCCA	*Cyp3A1*	Ogino et al. 1999	9917326
H4.33	TGGACTtAGTTCA	*CYP7A1*	Crestani et al. 1998	9799805
H4.34	AGGTCCaAGGGCA	*Cyp8b1*	del Castillo-Olivares and Gil 2001	11574686
H4.35	TGTCCAaAGTCCA	*AKR1C4*	Ozeki et al. 2001	11284743
H4.36	AGGTCGaGAGGTC	*EPO*	Galson et al. 1995	7891708
H4.37	CTAGCAaAGGTTA	*F9*	Naka and Brownlee 1996	8562402

Table 6.1. Literature derived HNF4α binding sequences cont.

H4 No.	**Site Sequence**	**Symbol**	**Reference**	**PMID**
H4.38	GTACCAaAGTACA	*F9*	Reijnen et al. 1992	1631121
H4.39	AGTGGTaAGGTCG	*F9*	Naka and Brownlee 1996	8562402
H4.40	CGGGCAaAGTTCT	*F7*	Arbini et al. 1997	8978290
H4.41	AGGGCAaAGGTCA	*F7*	Stauffer et al. 1998	9442072
H4.42	GGGGCAtAAGTCT	*F8*	Figueiredo and Brownlee 1995	7744832
H4.43	GGAGCAaAGTCCA	*F10*	Miao et al. 1992	1313796
H4.44	AAACCAaAGTTCA	*GUCY2C*	Swenson et al. 1999	10070050
H4.45	GGGGTAaAGGTTC		Garcia et al. 1993	8389913
H4.46	AGTCCAaGAGTCC		Guo et al. 1993	8417343
H4.47	AGGTTAaAGGTCT		Raney et al. 1997	8995626
H4.48	AGGTCAgGGTCCA	*MST1*	Waltz et al. 1996	8621550
H4.49	GGTCCAaAGTTCA	*HNF1a*	Zapp et al. 1993	8413240
H4.50	AGTCCAaAGTTCA	*TCF1*	Gragnoli et al. 1997	9313764
H4.51	GGGCTGaAGTCCA	*TCF1*	Sladek, unpublished	
H4.52	CGGGCAaAGGCCA	*Onecut1*	Lahuna et al. 2000	10674400
H4.53	GGGCCAaGGGTCA	HIV LTR	Ladias 1994	8119938
H4.54	AGTTCAaAGTTCA	*Fabp2*	Rottman and Gordon 1993	8505324
H4.55	GGGCCAgAGTCCA	*Pklr*	Diaz Guerra et al. 1993	8246989
H4.56	AGGTCTcAGGTCA	*MST1*	Ueda et al. 1998	9668124
H4.57	GGGTCAcAGTGCA	*MST1*	Ueda et al. 1998	9668124
H4.58	CGGGTAaAGGTGA	*ACADM*	Carter et al. 1993	8314750
H4.59	GGTTTAaAGTTCA	*Otc*	Nishiyori et al. 1994	8288597
H4.60	GGATCAaAGGTCC	*Otc*	Kimura et al. 1993	8496174
H4.61	AGTTCAgAGGTTA	*Otc*	Nishiyori et al. 1994	8288597

Table 6.1. Literature derived HNF4α binding sequences cont.

H4 No.	**Site Sequence**	**Symbol**	**Reference**	**PMID**
H4.62	GGCTTAaAGTTCA	*Otc*	Kimura et al. 1993	8496174
H4.63	CGGCCAaAGGTCA	*Pck1*	Hall et al. 1992	1333043
H4.64	GGGGCAaAGTCAA	*Prlr*	Moldrup et al. 1996	8776726
H4.65	GGGTTAaAGGTTG	*SHBG*	Janne and Hammond 1998	9852068
H4.66	GGGTCAaGGGTCA	*SHBG*	Janne and Hammond 1998	9852068
H4.67	AGGTCAaAGATTG	*Trf*	Schaeffer et al. 1993	8226864
H4.68	GGCAAGgTTCATA	*Ttr*	Sladek et al. 1990	2279702
H4.69	AGATCAaAGAGCA	*Tat*	Nitsch et al. 1993	8100067
H4.70	AGTCCAaAGGTCC	*WHVEnII*	Ueda et al. 1996	8676498
H4.71	GGAGTAaAGTTCA	*Aldob*	Gregori et al. 1998	9737987
H4.142	TGGGCAaAGGTCG	*GHR*	Jiang and Lucy 2001	11376119
H4.143	TGGGCAaAGAGCA	*GHR*	Jiang and Lucy 2001	11376119
H4.144	CTGAAGgGCTCAC	*MTTP*	Hagan et al. 1994	7961826
H4.145	TGGGAGgGCTGAC	*MTTP*	Hagan et al. 1994	7961826
H4.146	AGGTCAgAGACCT	*APOE*	Dang et al. 1995	7673250
H4.147	ATACCAaAGTTCA	*AFP*	Nakabayashi et al. 2004	15144905
H4.148	AGGACAaAGGCCA	*CYP4A6*	Muerhoff et al. 1992	1605646
H4.149	TGGGCAaGGGTCA	*CYP4A6*	Muerhoff et al. 1992	1605646
H4.150	TAGGCAaGAGGCA	*Cyp4A1*	Muerhoff et al. 1992	1605646
H4.151	GGGACAaAGTTCA	*HNF1*	McNair et al. 2000	11085951
H4.152	TGAGCAaAGTCTT	*Hnf4a*	Bailly et al. 2001	11522818
H4.153	AGACCTtTGAGTT	*PAX4*	Smith et al. 2000	10967107
H4.154	AGGGCAaGGTCCA	*CYP8B1*	Zhang and Chiang 2001	11535594
H4.155	TGGGCAaAGTCCT	*NR0B2*	Shih et al. 2001	11679424

Table 6.1. Literature derived HNF4α binding sequences cont.

H4 No.	**Site Sequence**	**Symbol**	**Reference**	**PMID**
H4.156	TGATTAaAGTCCA	*HP-25*	Kojima et al. 2000	10903495
H4.157	CGACCAaAGTCCA	*Cyp3a16*	Nakayama et al. 2001	11573935
H4.158	AGGTCAaGCTCCT	*FMO1*	Luo and Hines 2001	11723251
H4.159	AGGCTAaAGTACA	*FMO1*	Luo and Hines 2001	11723251
H4.160	AGATCAgACTCCT	*FMO1*	Luo and Hines 2001	11723251
H4.161	GGGGCAaAGTTCA	*PPARA*	Pineda Torra et al. 2002	11981036
H4.162	GGGACAaAGAGCA	*F2*	Ceelie et al. 2003	12911579
H4.163	ACGGCAaAGTCCA	*Ins1*	Bartoov-Shifman et al. 2002	11994285
H4.164	GGAACCaGGGCCA	*G6pc*	Rajas et al. 2002	11864989
H4.165	TGACCCcAGGTCC	*G6pc*	Rajas et al. 2002	11864989
H4.166	AGGTCAgGGGACA	*Nos2*	Guo et al. 2002	11741883
H4.167	CAATTAaAGGTCA	*CYP3A4*	Tirona et al. 2003	12514743
H4.168	AGTCCAaAGGTCA	*Hnf1b*	Power and Cereghini 1996	8622679
H4.169	AGGTCAaAGGTCA	synthetic	Jiang et al. 1997	9126270
H4.170	AGGTCAaAGGTTA	*Aldob*	Garrison et al.2006	16618389
H4.171	AGAGCAaAGGTGT	*Apoc2*	Garrison et al.2006	16618389
H4.172	AGGCCAgAGGTCA	*Crb3*	Garrison et al.2006	16618389
H4.173	AGGTCAgAGGACA	*Apoc2*	Garrison et al.2006	16618389
H4.174	AGAGCAaAGGTCT	*Aqp4*	Garrison et al.2006	16618389
H4.175	AGGTCAgAGGCCT	*Cldn2*	Garrison et al.2006	16618389
H4.176	AGGGCAaGGAGCA	*Gatm*	Garrison et al.2006	16618389
H4.177	GGTTCAaAGGGCA	*Mucdhl*	Garrison et al.2006	16618389
H4.178	CTAGCAaAGTCCA	*Neu3*	Garrison et al.2006	16618389
H4.179	GTACCAaAGGTCC	*Saa1*	Garrison et al.2006	16618389

Table 6.1. Literature derived HNF4α binding sequences cont.

H4 No.	Site Sequence	Symbol	Reference	PMID
H4.180	GGTCCAcAGTTCA	*Slc39a4*	Garrison et al.2006	16618389
H4.181	AGGGCTaGGGTCA	*Slc39a4*	Garrison et al.2006	16618389
H4.182	AGTCCAaAGGCCG	*Cdh1*	Battle et al. 2006	16714383
H4.183	AGATCAaAGTGCA	*Cxadr*	Battle et al. 2006	16714383
H4.184	AGATCAaTGTCCA	*Cxadr*	Battle et al. 2006	16714383
H4.185	AGTCCAaAGGTTC	*Gjb1*	Battle et al. 2006	16714383
H4.186	GGGTCAgAGGTCA	*Gpld1*	Battle et al. 2006	16714383
H4.187	TGGGCAaAGGTCT	*Lgals9*	Battle et al. 2006	16714383
H4.188	GGCTCAaAGTTCA	*Npnt*	Battle et al. 2006	16714383
H4.189	AGGTCAgAGGGCA	*Npnt*	Battle et al. 2006	16714383
H4.190	GGGTTAgAGTCCA	*Pkp2*	Battle et al. 2006	16714383
H4.191	GGTCCAgAGTTCA	*Rhpn2*	Battle et al. 2006	16714383
H4.192	AGGGCAaAGGTTT	*Vtn*	Battle et al. 2006	16714383
H4.193	AGTCCAaAGTCAA	*Vtn*	Battle et al. 2006	16714383
H4.194	AGGTTAaAGGTCA	*APOA5*	Prieur et al. 2003	16051671
H4.195	GGGACAaATTCCA	*UGT1A9*	O Barbier et al., 2005	15470081
H4.196	AGACCAaAGGACA	*CYP2C9*	Kawashima et al. 2006	16540586
H4.197	AGACCAaAGGGCA	*Kcnj11*	Gupta et al. 2005	15761495
H4.198	GTGGTAaAGGTCT	*Pck1*	Scribner et al. 2006	16713227
H4.199	AGGTCAAAAGTAC	*Acmsd*	Shin, Kimura, 2006	16807375
H4.200	AGGTCAaAGGCCT	*Fasn*	Adamson et al. 2006	16800817
H4.201	AGTTTGgAGTCTG	*MTTP*	Hirokane et al. 2004	15337761
H4.202	TGGAACtGGGTCA	*G6pc*	Rajas et al. 2002	11864989
H4.203	AGGGCAaAGACCT	*Nr1i3*	Ding et al. 2006	16825189

Table 6.1. Literature derived HNF4α binding sequences cont.

H4 No.	Site Sequence	Symbol	Reference	PMID
H4.204	TGTGCAaGGTTCA	*Ces2*	Furihata et al. 2006	16527247
H4.205	TGTCACAAGGTCA	*Gck*	Roth et al. 2002	11950391
H4.206	AGGACAGAGTCCA	*Slc27a5*	Inoue et al. 2003	14583614
H4.207	AGTTCCAAGGTCT	*Baat*	Inoue et al. 2003	14583614
H4.208	TGGGCAaGGGGCA	*Prodh2*	Kamiya et al. 2004	15581617
H4.209	GGGACAgAGGTCA	*Prodh2*	Kamiya et al. 2004	15581617
H4.210	AGTGCAAGGGTCT	*G6PC*	Rhee et al 2003	12651943
H4.211	AGGACAGAGTCTA	*G6PC*	Hirota et al 2005	15702243
H4.212	TGAACTtAGGTCC	*MTTP*	Sheena et al. 2005	15547294
H4.213	TGGGGAaAGGTCA	*MTTP*	Sheena et al. 2005	15547294
H4.214	TGATTAaAGTTCA	*PAK7*	Niehof and Borlak, 2005	15615695
H4.215	GGGTACaATGTTC	*PAK7*	Niehof and Borlak, 2005	15615695
H4.216	CTGACTaAGGTAC	*PAK7*	Niehof and Borlak, 2005	15615695
H4.217	AGGGCAtAGGTCA	*RSK4*	Niehof and Borlak, 2005	15615695
DR2.1	CGCTCAAAAGGTTG	*RSK4*	Niehof and Borlak, 2005	15615695
DR2.2	AGCCCTatTGACCC	*SLC22A1*	Saborowski et al 2005	16436500
DR2.3	TGATCTctTGTCCT	*SLC22A1*	Saborowski et al 2005	16436500

Table 6.2. HNF4a motifs H4.72 to H4.141.S1B in nrmotif.ucr.edu

H4 No.	Sequence	Reference	Pubmed id
H4.72	GGGGCAaAGTCCA	Used in Ellrott et al. 2002	12385991
H4.73	GGGTCAaAGGTCA	Used in Ellrott et al. 2002	12385991
H4.74	GGGCCAaAGTCCA	Used in Ellrott et al. 2002	12385991
H4.75	AGGCCAaAGTTCA	Used in Ellrott et al. 2002	12385991

Table 6.2. HNF4a motifs H4.72 to H4.141 cont.(S1B)

H4 No.	Sequence	Reference	Pubmed id
H4.76	GGAGCAaAGTTCA	Used in Ellrott et al. 2002	12385991
H4.77	AGGTCAaAGGGCA	Used in Ellrott et al. 2002	12385991
H4.78	GGTTTAaAGTTCA	Used in Ellrott et al. 2002	12385991
H4.79	AGTTCAaAGAGCA	Used in Ellrott et al. 2002	12385991
H4.80	GGGTCAaGGGTCG	Used in Ellrott et al. 2002	12385991
H4.81	AGGTCAgAGGTCA	Used in Ellrott et al. 2002	12385991
H4.82	AGGCCAaAGGGCA	Used in Ellrott et al. 2002	12385991
H4.83	AGAGCAaAGGTCA	Used in Ellrott et al. 2002	12385991
H4.84	AGTCCAaAAGTCC	Used in Ellrott et al. 2002	12385991
H4.85	AGTTCAaAGTGCA	Used in Ellrott et al. 2002	12385991
H4.86	GGGTCAaAGGCCA	Used in Ellrott et al. 2002	12385991
H4.87	GGCTCAaAGTCCA	Used in Ellrott et al. 2002	12385991
H4.88	AGTCCAaAGTTTA	Used in Ellrott et al. 2002	12385991
H4.89	AGGTCAaAGGCCA	Used in Ellrott et al. 2002	12385991
H4.90	AGTCCAaAGGCCA	Used in Ellrott et al. 2002	12385991
H4.91	CGAGCAaAGGCCA	Used in Ellrott et al. 2002	12385991
H4.92	AGTTCAgAGGTCT	Used in Ellrott et al. 2002	12385991
H4.93	AGCTTAaAGTTCA	Used in Ellrott et al. 2002	12385991
H4.94	AGGTCAgAGTTCA	Used in Ellrott et al. 2002	12385991
H4.95	AGTTCAcAGTTCA	Used in Ellrott et al. 2002	12385991
H4.96	GGTTTAaAGGTTT	Used in Ellrott et al. 2002	12385991
H4.97	AGGCCAaGGTTCA	Used in Ellrott et al. 2002	12385991
H4.97	AGGCCAaGGTTCA	Used in Ellrott et al. 2002	12385991
H4.98	AGGGCAaAAGCCA	Used in Ellrott et al. 2002	12385991

Table 6.2. HNF4a motifs H4.72 to H4.141 cont.(S1B)

H4 No.	Sequence	Reference	Pubmed id
H4.99	AGGTCTaAGGTCA	Used in Ellrott et al. 2002	12385991
H4.100	AGCTCAaGGTTCA	Used in Ellrott et al. 2002	12385991
H4.101	GGGTTCaAGGTTA	Used in Ellrott et al. 2002	12385991
H4.102	AGGTCAgAGGTTA	Used in Ellrott et al. 2002	12385991
H4.103	AGGTTAaAGATCG	Used in Ellrott et al. 2002	12385991
H4.104	AGGTTCaAGGTTA	Used in Ellrott et al. 2002	12385991
H4.105	GGGGCAaAGTTTA	Used in Ellrott et al. 2002	12385991
H4.106	AGGGAAaAGGTCA	Used in Ellrott et al. 2002	12385991
H4.107	GGGCTAaAGTTTA	Used in Ellrott et al. 2002	12385991
H4.108	AGTCCAaGGGTTC	Used in Ellrott et al. 2002	12385991
H4.109	AGGCTAaAGTGCA	Used in Ellrott et al. 2002	12385991
H4.110	GAGTCAaGGGTCA	Used in Ellrott et al. 2002	12385991
H4.111	AGGTCAgGGGTCA	Used in Ellrott et al. 2002	12385991
H4.112	GGTCCAaAGGTTA	Used in Ellrott et al. 2002	12385991
H4.113	AAAGCAaAGGTCA	Used in Ellrott et al. 2002	12385991
H4.114	GGGCCAgGGTCCA	Used in Ellrott et al. 2002	12385991
H4.115	AGGTCAaAATGCA	Used in Ellrott et al. 2002	12385991
H4.116	GGGCCAaTGTTCA	Used in Ellrott et al. 2002	12385991
H4.117	GAGGCAaAAGTCC	Used in Ellrott et al. 2002	12385991
H4.118	AGGCAAaAGTCCA	Used in Ellrott et al. 2002	12385991
H4.119	GGTCAAaAGGGCA	Used in Ellrott et al. 2002	12385991
H4.120	AGGGCAgAGGCCA	Used in Ellrott et al. 2002	12385991
H4.121	AGACCAaAATCCG	Used in Ellrott et al. 2002	12385991
H4.122	AGAATAaAGATTA	Used in Ellrott et al. 2002	12385991

Table 6.2. HNF4a motifs H4.72 to H4.141 cont.(S1B)

H4 No.	Sequence	Reference	Pubmed id
H4.123	AGGTCAgGGGCCA	Used in Ellrott et al. 2002	12385991
H4.124	GGGTCAgGGGCCC	Used in Ellrott et al. 2002	12385991
H4.125	GGGTCAaGAGTCG	Used in Ellrott et al. 2002	12385991
H4.126	GGGTAGaTTGTTG	Used in Ellrott et al. 2002	12385991
H4.127	AGGTCAaGGGTTT	Used in Ellrott et al. 2002	12385991
H4.128	AGTTCAaGGGTTT	Used in Ellrott et al. 2002	12385991
H4.129	GGGCTGaAGGGCA	Used in Ellrott et al. 2002	12385991
H4.130	AGAGCAaTGGTCA	Used in Ellrott et al. 2002	12385991
H4.131	GGGTCAgTGGCCA	Used in Ellrott et al. 2002	12385991
H4.132	AGGTCAaGGGCTG	Used in Ellrott et al. 2002	12385991
H4.133	GGGTCAgAGGCCA	Used in Ellrott et al. 2002	12385991
H4.134	AGGGTAaAGGCGA	Used in Ellrott et al. 2002	12385991
H4.135	AGGGCAgAAGCCA	Used in Ellrott et al. 2002	12385991
H4.136	AGAGCAaAGGCAA	Used in Ellrott et al. 2002	12385991
H4.137	AGTTCTaAGGTCT	Used in Ellrott et al. 2002	12385991
H4.138	GGGGAAaAGGTCC	Used in Ellrott et al. 2002	12385991
H4.139	AGTCCAgAATTCA	Used in Ellrott et al. 2002	12385991
H4.140	GGATGTaAGGTCC	Used in Ellrott et al. 2002	12385991
H4.141	AGGCTGaAGTGCA	Used in Ellrott et al. 2002	12385991

Table 6.3. First 27/3000 probe sequences from PBM2 array design. Partial table, for complete table see nrmotif.ucr.edu. S2A in nrmotif.ucr.edu

ID	Descriptor	Linker	Variable	Cap
PBM1.1	agct	TCGACCGATACTCTAATCTCCCTAGGC	AGCTAGCTAGCTA	GCGCG
PBM1.2	at's	TCGACCGATACTCTAATCTCCCTAGGC	ATATATATATATA	GCGCG
PBM1.3	DR0	TCGACCGATACTCTAATCTCCCTAGGC	AGGTCAAGGTCA	GCGCG
PBM1.4	Dr2	TCGACCGATACTCTAATCTCCCTAGGC	AGGTCAAGAGGTCA	GCGCG
PBM1.5	DR2classic	TCGACCGATACTCTAATCTCCCTAGGC	AGGTCAAAAGGTCA	GCGCG
PBM1.6	DR3	TCGACCGATACTCTAATCTCCCTAGGC	AGTTCAAAAAGGTCA	GCGCG
PBM1.7	DR4	TCGACCGATACTCTAATCTCCCTAGGC	AGTTCAAAAAAGGTCA	GCGCG
PBM1.8	DR5	TCGACCGATACTCTAATCTCCCTAGGC	AGTTCAAAAAAAGGTCA	GCGCG
PBM1.9	EBace1	TCGACCGATACTCTAATCTCCCTAGGC	GAGCAAAGTCCA	GCGCG
PBM1.10	EBace10	TCGACCGATACTCTAATCTCCCTAGGC	GAGCAGAGGACA	GCGCG
PBM1.11	EBace100	TCGACCGATACTCTAATCTCCCTAGGC	GGGGCCAAGGGCA	GCGCG
PBM1.12	EBace101	TCGACCGATACTCTAATCTCCCTAGGC	GGGGGAAGGGACA	GCGCG
PBM1.13	EBace102	TCGACCGATACTCTAATCTCCCTAGGC	GGGGGCAGGGGCA	GCGCG
PBM1.14	EBace103	TCGACCGATACTCTAATCTCCCTAGGC	GGGTCAAAGGTCA	GCGCG
PBM1.15	EBace104	TCGACCGATACTCTAATCTCCCTAGGC	GGGTCAAGGGTCA	GCGCG
PBM1.16	EBace105	TCGACCGATACTCTAATCTCCCTAGGC	GGACCCAAGGCCA	GCGCG
PBM1.17	EBace106	TCGACCGATACTCTAATCTCCCTAGGC	AGAACAAAGGTCA	GCGCG
PBM1.18	EBace11	TCGACCGATACTCTAATCTCCCTAGGC	GAGCAGAGGCCA	GCGCG
PBM1.19	EBace12	TCGACCGATACTCTAATCTCCCTAGGC	GGTCAAAGGGCA	GCGCG
PBM1.20	EBace13	TCGACCGATACTCTAATCTCCCTAGGC	GATCAAAGTGCA	GCGCG
PBM1.21	EBace14	TCGACCGATACTCTAATCTCCCTAGGC	GGTCAAAGTCCA	GCGCG
PBM1.22	EBace15	TCGACCGATACTCTAATCTCCCTAGGC	GGTCAAAGTTCA	GCGCG
PBM1.23	EBace16	TCGACCGATACTCTAATCTCCCTAGGC	GGTCAGAGGGCA	GCGCG
PBM1.24	EBace17	TCGACCGATACTCTAATCTCCCTAGGC	GATCAGAGTCCA	GCGCG
PBM1.25	EBace18	TCGACCGATACTCTAATCTCCCTAGGC	GGGCAGAGGTCA	GCGCG
PBM1.26	EBace19	TCGACCGATACTCTAATCTCCCTAGGC	GGGCAGAGTTCA	GCGCG
PBM1.27	EBace2	TCGACCGATACTCTAATCTCCCTAGGC	GGTCAAAGGTCA	GCGCG

Table 6.4. First 27/3000 probe sequences from PBM2 array design. Partial table, for complete table see nrmotif.ucr.edu.(S2b)

ID	Descriptor	Linker	Variable	Cap
PBM2.1	H4.1	TCGACCGATACTCTAATCTCCCTAGGC	GGGCCAAAGGTCT	GCGCG
PBM2.2	H4.2	TCGACCGATACTCTAATCTCCCTAGGC	GGTTCAAAGGTCT	GCGCG
PBM2.3	H4.3	TCGACCGATACTCTAATCTCCCTAGGC	GGGGCTAAGTCCA	GCGCG
PBM2.4	H4.4	TCGACCGATACTCTAATCTCCCTAGGC	AGTCAAAAGTCCA	GCGCG
PBM2.5	H4.5	TCGACCGATACTCTAATCTCCCTAGGC	CTGCCAAGGGCCA	GCGCG
PBM2.6	H4.6	TCGACCGATACTCTAATCTCCCTAGGC	GTCTAAGAGTCCA	GCGCG
PBM2.7	H4.7	TCGACCGATACTCTAATCTCCCTAGGC	AGGACAAAGGTCA	GCGCG
PBM2.8	H4.8	TCGACCGATACTCTAATCTCCCTAGGC	AGGTCAGGGTCCC	GCGCG
PBM2.9	H4.9	TCGACCGATACTCTAATCTCCCTAGGC	GGGTCAAAGGCAC	GCGCG
PBM2.10	H4.10	TCGACCGATACTCTAATCTCCCTAGGC	AGGGCAGAGGGCA	GCGCG
PBM2.11	H4.11	TCGACCGATACTCTAATCTCCCTAGGC	GGGGCCAAGGTTC	GCGCG
PBM2.12	H4.12	TCGACCGATACTCTAATCTCCCTAGGC	AGGTCAAAGGCTG	GCGCG
PBM2.13	H4.13	TCGACCGATACTCTAATCTCCCTAGGC	AGTGTAGAGCCCA	GCGCG
PBM2.14	H4.14	TCGACCGATACTCTAATCTCCCTAGGC	AGTTCAAGGATCA	GCGCG
PBM2.15	H4.15	TCGACCGATACTCTAATCTCCCTAGGC	GGGGTCAAGGGTT	GCGCG
PBM2.16	H4.16	TCGACCGATACTCTAATCTCCCTAGGC	AGGGTAAAGGTTG	GCGCG
PBM2.17	H4.17	TCGACCGATACTCTAATCTCCCTAGGC	GTCACAAAAGTCC	GCGCG
PBM2.18	H4.18	TCGACCGATACTCTAATCTCCCTAGGC	GGTCCAAAGGGCG	GCGCG
PBM2.19	H4.19	TCGACCGATACTCTAATCTCCCTAGGC	AGAACAAAGAGCA	GCGCG
PBM2.20	H4.20	TCGACCGATACTCTAATCTCCCTAGGC	AGGCCAAAGTCCT	GCGCG
PBM2.21	H4.21	TCGACCGATACTCTAATCTCCCTAGGC	TGGGCAAAGGTCA	GCGCG
PBM2.22	H4.22	TCGACCGATACTCTAATCTCCCTAGGC	AGTCCAGAGGTCA	GCGCG
PBM2.23	H4.23	TCGACCGATACTCTAATCTCCCTAGGC	GGTCCAGAGGGCA	GCGCG
PBM2.24	H4.24	TCGACCGATACTCTAATCTCCCTAGGC	TGATCAGACAAAG	GCGCG
PBM2.25	H4.25	TCGACCGATACTCTAATCTCCCTAGGC	GAGTCAAAGGTCA	GCGCG
PBM2.26	H4.26	TCGACCGATACTCTAATCTCCCTAGGC	AGACCAAAGTCCG	GCGCG
PBM2.27	H4.27	TCGACCGATACTCTAATCTCCCTAGGC	GGACCAAAGTCCA	GCGCG

Table 6.5. Expression profiling of HNF4a RNAi in HepG2 cells using Affymetrix HGU 133 plus 2.0 arrays. Partial table, for complete table see nrmotif.ucr.edu. (S3A)

HNF4(+)	HNF4(+)	HNF4(-)	HNF4(-)	Log Fc	Gene Symbol	Entrez Gene	P Value
8.55618137	8.55885719	4.7081182	5.155934	-3.625493043	*NINJ1*	4814	0.138
7.96193561	4.88032065	3.4895199	2.213146	-3.569794933	*C12orf46*	121506	0.77
7.66880941	8.11345193	4.4767823	5.131102	-3.087188446	*PCGF5*	84333	0.20
7.47935957	7.25302111	3.7227658	4.881787	-3.063913975	*DIO1*	1733	0.30
6.66152686	6.35428497	3.3051317	3.844362	-2.933159146	*AKR1D1*	6718	0.17
5.45876373	5.8945928	2.4879498	3.014599	-2.925403718	*DNASE1L1*	1774	0.19
5.85109357	4.87972419	2.8715479	2.301142	-2.779064083	*SOAT2*	8435	0.31
8.71349127	9.00157815	6.0587561	6.357091	-2.649611121	—		0.15
7.97648292	8.10145482	5.1814388	5.746395	-2.575052054	*PCGF5*	84333	0.18
6.29939995	6.11562829	4.3484988	3.096073	-2.485228383	—		0.39
6.59519217	5.95780443	3.4411485	4.163994	-2.47392682	*TRIM4*	89122	0.30
6.49504343	5.98818357	4.1642998	3.446417	-2.436254922	—		0.28
7.73397385	8.50204087	5.255261	6.148113	-2.416320476	*LOC284422*	284422	0.37
8.22890826	7.37880179	6.1422982	4.701191	-2.382110534	*CDH1*	999	0.56
6.78938191	6.67945999	4.382235	4.41202	-2.337293505	*DGAT1*		0.13
10.6155075	10.8018987	8.4071931	8.336735	-2.336738995	*HN1*	51155	0.13
9.08197448	9.31902349	6.5793382	7.188213	-2.316723542	*CLU*	1191	0.22
6.63090296	7.26675122	4.0980562	5.228563	-2.285517553	*HKDC1*	80201	0.44
4.91471788	5.64216451	3.1831968	2.821969	-2.275858472	*SLC7A9*	11136	0.27
6.00251085	5.70276483	3.6924988	3.463007	-2.274884789	*CYP3A5*	1577	0.15
9.40022627	9.45695805	7.0863205	7.229227	-2.270818602	*HYAL1*	3373	0.13
10.6265566	10.4601034	8.4949198	8.098633	-2.246553705	*CDC42*	998	0.16
7.59103942	7.47390702	5.1149398	5.460229	-2.24488872	*FAM62A*	23344	0.15
6.57485752	6.29433088	4.3213953	4.091838	-2.227977717	*IVD*	3712	0.15

Table 6.6. Primers used for RT-PCR in HNF4a RNAi in HepG2 cells.

Gene Symbol	Accession no.	Sequence 5'-3'	Location
HNF4A	NM_178849	F' ACATGGACATGGCCGACTACA	115-135
		R' AGCTCGCAGAAAGCTGGGAT	721-702
ACTB	X00351	F' CGTACCACTGGCATCGTGAT	480-499
		R' GTGTTGGCGTACAGGTCTTTG	931-912
APOC3	NM_000040	F' TACTCCTTGTTGTTGCCCTCC	60-80
		R' CACGGCTGAAGTTTGGTCTGA	337-318
APOA4	NM_000482	F' CCTACGCTGACGAATTCAAA	722-741
		R' AAGCTCAAGTGGCCTTCCA	1139-1121
APOA2	NM_001643	F' TCGCAGCAACTGTGCTACTC	69-88
		R' AGGCTGTGTTCCAAGTTCCA	349-330
CYP2D6	X08006	F' CTAAGGGAACGACACTCATCAC	1169-1190
		R' CTCACCAGGAAAGCAAAGACAC	1457-1436
CYP7A1	NM_000780	F' GCTGTGCTCTGCAATTTGGT	192-211
		R' CATCACTCGGTAGCAGAAAGA	601-581
APOC4	NM_001646	F' AGATGAGTCGCTGGAGCCT	168-186
		R' TGTCCCCACAGACAAGCCT	410-392
RDH16	NM_003708	F' TGACTCTGGCTTCGGGAAA	963-981
		R' GCTTGGAGAGACCCAGTACAT	1788-1768
APOM	NM_019101	F' ATTTGGGCAGCTCTGCTCTA	86-105
		R' TTATTGGACAGCTCACAGGC	636-617
APOH	NM_000042	F' ATGTTGCTATTGCAGGACGG	77-96
		R' CATCGCATGTTGTGGCAAAC	573-554
SPSB2	NM_032641	F' CTGTACCCTGACCTCTCCTGT	194-214
		R' TGAGTTCCCGCTGGATACTG	609-590
UBD	NM_006398	F' TGCAGGACCAGGTTCTTTTG	358-377
		R' TGCCAGGAAGAGTAAGTTGC	701-682
ZDHHC11	NM_024786	F' CACCCCAGAAGCCATACTCA	417-436
		R' GGCATGGGCTGAGAATAGTT	716-697

Table 6.7. Primers used for PCR in HNF4a ChIP on human NINJ1 gene in HepG2 cells.

Primer Set	Direction	Sequence 5’-3’	Location
1	forward	F’ TGGGTAAACAGCATTGAGCA	-2030
	reverse	R’ AGCTGGGACTACAGGTGTGC	-1599
2	forward	F’ AGCTTGCAGTGAGCCAAGAT	-1553
	reverse	R’ AATCCAGCACCATTCACACA	-1166
3	forward	F’ TGTGTGAATGGTGCTGGATT	-1185
	reverse	R’ TATTTCCAGAAGGGCAGTGG	-826
4	forward	F’ CCACTGCCCTTCTGGAAATA	-845
	reverse	R’ GCCCCTAGTAACAGCGTCAG	-394
5	forward	F’ CAGCAGTCTGTGCCCTCATA	-522
	reverse	R’ CTGGAGGCTGTACGCTGAG	-12

Table 6.8. Training set for SVM1 from PBM1. Partial table, for complete table see nrmotif.ucr.edu (S4A).

Descriptor	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	Binding Score Array 1	StDevArray 1
Msb439	G	G	G	G	C	A	A	A	G	T	C	C	A	9.483528411	1.425387622
EBbiopros31	G	G	G	T	C	A	A	A	G	T	C	C	A	10.77488137	1.388627127
EBbiopros28	G	G	G	T	C	A	A	A	G	G	T	C	A	7.544594734	1.755858692
Msb716	G	G	G	T	T	A	A	A	G	G	T	C	A	8.685004918	0.584401886
EBbiopros49	G	G	G	T	C	A	A	A	G	G	T	C	A	7.469879874	0.477724632
EBbiopros112	G	G	G	G	C	A	A	A	G	T	C	C	A	7.202565323	0.529150019
EBbiopros53	G	G	G	G	C	A	A	A	G	T	C	C	A	6.66263406	0.714311543
EBbiopros5	G	G	A	C	C	A	A	A	G	T	C	C	A	6.489077779	0.713803
rada16	G	G	G	T	C	A	A	A	G	T	C	C	A	6.184556806	0.247849744
EBbiopros25	G	G	G	T	C	A	A	A	G	T	T	C	A	5.217810454	0.674310861
EBace103	G	G	G	T	C	A	A	A	G	G	T	C	A	4.833228272	1.329329195
lit88	G	G	G	T	C	A	A	A	G	G	T	C	G	4.708471423	0.844929992
lit62	G	G	G	T	C	A	A	A	G	G	T	C	C	5.744625435	0.752816803
EBbiopros109	G	G	T	T	C	A	A	A	G	G	T	C	A	5.286823521	0.358419944
EBbiopros78	G	G	T	T	C	A	A	A	G	G	T	C	A	5.34068094	0.436729304
Msb184	G	G	G	G	C	A	G	A	G	T	C	C	A	5.515151745	0.828418479
Msb53	G	G	G	T	C	A	A	A	G	T	A	C	C	4.996437677	0.328968185
EBbiopros2	G	G	A	G	C	A	A	A	G	T	C	C	A	4.512967215	0.415998413
H4.28	G	G	T	C	C	A	A	A	G	T	C	C	A	4.70623637	0.796707101
Msb33	G	G	A	T	C	A	A	A	G	G	T	C	A	4.678171941	0.454847291
EBbiopros81	G	G	A	G	C	A	A	A	G	T	C	C	A	4.271126445	0.361652303
Msb709	G	G	G	G	C	A	A	A	G	G	T	G	A	4.497792664	0.402984467
H4.72	G	G	G	G	C	A	A	A	G	T	C	C	A	4.012929001	0.352565193
EBbiopros48	G	G	T	T	C	A	A	A	G	G	T	C	A	3.7212593	0.162980617

Table 6.9. Training set for SVM2 from PBM2. Partial table, for comlete table see nrmotif.ucr.edu (S4B)

Descriptor	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	AveArray 1
EBbiopros31.top	G	G	G	T	C	A	A	A	G	T	C	C	A	
H4.73	G	G	G	T	C	A	A	A	G	G	T	C	A	
EBvarythree522.top	A	G	G	T	C	A	A	A	G	T	C	C	A	15.57142225
H4.28	G	G	T	C	C	A	A	A	G	T	C	C	A	
lit88.top	G	G	G	T	C	A	A	A	G	G	T	C	G	14.84151073
rada21.top	A	G	T	C	C	A	A	A	G	T	C	C	A	13.43317691
H4.169	A	G	G	T	C	A	A	A	G	G	T	C	A	
Ebnonadj145.top	A	G	G	T	C	A	A	A	G	G	T	C	G	11.15754263
H4.72	G	G	G	G	C	A	A	A	G	T	C	C	A	11.29340131
EBbiopros21.top	A	G	G	G	C	A	A	A	G	T	C	C	A	9.921637286
Classification1.4promoters.SVM.271	A	G	T	C	C	A	A	A	G	T	C	C	G	10.48746299
H4.168	A	G	T	C	C	A	A	A	G	G	T	C	A	11.34211144
EBbiopros25.top	G	G	G	T	C	A	A	A	G	T	T	C	A	7.986283769
Ebnonadj11.top	T	G	G	T	C	A	A	A	G	G	T	C	A	10.2690305
rada26.top	G	G	T	T	C	A	A	A	G	T	C	C	A	9.424396299
EBbiopros109.top	G	G	T	T	C	A	A	A	G	G	T	C	A	9.520631574
Classification1.4promoters.SVM.13	G	G	T	A	C	A	A	A	G	T	C	C	A	
ChipchipMSB.SVM.134	A	G	T	T	C	A	A	A	G	T	C	C	A	8.441423611
EBace14.top	C	G	G	T	C	A	A	A	G	T	C	C	A	8.213686111
Ebnonadj111.top	A	G	G	T	C	A	A	A	G	T	T	C	A	8.874146132
lit91.top	C	G	G	T	C	A	A	A	G	G	T	C	G	8.862875429
H4.41	A	G	G	G	C	A	A	A	G	G	T	C	A	8.410922709
Ebnonadj7.top	C	G	G	T	C	A	A	A	G	G	T	C	A	
Ebnonadj1.top	G	G	G	G	C	A	A	A	G	G	T	C	A	9.209047512

Table 6.10. Gene Ontology overrepresented categories of biological processes and scores (p-values) in HNF4a targets identified by PBM2 search, RNAi expression profiling, ChIP-chip or SVM2 (¿3 H4 motifs). Partial table, for full table see. nrmotif.ucr.edu (S5).

Go Term	PBM search	RNAi expression	ChIP-chip	SVM search 4+
GO:0006810 transport	3.25313E-08	4.80177E-05	0.000108705	0.000155349
GO:0051179 localization	8.63387E-08	0.000976066	0.00047885	6.13317E-05
GO:0065007 biological regulation	1.14047E-07		0.008724394	0.013499018
GO:0051234 establishment of localization	3.053E-07	3.5952E-05	7.39424E-05	0.000179856
GO:0048856 anatomical structure development	4.30571E-07			
GO:0032502 developmental process	6.91675E-07			
GO:0050789 regulation of biological process	1.0031E-06		0.024233044	0.028202566
GO:0009966 regulation of signal transduction	1.65436E-06	0.019449233		0.044035903
GO:0044255 cellular lipid metabolic process	1.83559E-06	1.09225E-06	2.687E-19	3.65438E-08
GO:0009653 anatomical structure morphogenesis	7.86613E-06			0.062432781
GO:0015674 di-, tri-valent inorganic cation transport	9.77152E-06			
GO:0048731 system development	1.08239E-05			
GO:0007275 multicellular organismal development	1.34647E-05			
GO:0006886 intracellular protein transport	1.78969E-05	0.005064189	2.29026E-09	0.099366859
GO:0050794 regulation of cellular process	1.79493E-05		0.038239663	
GO:0030154 cell differentiation	2.29227E-05			
GO:0048869 cellular developmental process	2.29227E-05			
GO:0016043 cellular component organization and biogenesis	2.46689E-05	0.018828036	2.15143E-12	
GO:0006629 lipid metabolic process	2.60666E-05	1.87787E-06	7.9439E-20	2.17359E-08
GO:0030001 metal ion transport	4.43641E-05			0.087147259
GO:0046907 intracellular transport	8.91617E-05	0.00740417	7.90286E-10	
GO:0009987 cellular process	0.00011019		1.86507E-06	
GO:0006066 alcohol metabolic process	0.000135653	4.2616E-05	3.0278E-11	0.000731386
GO:0006605 protein targeting	0.000175142	0.096603801	0.000100957	0.024625829

Table 6.11. 198 human genes in the intersection of the PBM2, ChIP-chip and expression profiing of HNF4a in hepatic cells. Partial table, for a complete table see nrmotif.ucr.edu .

Gene ID	Gene Symbol	Gene name
368	*ABCC6*	ATP-binding cassette, sub-family C (CFTR/MRP), member 6
30	*ACAA1*	acetyl-Coenzyme A acyltransferase 1
8309	*ACOX2*	acyl-Coenzyme A oxidase 2, branched chain
183	*AGT*	angiotensinogen (serpin peptidase inhibitor, clade A, member 8)
51390	*AIG1*	androgen-induced 1
202	*AIM1*	absent in melanoma 1
8165	*AKAP1*	A kinase (PRKA) anchor protein 1
64400	*AKTIP*	AKT interacting protein
214	*ALCAM*	activated leukocyte cell adhesion molecule
10840	*ALDH1L1*	aldehyde dehydrogenase 1 family, member L1
29123	*ANKRD11*	ankyrin repeat domain 11
290	*ANPEP*	alanyl (membrane) aminopeptidase
307	*ANXA4*	annexin A4
116519	*APOA5*	apolipoprotein A-V
338	*APOB*	apolipoprotein B (including Ag(x) antigen)
345	*APOC3*	apolipoprotein C-III
55937	*APOM*	apolipoprotein M
427	*ASAH1*	N-acylsphingosine amidohydrolase (acid ceramidase) 1
445	*ASS1*	argininosuccinate synthetase 1
23130	*ATG2A*	ATG2 autophagy related 2 homolog A (S. cerevisiae)
93974	*ATPIF1*	ATPase inhibitory factor 1
622	*BDH1*	3-hydroxybutyrate dehydrogenase, type 1
91272	*BOD1*	biorientation of chromosomes in cell division 1
9577	*BRE*	brain and reproductive organ-expressed (TNFRSF1A modulator)
119504	*C10orf104*	chromosome 10 open reading frame 104
282969	*C10orf125*	chromosome 10 open reading frame 125
80017	*C14orf159*	chromosome 14 open reading frame 159
79762	*C1orf115*	chromosome 1 open reading frame 115
83606	*C22orf13*	chromosome 22 open reading frame 13
720	*C4A*	complement component 4A (Rodgers blood group)
722	*C4BPA*	complement component 4 binding protein, alpha
84273	*C4orf14*	chromosome 4 open reading frame 14

Table 6.12. 135 human genes in the intersection of the SVM2 (¿3 H4 motifs), ChIP-chip and expression profiing of HNF4a in hepatic cells. Partial table, for a complete table see nrmotif.ucr.edu.

Gene Id	Gene Symbol	Gene Name
84836	*ABHD14B*	abhydrolase domain containing 14B
95	*ACY1*	aminoacylase 1
202	*AIM1*	absent in melanoma 1
8165	*AKAP1*	A kinase (PRKA) anchor protein 1
8659	*ALDH4A1*	aldehyde dehydrogenase 4 family, member A1
290	*ANPEP*	alanyl (membrane) aminopeptidase
338	*APOB*	apolipoprotein B (including Ag(x) antigen)
55937	*APOM*	apolipoprotein M
427	*ASAH1*	N-acylsphingosine amidohydrolase (acid ceramidase) 1
445	*ASS1*	argininosuccinate synthetase 1
23130	*ATG2A*	ATG2 autophagy related 2 homolog A (S. cerevisiae)
8678	*BECN1*	beclin 1, autophagy related
635	*BHMT*	betaine-homocysteine methyltransferase
664	*BNIP3*	BCL2/adenovirus E1B 19kDa interacting protein 3
9577	*BRE*	brain and reproductive organ-expressed (TNFRSF1A modulator)
282969	*C10orf125*	chromosome 10 open reading frame 125
64776	*C11orf1*	chromosome 11 open reading frame 1
80017	*C14orf159*	chromosome 14 open reading frame 159
720	*C4A*	complement component 4A (Rodgers blood group)
722	*C4BPA*	complement component 4 binding protein, alpha
57827	*C6orf47*	chromosome 6 open reading frame 47
821	*CANX*	calnexin
1235	*CCR6*	chemokine (C-C motif) receptor 6
8837	*CFLAR*	CASP8 and FADD-like apoptosis regulator
11261	*CHP*	calcium binding protein P22
27141	*CIDEB*	cell death-inducing DFFA-like effector b
149461	*CLDN19*	claudin 19
1314	*COPA*	coatomer protein complex, subunit alpha
84699	*CREB3L3*	cAMP responsive element binding protein 3-like 3
51496	*CTDSPL2*	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A)
28960	*DCPS*	decapping enzyme, scavenger
1644	*DDC*	dopa decarboxylase (aromatic L-amino acid decarboxylase)
84649	*DGAT2*	diacylglycerol O-acyltransferase homolog 2 (mouse)
285381	*DPH3*	DPH3, KTI11 homolog (S. cerevisiae)

Table 6.13. HNF4a PBM2 search of all annotated human promoters -2kb to +1kb relative to the transcription start site (+1) (UCSC hg18). Partial table, for a comlete table see nrmotif.ucr.edu .

Gene Id	Refseq Id	Distance to +1	Descriptor	Motif Sequence	Gene Symbol
35	NM_000017	287	Classification1.4promoters.SVM.436	CGGCCTTTGCCCC	*ACADS*
37	NM_000018	522	Msb625.top	TGACCTTTCCCCT	*ACADVL*
38	NM_000019	958	Msb149.top	TGGACTTGACCCT	*ACAT1*
183	NM_000029	198	H4.11	GAACCTTGGCCCC	*AGT*
212	NM_000032	-1896	Ebnonadj40.top	TGACCTTTGAGCT	*ALAS2*
215	NM_000033	-1806	ChipchipMSB.SVM.514	TGGACCCTGGAAT	*ABCD1*
286	NM_000037	-208	Classification1.4promoters.SVM.1075	CGACCCCGGGGCA	*ANK1*
286	NM_000037	289	Msb41.top	TGGACTCTGCCCT	*ANK1*
335	NM_000039	-206	H4.15	AACCCTTGACCCC	*APOA1*
335	NM_000039	-200	ChipchipMSB.SVM.169	TGACCCCTGCCCT	*APOA1*
345	NM_000040	-81	H4.21	TGACCTTTGCCCA	*APOC3*
348	NM_000041	-1842	Msb206.top	TGACCTTGAGACC	*APOE*
348	NM_000041	-1793	EBbiopros67.top	AGGACTTTGTCCC	*APOE*
348	NM_000041	-1546	Msb451.top	TGACCCCAGCCCT	*APOE*
348	NM_000041	566	Msb394.top	TGGAATTTGAACC	*APOE*
443	NM_000049	42	ChipchipMSB.SVM.323	TGTACTTTGCCCT	*ASPA*
445	NM_000050	-1900	Classification1.4promoters.SVM.385	GGACCCCTGACCT	*ASS1*
445	NM_000050	-535	ChipchipMSB.SVM.457	CGGGCCTGGGGTC	*ASS1*
445	NM_000050	-534	ChipchipMSB.SVM.570	TGACCCCAGGCCC	*ASS1*
445	NM_000050	544	Classification1.4promoters.SVM.400	CGGACCCGGGGAC	*ASS1*
540	NM_000053	-1206	Classification1.4promoters.SVM.515	CGACCCCTCGGCC	*ATP7B*
594	NM_000056	265	ChipchipMSB.SVM.279	TGAGCCCTGGGAC	*BCKDHB*
717	NM_000063	-166	H4.192	AAACCTTTGCCCT	*C2*
717	NM_000063	-17	H4.106	TGACCTTTTCCCT	*C2*
729	NM_000065	-851	EBvarythree30.top	TGACCTTTGAGTG	*C6*

Table 6.14. HNF4a SVM2 search of all annotated human promoters -2kb to +1kb relative to the transcription start site (+1) (UCSC hg18). Partial table, for a complete table see nrmotif.ucr.edu

Refseq Id	**Distance to +1**	**Motif score**	**Sequence**	**Gene Symbol**
NM_130786	441	1.54615261237207	GAGTCAAGGTGCA	*A1BG*
NM_130786	326	1.78449112425934	TGGGCAGAGTCCG	*A1BG*
NM_138932	-94	1.62934942892884	TGGTCAAAGGGCT	*A1CF*
NM_138932	851	1.84656297078716	TGGTAAATGTCCA	*A1CF*
NM_138933	851	1.84656297078716	TGGTAAATGTCCA	*A1CF*
NM_138933	-94	1.62934942892884	TGGTCAAAGGGCT	*A1CF*
NM_014576	851	1.84656297078716	TGGTAAATGTCCA	*A1CF*
NM_014576	-94	1.62934942892884	TGGTCAAAGGGCT	*A1CF*
NM_001142334	-1517	1.67201174193349	TGTACTAGGGTCA	*A2BP1*
NM_001142334	292	1.52361673763502	TACAAAAAGTCCA	*A2BP1*
NM_001142334	-1621	1.52694700919379	GGCTCACAGTCCA	*A2BP1*
NM_001142334	-1145	1.91640367924097	CACTCAAAGTCCA	*A2BP1*
NM_001142333	-1495	1.54526154267027	TTTGCAAAATCCA	*A2BP1*
NM_001142333	-1059	1.52148900909433	TTGTTCAAGGTCA	*A2BP1*
NM_001142333	157	1.82136037027284	GGTTTGAAGGTCA	*A2BP1*
NM_018723	-1059	1.52148900909433	TTGTTCAAGGTCA	*A2BP1*
NM_018723	157	1.82136037027284	GGTTTGAAGGTCA	*A2BP1*
NM_018723	-1495	1.54526154267027	TTTGCAAAATCCA	*A2BP1*
NM_145891	-1681	1.66231769981745	TGGCCAAAATTCA	*A2BP1*
NM_145892	-1681	1.66231769981745	TGGCCAAAATTCA	*A2BP1*
NM_145893	-1537	1.52331252665727	AATCCAAATGTCA	*A2BP1*
NM_145893	-1681	1.66231769981745	TGGCCAAAATTCA	*A2BP1*
NM_145891	-1537	1.52331252665727	AATCCAAATGTCA	*A2BP1*
NM_145892	-1537	1.52331252665727	AATCCAAATGTCA	*A2BP1*

Table 6.15. Sequences of oligonucleotides used in gel shift assays. Partial table, for the complete table see nrmotif.ucr.edu .

Designation	5’ flank1	Test sequence2	3’ flank	Reference4
Oligo’s used as probes				
ApoB (-85 to -47) (human)	gatccgggagg	CGCCCTTTGGACC	ttttgcaatcctggcgctc	Maeda et al., 2002
ApoA1 site A (-210 to -188) (human)	tcgaagggc	AGGGGTCAAGGGT	tcagt	Jiang et al., 1997
Nonspecific oligo				
TTR (-175 to -153) (mouse)	tcgaccga	TACTCTAATCTCC	ctaggc	Sladek et al., 1990
Oligo’s used as competitors in Fig. 5D				
Oligo 4n (ninj SVMnat top)	tggcagga	AAACCTAAGGTCA	gggagt	this study
Oligo 4p (ninj SVMpbm top)	tccctaggc	AAACCTAAGGTCA	gcgcg	this study
Oligo 1r (ninj random top)	tacatcag	CCTCGGGGTTTGT	aaacct	this study
Oligo’s used as competitors in Fig. S6				
YCH1	ctcatgg	GGGGCAAAGTCCA	catctcc	this study3
YCH2	acaccag	GGGTCAAAGGTCA	cactatg	this study3
YCH3	cccaaaa	GGGCCAAAGGTCT	ctatctt	this study3
YCH4	actgccc	TGGACTTAGTTCA	agtatca	this study3
YCH5	ggggtgt	GGGTCAAAGTTCA	ggtcagg	this study3
YCH6	gaggtgc	AGGGCAAAGGTCA	gattctg	this study3