Supplementary Information

for

Identification of a binding motif specific to HNF4 by comparative analysis of multiple nuclear receptors

Bin Fang, Daniel Mane-Padros, Eugene Bolotin, Tao Jiang and Frances M. Sladek

Running title: HNF4, RXR, COUPTF2 DNA binding

SUPPLEMENTARY MATERIALS AND METHODS

Transient transfection and nuclear extracts

Cos-7 cells were cultured at 37 ºC and 5% CO_2 in Dulbecco's modified Eagle's medium (DMEM), supplemented with 10% Bovine Calf Serum (BCS) and 1% Penicillin/Streptomycin. 3.5 x 10^6 cells per 150-mm plate were transiently transfected via $CaPO_4$ precipitation as previously described (1) with 25 μg of expression vector containing human NR cDNA (pcDNA3.1.hHNF4α2 (2,3), pMT2.hCOUPTF2 (4) or pMT2.hRXRα(1)). Cells were shocked 8 hr after the addition of the precipitate with 15% glycerol for 3 min and then fed with 14-ml of supplemented media. For RXRα, 9cis-retinoic acid (Sigma) was added to a final concentration of 10 μM. Nuclear extracts were prepared 24 hr after glycerol shock as previously described (1). The amount of NR protein in the extracts was determined by semi-quantitative immunoblot analysis using the recombinant HNF4α fragment containing the LBD and F domain (HNF4α LBD/F) and different flag-tagged constructs (pcDNA5-hCOUPTF2-2xFLAG, a generous gift of Ming Tsai, Baylor College of Medicine, and pCMV.RXRα.Flag from Origene) expressed in Cos-7 cells and purified, recombinant carboxy-terminal Flag-BAP fusion protein (Sigma). The Multitag Protein Marker (Genscript) was used as a standard for the His-tagged RARα vector (pCDNA6-His-hRARα, generous gift from Jonathan M. Kurie, M. D. Anderson Cancer Center).

Reporter constructs and site-directed mutagenesis

pGL4.23.hApoA1.DR1: The DR1 site (AGGGCAgGGGTCA) at -201 to -189 from the transcription start site (+1) of the human *APOA1* gene was cloned upstream of the core promoter in the luciferase reporter construct pGL4.23 (Promega). Synthetic 5'phosphorylated oligonucleotides (2 μg each) purchased from IDT (Integrated DNA Technologies) containing the ApoA1 DR1 site plus NheI/HindIII overhangs (italics) and an added AseI site for cloning purposes (underlined) (Top: 5'*CTAG*AGGGCAGGGGTCAATTAAT and Bottom: 5'*AGCTA*TTAATTGACCCCTGCCCT) were annealed by heating at 85 ºC for 5 min in 1x Hybridization Buffer (2 M NaCl, 200 mM TrisHCl pH 8.0 and 5 mM EDTA, final volume 500 μl) and then slowly cooling to room temperature. The double-stranded oligo (1 ng) was ligated into the dephosphorylated, NheI/HindIII- digested, pGL4.23 vector (100ng) for 10-12hr at 16 ºC using T4-Ligase (New England Biolabs).

pGL4.10.hApoC2: The human *APOC2* promoter(-115 bp to +18 bp) was subcloned in a similar fashion using NHeI/HindIII sites into pGL4.10 which lacks a core promoter (Promega) from a pSMART.ApoC2 (-115 to +18) vector purchased from the custom vector service of IDT.

Mutant constructs for reporter constructs and HNF4α2 (D69E, R76K and D69E/R76K) were generated using the QuikChange Mutagenesis Kit (Stratagene) following the manufacturer's protocol. All constructs were sequenced verified by the Genomics Core in the UCR Institute for Integrated Genome Biology.

Luciferase reporter assay

Human embryonic kidney, HEK293T, cells, maintained at 37 ºC and 5% CO_2 in DMEM plus 10% fetal bovine serum and 1% Penicillin/Streptomycin, were plated in 12-well plates (350,000 cells per well). Approximately 24 hr later, 500 ng of reporter construct (pGL4.23.hApoA1.DR1,

pGL4.10.hApoC2), 100 ng of β-galactosidase construct (pCMV.β-Gal) and 50 ng of expression vector (pcDNA3.1.HNF4α2, pMT2.COUPTF2 or pMT2.RXRα) or empty vector (pcDNA3.1) were co-transfected in each well using Lipofectamine2000 (Invitrogen) according to the manufacturer instructions. 24 hr later, cells were washed twice in PBS, and lysed in 100 μl of lysis buffer (25 mMGly-Gly, 15 mM MgSO4, 4 mM EGTA, 1%Triton X-100, 1 mM DTT). Extracts were nutated for 45 min at 4°C and then centrifuged in a microfuge at 12,000 rpm for 10 min. The supernatants (15 μl) were measured in a luminometer (Turner Biosystems) with the addition of 100 μl of Luciferin Buffer: 33.3 mM DTT, 270 μM Co-enzyme A, 530 μM ATP and 470 mM Luciferin in LAR buffer (20 mM Tricine pH 7.8, 2.67mM MgSO4, 0.1 mM EDTA, 1.07 mM $(MgCO_3)_4 \cdot Mg(OH)_2 \cdot 5H_2O$). β-Galactosidase activity was measured using a microplate reader (SpectraMax-190, Molecular Devices). Relative levels of gene induction were calculated by using relative lights units (RLU) normalized to β-galactosidase activity. Luciferase assays were performed in triplicate and each transfection was repeated at least three times. Where indicated, 9-cis-retinoic acid from Sigma (or ethanol as vehicle) at a final concentration of 1 μM was added to the cells during the transfection.

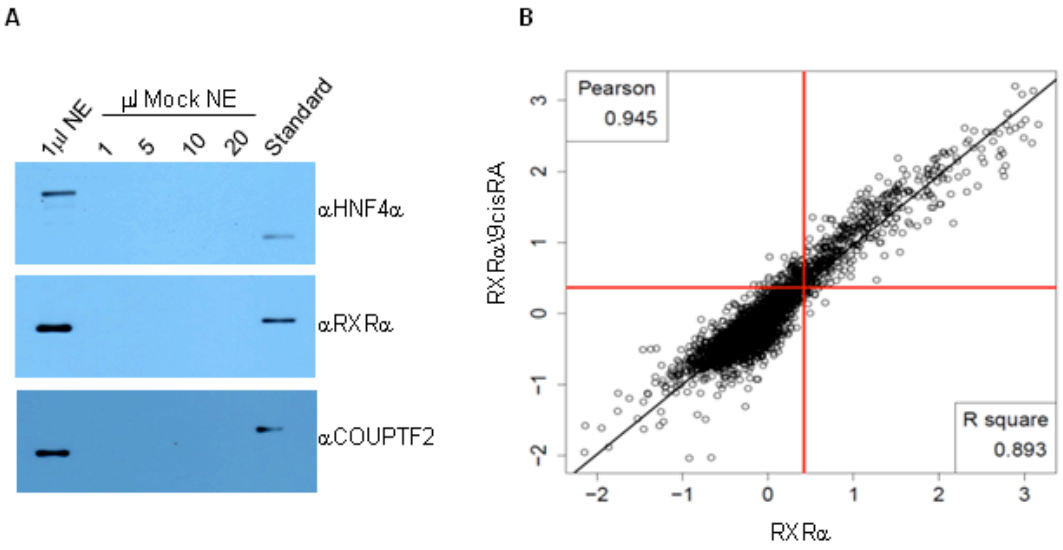
Prediction of HNF4-specific binding sequences

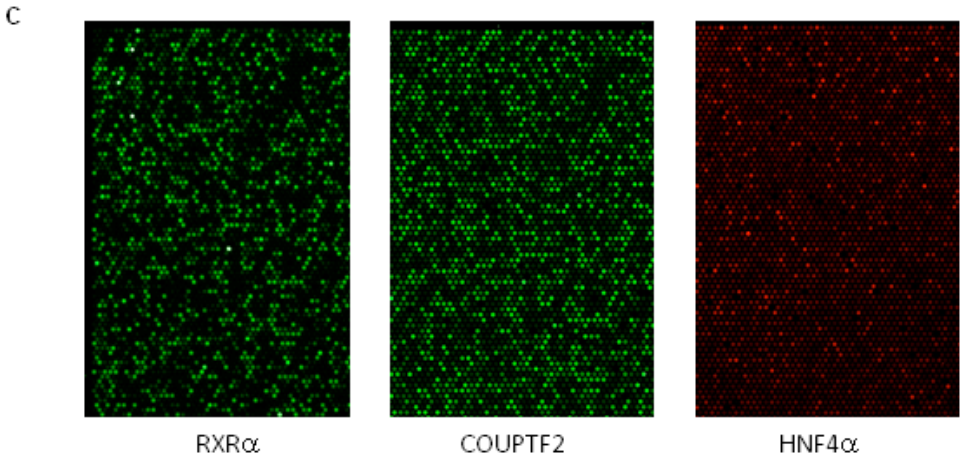
A three-step SVM model was built to predict HNF4-specific binding sequences. First, a classification SVM (SVC1) with besseldot kernel trained with the HNF4α PBM2 data including 1371 binders and 1530 non-binders predicted ~14,000 HNF4 binding sequences. These candidates were further selected by a secondary SVM classifier (SVC2) trained with 216 HNF4-specific binding sequences from the PBM, producing ~4,000 potential HNF4-specific binding sequences. In order to maintain a high confidence level, only those sequences with "TC" at p10 and p11 were selected in the last step. Finally, 2887 13-nt HNF4-specific binding sequences were selected and used in the subsequent studies. The SVM models were implemented in R using the Kernlab package. The accuracies in a 10-fold cross validation were 0.86 and 0.93 for SVC1 and SVC2, respectively.

Protein structural analysis

DBD structures of HNF4 (3CBB) (5) and RXR(1BY4) (6) were downloaded from the PDB database. The two structures were superimposed according to Cα atoms in Coot (7). The structural figure was made by Pymol 0.99rc6 (http://www.pymol.org). Multiple sequence alignment was performed by ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2).

SUPPLEMENTARY FIGURES





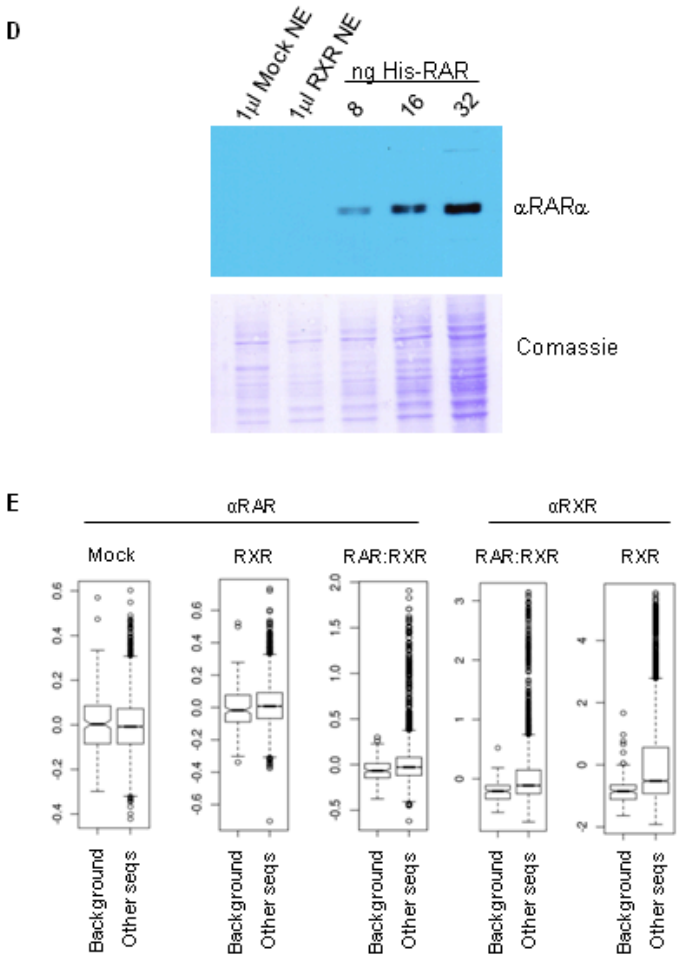


Figure S1. Experimental details of NR protein-related aspects of the PBMs.

(A) HNF4α, RXRα and COUPTF2 expressed in Cos-7 cells used in PBMs. Crude nuclear extracts (NE, 1µl) containing ectopically expressed human HNF4α2, RXRα or COUPTF2 proteins were analyzed by immunoblotting (IB) along side 1, 5, 10 and 20 µl of mock-transfected NE and appropriate standards (25 ng of HNF4α LBD/F peptide, 70 ng Flag-tagged RXRα or 80 ng 2xFlag-tagged COUPTF2) with the same antibodies used for detection in the PBM (see Methods in main text). The results allow for quantification of the amount of NR applied to the PBM and verify the specificity of the antibodies used. They also show that the ectopically RXRα is expressed in far excess of endogenous RXRα (Mock NE lanes).

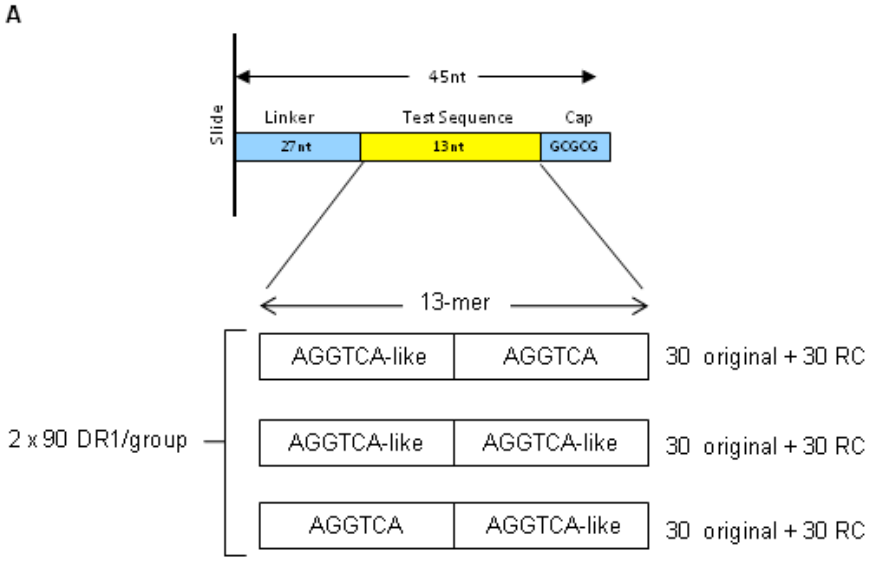
(B) 9-cis retinoic acid (9cisRA) does not affect DNA binding of RXRα on the PBM. DNA binding specificities of RXRα, in the presence or absence of 9-cisRA, are highly correlated (Pearson correlation coefficient is 0.945 and R^2 is 0.893).

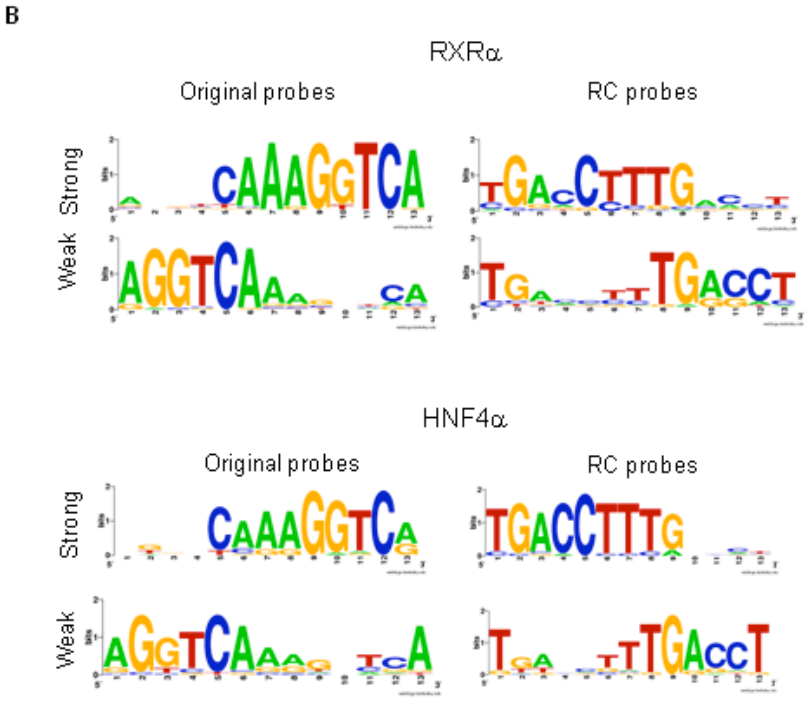
(C) Actual images of NR binding to the PBMs used for analysis. Shown is a portion of one representative grid of 15,000 DNA spots used for the analysis. The secondary antibody for RXRα and COUPTF2 is conjugated to Cy3 (green), while that of HNF4α is conjugated to Cy5

(red). Both of them work very well for bioinformatics analysis. The results show the excellent binding of RXRα, even in the absence of an ectopically expressed heterodimeric partner.

(D) Endogenous RARα in Cos-7 cells. Levels of endogenous RARα in mock-transfected Cos-7 nuclear extracts (Mock NE) and RXRα-transfected NE (sample used for RXR PBM) were investigated using a monoclonal antibody to RARα. As positive controls, nuclear extracts containing 8, 16, and 32 ng His-tag RARα were detected using the same antibody. This and similar blots indicated that there is <1 ng endogenous RARα per 1 µl of Cos-7 nuclear extracts (~2.5 µg/µl total nuclear protein).

(E) Endogenous RARα in Cos-7 cells is not enough to be detected in the PBM. Background binding signals (signals from random control probes) and signals from all other PBM probes are shown. In the mock PBM, no DNA probes showed a significantly stronger signal than the background, suggesting that binding events of endogenous RARα, if any, were not enough to be detected in the PBM. Similarly, RARα was not detected in the RXRα PBM. (~20 µl of crude nuclear extract from mock- and RXRα-transfected cells were applied to each grid; there was ~1.2 µg RXRα protein present in the latter). In contrast, when RARα is co-expresssed with RXRα in Cos-7 (RAR:RXR), both RARα and RXRα can be detected (~10 µl of crude extract containing ~1.2 µg RXRα protein.)


A
45nt
Slide
Linker
Test Sequence
Cap
27nt
13nt
GCGCG
13-mer
2 x 90 DR1/group
AGGTCA-like
AGGTCA
30 original + 30 RC
AGGTCA-like
AGGTCA-like
30 original + 30 RC
AGGTCA
AGGTCA-like
30 original + 30 RC


B
RXRα
Original probes
RC probes
Strong
Weak
HNF4α
Original probes
RC probes
Strong
Weak

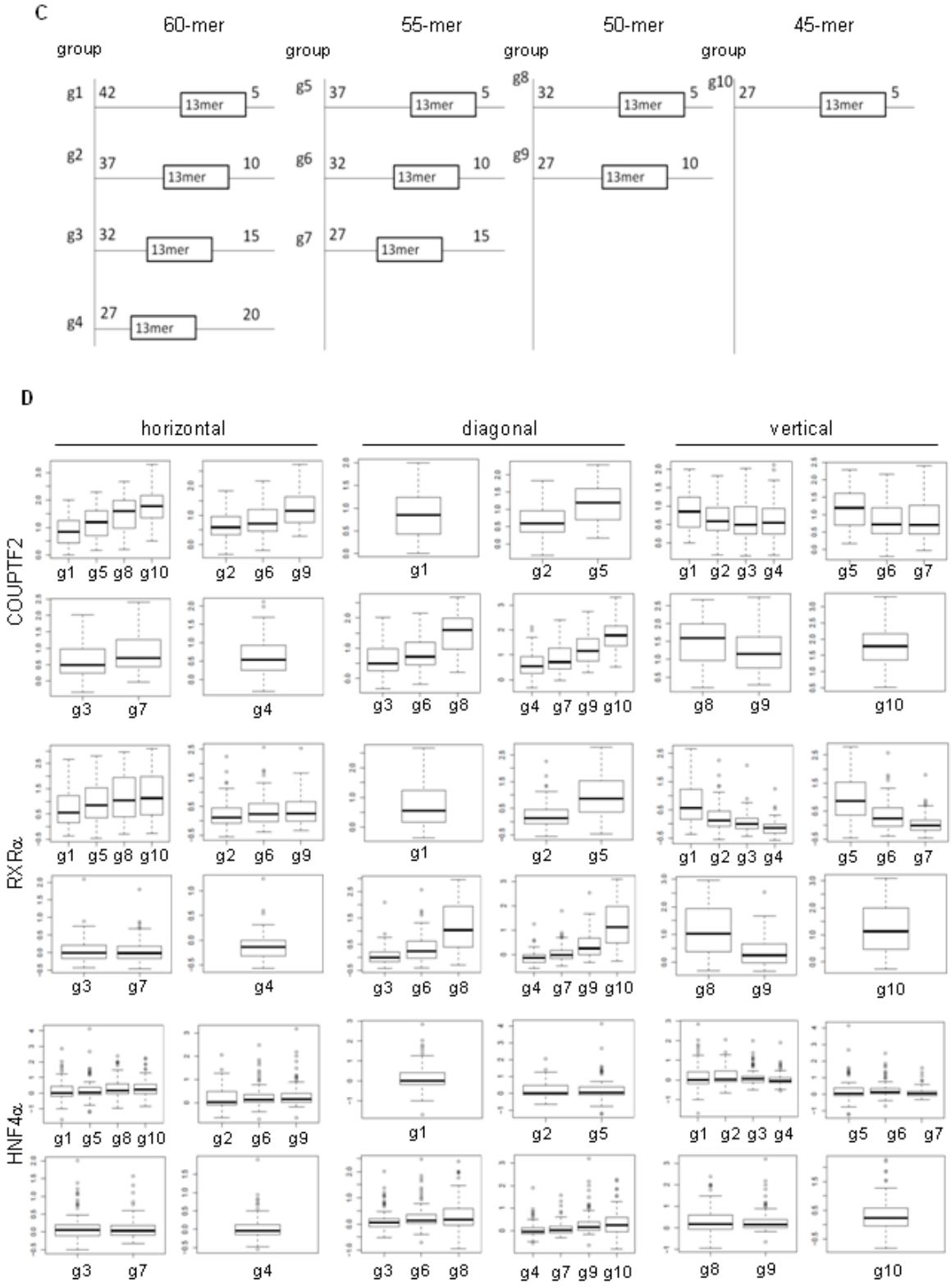

C
60-mer
55-mer
50-mer
45-mer
group
g1 42 13mer 5
g2 37 13mer 10
g3 32 13mer 15
g4 27 13mer 20
g5 37 13mer 5
g6 32 13mer 10
g7 27 13mer 15
g8 32 13mer 5
g9 27 13mer 10
g10 27 13mer 5
D
horizontal
diagonal
vertical
COUPTF2
RXRα
HNF4α
g1 g5 g8 g10
g2 g6 g9
g3 g7
g4
g1
g2 g5
g1 g2 g3 g4
g5 g6 g7
g3 g6 g8
g4 g7 g9 g10
g8 g9
g10

Figure S2. Analysis of polarity of NR DNA binding using PBMs.

(A) PBM 6.1 designed to examine the effect of DNA probe orientation and proximity to the glass slide. **Top,** schematic diagram of the DNA probes on the PBM. **Bottom,** 90 unique 13-bp test sequences (referred to as "original probes") are divided into three groups of 30 sequences each. One group contains sequences bearing the canonical DR1 half site sequence AGGTCA on the3' (right) half site with variations in AGGTCA-like sequence on the 5' (left) half sites. A second group contains AGGTCA-like sequences in both half sites. The third group contains the canonical DR1 half site sequence on the 5' (left) side and variations in AGGTCA-like sequences on the 3' (right) side. In addition, the reverse complement sequence of each probe (RC probes) was also added to the PBM, bringing the total number of probes to $2 \times 90 = 180$. All test sequences are 27 nt from the glass slide, the length of the linker sequence used for priming the *in vitro* extension of the single-stranded DNA spotted onto the slide.

(B) Polarity of the DR1 motif bound by RXRα and HNF4α. Shown are motifs produced from the DNA probes with the strongest (Strong) and weakest (Weak) binding to the indicated NR, respectively (~30 sequences per category). The motifs for the Original probes and RC probes represent essentially identical sequences but in opposite orientations, indicating that the probe orientation (and hence the distance to the slide) does not affect the binding specificity of RXRα and HNF4α. Similar results for COUPTF2 are shown in the main text.

(C) Schematic diagram of PBM probes designed to test potential effects of the length of linker and cap sequences. Each probe contains a linker sequence attached to the glass slide, a 13-bp test sequence and a GC-rich cap as shown in (**A**). Ten groups of probes (g1 to g10) varying only in the length of the linker and cap sequences as illustrated were examined. Each group contains 180 probes that differ only in the composition of the 13-bp test sequence. Numbers refer to number of nt in the linker and cap regions. Total length of probes is given on the top.

(D) Binding scores of DNA probes in groups g1 to g10 for COUPTF2, RXRα, and HNF4α. Groups in which only the length of the linker is varied (displayed in (**C**) in a horizontal fashion) are shown in the panels on the left; groups in which the length of the cap only varies (displayed in (**C**) diagonally) are shown in the middle panels; and groups in which the relative position of the 13-mer test sequence is varied (displayed in (**C**) vertically) are shown in the right panels. Overall, probes in g10 yield the best scores (higher number), indicating that these are the optimal lengths of linker and cap. The group 10 design was the default for the PBMs used in this work.

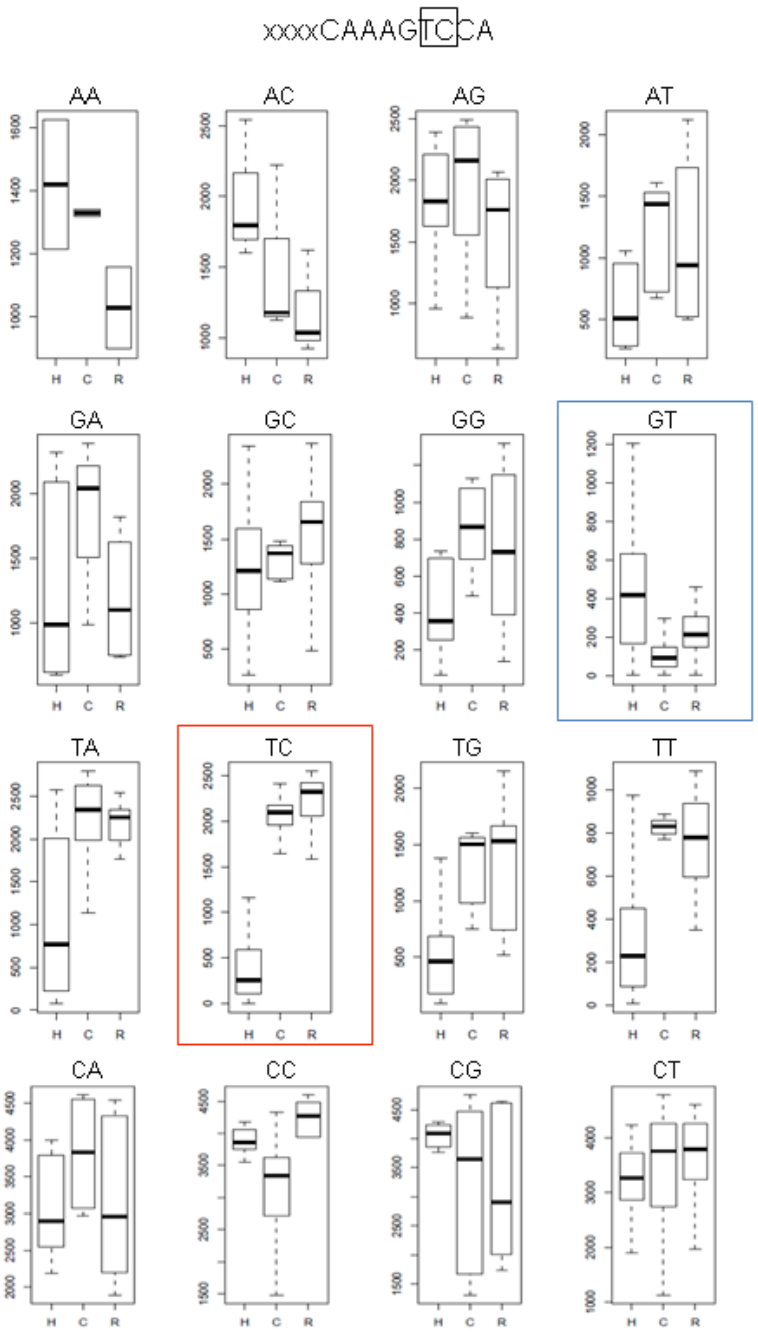


Figure S3. Preference of HNF4α for the H4-SBM and lack of binding by RXRα and COUPTF2.

PBM rankings of all 16 variations of the motif CAAAGxxCA for HNF4α (H), COUPTF2 (C), and RXRα (R) are shown. The lower the number, the higher the ranking (i.e., better binding); thresholds for HNF4α, COUPTF2, and RXRα binding are 1371, 1530, and 1285, respectively. Letters indicate nt in p10 and p11. The canonical DR1-like (CAAAGGTCA) and the H4-SBM (CAAAGTCCA) probes are highlighted by blue and red boxes, respectively.

A

NR distribution in digestive organs involved in intermediary metabolism.

	HNF4	RXR	PPAR	FXR	VDR	RAR	ER	ERR	LXR	LRH1	PXR	REV-ERB	GR	EAR2	ROR	TR	COUP-TF2
Duodenum	+	+	+	+	+	+		+	+				+	+	+	+	
Jejunum	+	+	+	+	+			+	+				+	+	+	+	
Ileum	+	+	+	+	+	+		+	+	+			+	+	+	+	
Colon	+	+	+	+	+	+		+	+	+			+	+	+	+	+
Gall bladder	+	+	+	+		+	+	+	+	+		+	+	+	+	+	+
Liver	+	+	+	+		+	+	+	+	+	+		+	+	+		+
Kidney	+	+	+	+	+	+	+	+	+			+	+	+	+	+	+

Data obtained from Nuclear Receptor Atlas (NURSA): www.nursa.org/10.1621/datasets.02001.
'+' refers to relatively high level of expression. NR isoforms are not distinguished.
Left panel: ChIP-seq data are available for these NRs in select tissues/cell lines.
Right panel: No ChIP-seq data are available for these NRs.

B

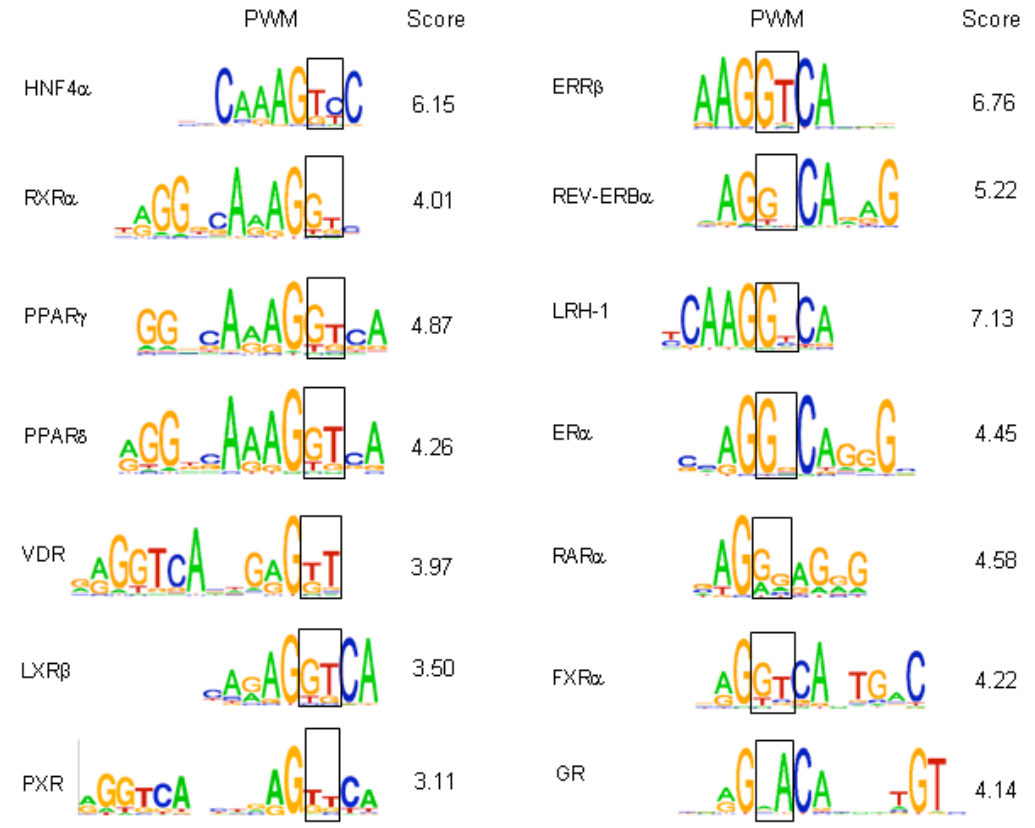


Figure S4. *In vivo* binding motifs of NRs in metabolic organs.

(A) Expression of 16 AGGTCA-binding NRs in digestive organs involved in intermediary metabolism. GR, which recognizes a different half site sequence (AGAACA), is included as a control. NRs expressed at a very low level are not included because they are less likely to compete with HNF4α. ChIP-seq data are available only for NRs in the left hand portion of the table. (See Supplementary Table S1 for details and references for ChIP-seq data for Fig.s S4-S7).

(B) DNA motif of over-represented motifs in ChIP-seq peaks of the indicated NRs. The equivalent p10-p11 positions of the H4-SBM in each motif are aligned and boxed. Motifs were mined using Gibbs Motif Sampler implemented in Cisgenome (8); the associated scores are shown.

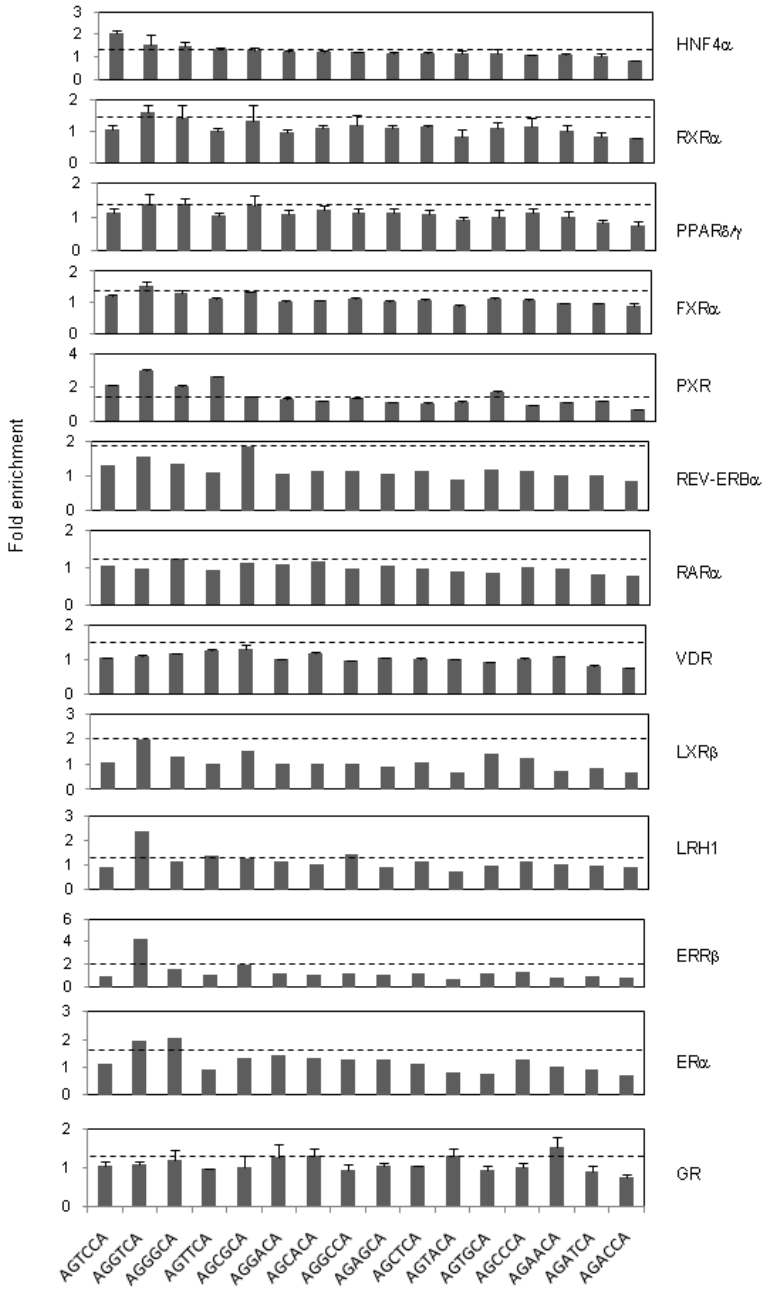


Figure S5. Fold enrichment of various half site sequences in ChIP-seq peaks of different NRs.

Fold-enrichment of 16 half site sequences derived from the canonical AGGTCA sequence with variations only in the 3rd and 4th positions in ChIP-seq peaks for the indicated NRs as described in Table S1. Error bars indicate SD from multiple datasets. Dashed lines indicate threshold for significant enrichment ($P < 0.05$ above the line), which is calculated by comparing with enrichment levels of 1000 6-nt random genomic sequences. The only NRs for which the HNF4-specific half site AGTCCA is significantly enriched are HNF4α and PXR.

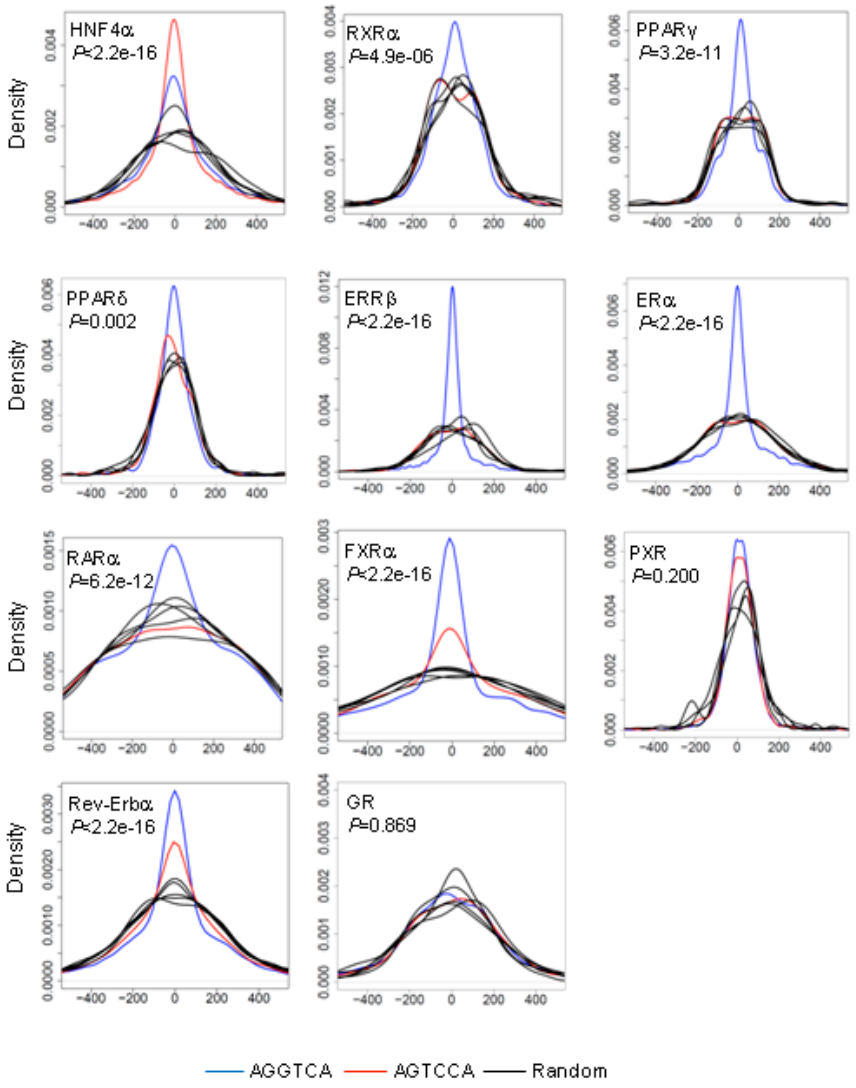
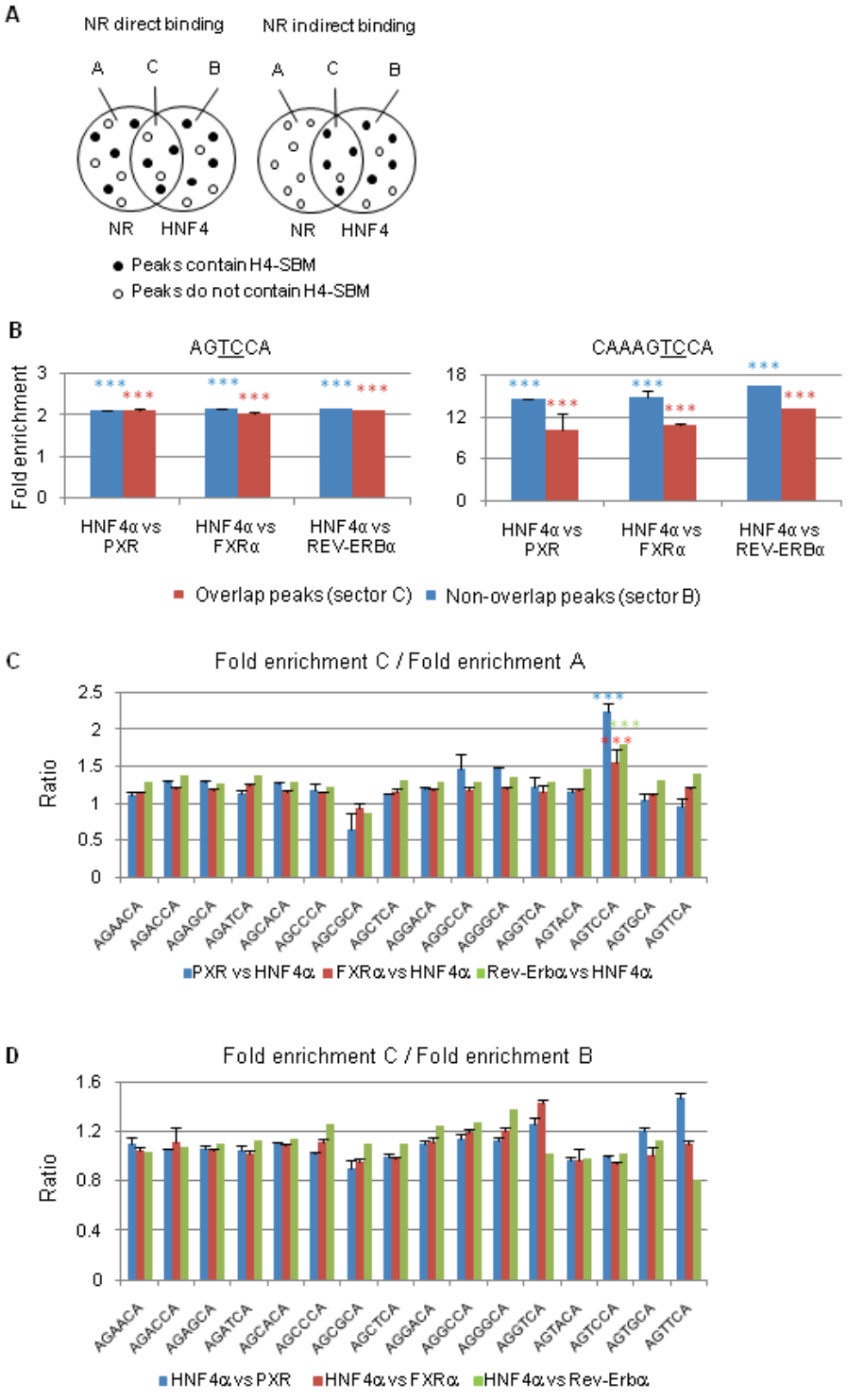
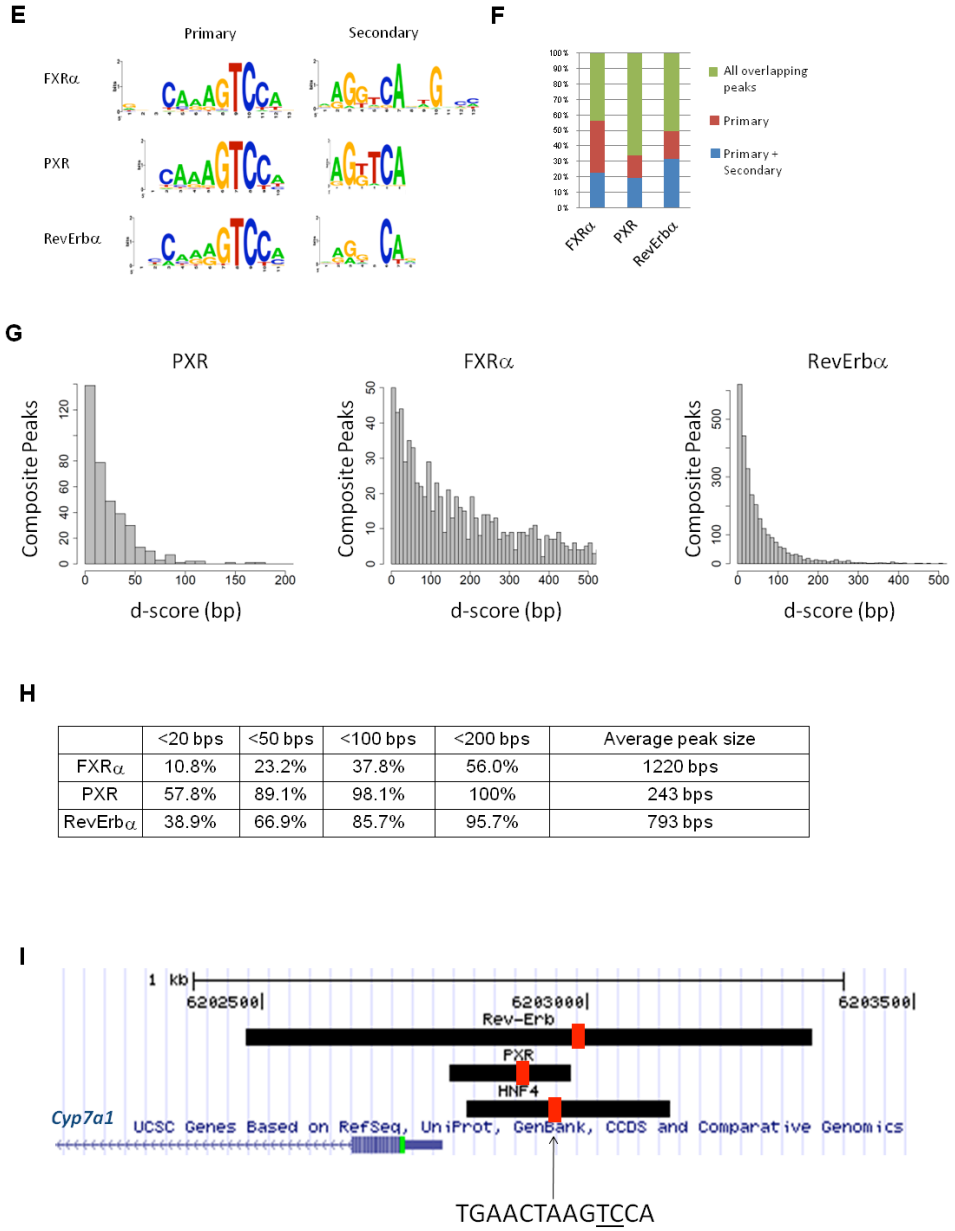


Figure S6. Distribution of half site sequences in ChIP-seq peaks of various NRs.

Distributions within ChIP-seq peaks of the canonical half site sequence AGGTCA (blue line) and the HNF4-specific half site sequences AGTCCA (red line) are shown in density plots. The x-axis indicates the distance from the peak center. Black lines represent distribution of five 6-nt random genomic sequences. P values, calculated using F-test in R, indicate differences between variances of the red and the blue line distributions. (See Table S1 for details on ChIP-seq datasets.)


A
NR direct binding
NR indirect binding
A C B
NR HNF4
Peaks contain H4-SBM
Peaks do not contain H4-SBM
B
AGTCCA
CAAAGTCCA
Fold enrichment
HNF4α vs PXR
HNF4α vs FXRα
HNF4α vs REV-ERBα
Overlap peaks (sector C)
Non-overlap peaks (sector B)
C
Fold enrichment C / Fold enrichment A
Ratio
AGAACA AGACCA AGAGCA AGATCA AGCACA AGCCCA AGCGCA AGCTCA AGGACA AGGCCA AGGGCA AGGTCA AGTACA AGTCCA AGTGCA AGTTCA
PXR vs HNF4α
FXRα vs HNF4α
Rev-Erbα vs HNF4α
D
Fold enrichment C / Fold enrichment B
Ratio
HNF4α vs PXR
HNF4α vs FXRα
HNF4α vs Rev-Erbα



	<20 bps	<50 bps	<100 bps	<200 bps	Average peak size
FXRα	10.8%	23.2%	37.8%	56.0%	1220 bps
PXR	57.8%	89.1%	98.1%	100%	243 bps
RevErbα	38.9%	66.9%	85.7%	95.7%	793 bps

Figure S7. PXR, FXRα, and Rev-Erbα bind the H4-SBM only in ChIP-seq peaks that overlap with HNF4α peaks.

(A) Schematic representation of direct and indirect binding of a NR to the H4-SBM. The direct binding is characterized by the even distribution of the H4-SBM in overlapping and non-overlapping ChIP-seq peaks (sectors C and A, respectively). In contrast, if a NR bound the H4-SBM via HNF4α, the H4-SBM would be found mainly in the overlapping peaks. This is the same diagram as in Figure 4D in the main text.

(B) Fold enrichment of the HNF4-specific half site (left panel) and the H4-SBM (right panel) in ChIP peaks of HNF4α. Enrichment levels in HNF4α peaks that overlap (sector C) or do not

overlap (sector B) with peaks of three NRs are indicated by red and blue bars, respectively. Empirical P values of enrichment (*P<.05, **P<.001, ***P<.0001) are estimated by comparison with enrichment levels of 1000 6-nt and 9-nt random genomic sequences for AGTCCA and CAAAGTCCA, respectively. Error bars indicate SD among multiple datasets (see Table S1).

(C) Enrichment of 16 half site sequences, varying at the 3rd and 4th positions, in ChIP-seq peaks of PXR, FXRα, and Rev-Erbα. The bar plot shows the ratio of the enrichment level of the indicated sequence in peaks that overlap with the HNF4α peaks (sector C) to those that do not overlap (sector A). The results show that the AGTCCA sequence is much more enriched in the overlapping peaks than the non overlapping peaks, suggesting that PXR, FXRα, and Rev-Erbα bind AGTCCA only when HNF4αis also bound.

(D) Enrichment of 16 half site sequences, varying at 3rd and 4th positions, in ChIP-seq peaks of HNF4α. The bar plot shows the ratio of the enrichment level of the indicated sequence in HNF4α peaks that overlap with peaks for PXR, FXRα, and Rev-Erbα (sector C) to those that do not overlap (sector B). The results show that the enrichment is roughly similar in both overlapping and non overlapping peaks, suggesting that HNF4α binding is not influenced by the binding of the other three NRs.

(E) Primary and secondary binding motifs over-represented in FXRα, PXR, and Rev-Erbα ChIP-seq peaks that overlap with HNF4α peaks in mouse liver. The motifs with "TC" in the 3rd and 4th positions (characteristic of the H4-SBM) are arbitrarily assigned as the primary motifs. The secondary motif was mined by masking the primary motif in each peak. An IR1-like secondary motif is over-represented in FXRα overlapping peaks, while the secondary motifs for PXR and Rev-Erbα resemble the canonical DR1 half site. The results indicate that in addition to the H4-SBM in the FXRα, PXR, and Rev-Erbα peaks, there are also sequences that resemble their own response elements.

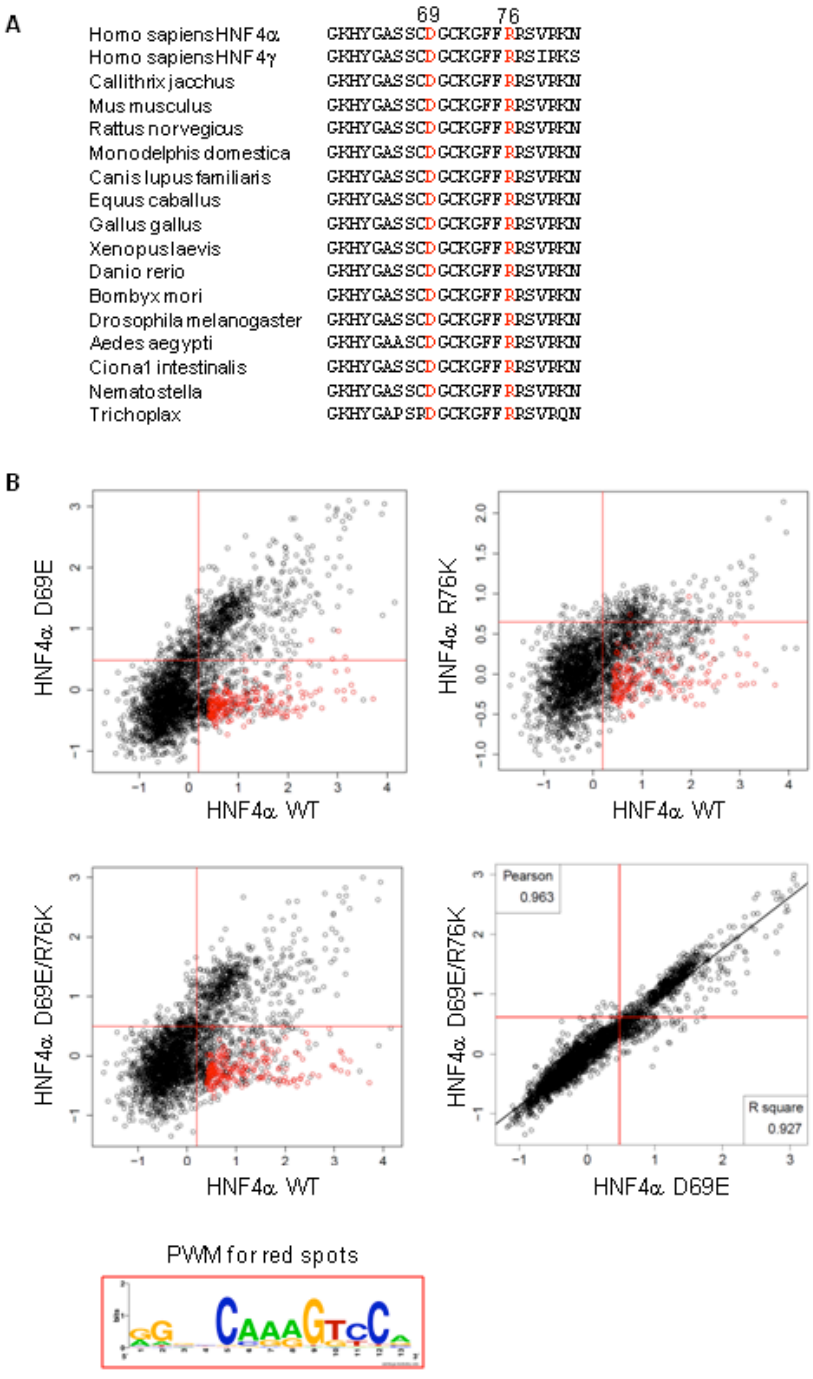
(F) Percent of all ChIP-seq peaks for FXRα, PXR, and Rev-Erbα that overlap with HNF4α peaks (green) compared to those that contain the primary motif (red) or both the primary and the secondary motifs (blue) as defined in (**E**). The results indicate that a secondary binding site can be found adjacent to an H4-SBM in about 40% to 60% of the overlapping peaks that contain an H4-SBM.

(G) Distance between the secondary and primary binding sites (as defined in (**E**)) in ChIP-seq peaks that contain both types of sites (termed composite peaks) for the indicated NRs. For each composite peak, the d-score is defined as the minimum distance in base pairs (bp) between a primary site and a secondary site. Bar plots show the number of composite peaks grouped by d-scores. The results indicate that the secondary binding sites are in close proximity to the primary sites, especially for PXR.

(H) Percentage of total composite peaks as defined in (**G**) with specified d-score cutoffs. For example, in 10.8% of the FXRα composite peaks, a secondary site can be found within 20 bps of

a primary site, which is much smaller than the average peak size of 1220 bps. Again, the PXR ChIP-seq peaks have secondary sites that are the closest to the primary site (57.8% are <20 bp away).

(I) Screenshot from UCSC Genome Browser of *in vivo* binding locations of HNF4α, PXR and Rev-Erbα in the promoter of the mouse *Cyp7a1* gene. Black bars indicate genomic regions covered by ChIP-seq peaks; red bars indicate locations of peak centers. The HNF4-specific binding site near the center of the HNF4α ChIP-seq peak is indicated; it is the same site as that originally found in the rat *Cyp7a1* promoter by Crestani et al., 1998 (9). FXR in two ChIP-seq datasets from mouse liver did not show a ChIP signal in this region (10,11).


A
69
76
Homo sapiens HNF4α GKHYGASSCDGCKGFFRRSVRKN
Homo sapiens HNF4γ GKHYGASSCDGCKGFFRRSIRKS
Callithrix jacchus GKHYGASSCDGCKGFFRRSVRKN
Mus musculus GKHYGASSCDGCKGFFRRSVRKN
Rattus norvegicus GKHYGASSCDGCKGFFRRSVRKN
Monodelphis domestica GKHYGASSCDGCKGFFRRSVRKN
Canis lupus familiaris GKHYGASSCDGCKGFFRRSVRKN
Equus caballus GKHYGASSCDGCKGFFRRSVRKN
Gallus gallus GKHYGASSCDGCKGFFRRSVRKN
Xenopus laevis GKHYGASSCDGCKGFFRRSVRKN
Danio rerio GKHYGASSCDGCKGFFRRSVRKN
Bombyx mori GKHYGASSCDGCKGFFRRSVRKN
Drosophila melanogaster GKHYGASSCDGCKGFFRRSVRKN
Aedes aegypti GKHYGAASCDGCKGFFRRSVRKN
Ciona1 intestinalis GKHYGASSCDGCKGFFRRSVRKN
Nematostella GKHYGASSCDGCKGFFRRSVRKN
Trichoplax GKHYGAPSRDGCKGFFRRSVRQN
B
HNF4α D69E
HNF4α WT
HNF4α R76K
HNF4α WT
HNF4α D69E/R76K
HNF4α WT
HNF4α D69E/R76K
HNF4α D69E
Pearson
0.963
R square
0.927
PWM for red spots

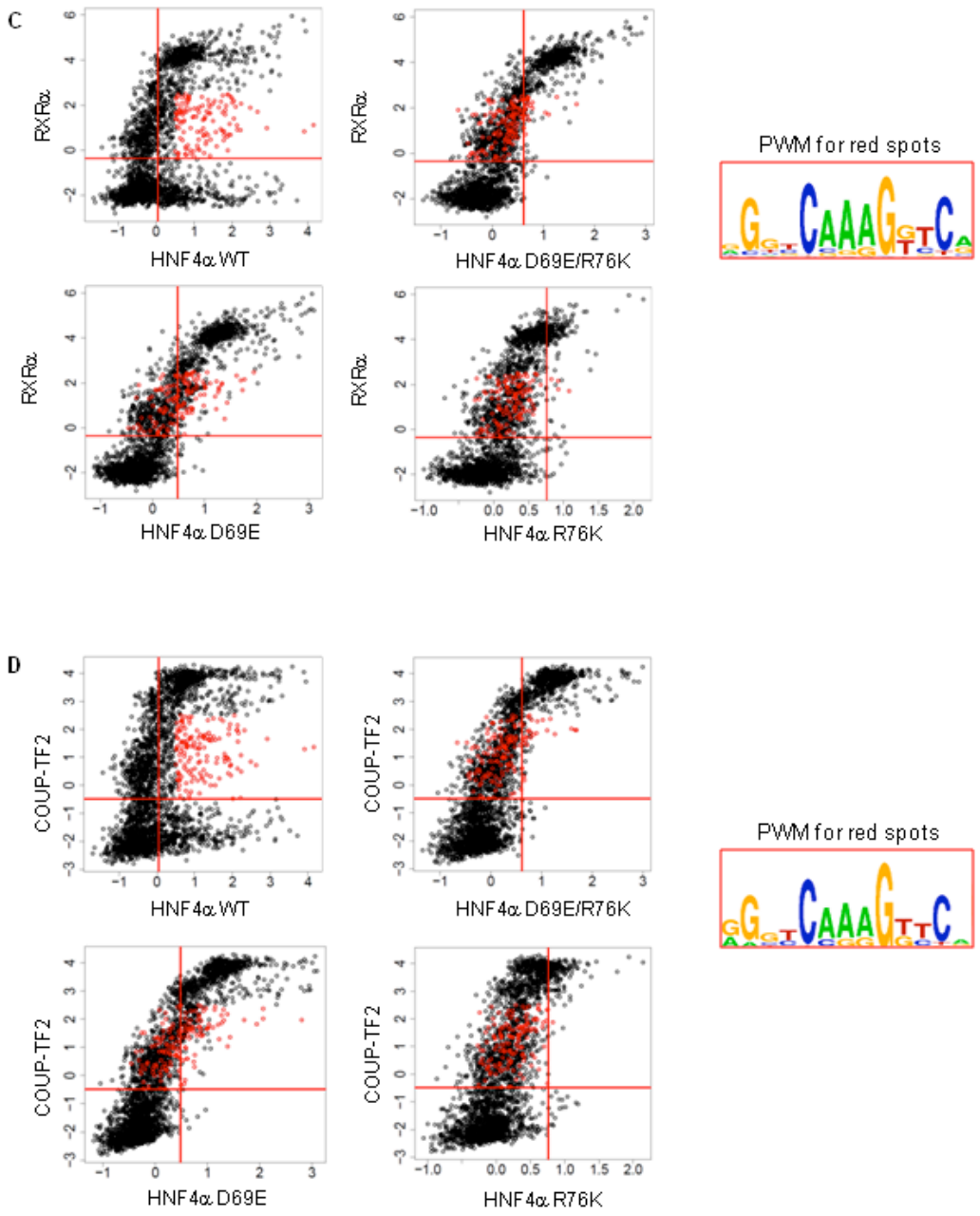


Figure S8. Asp69 and Arg76 are responsible for HNF4-specific DNA binding.

(A) Sequence alignment of the second half of the first zinc finger through helix one and including the P box (DGCKG) of the DBD is shown for human HNF4α, HNF4γ, and HNF4 orthologs in different species. Asp69 and Arg76 are highlighted in red. Numbering is from human HNF4α2 (see (12) for notes on numbering).

(B) Scatter plots comparing the binding specificity of HNF4α2 WT to the indicated mutants in the DBD. Each spot is the average of ~5 replicates for each unique DNA sequence on the PBM (~3000 total). Each plot is divided into four sub-grids by the binding thresholds of the two proteins (red lines). Spots in red refer to HNF4-specific sequences shown in Figure 5C and were used to derive the motif which is essentially the H4-SBM (CAAAGTCCA). The results show that the D69E mutant binds DNA well but not sequences containing the H4-SBM. The R76K mutant binds all DNA less well, including the H4-SBM. The double mutant D69E/R76K has a similar binding specificity to that of the D69E mutant with a Pearson correlation coefficient of 0.963 and an R^2 of 0.927.

(C) Scatter plots as in (B) comparing HNF4αWT and mutants to RXRα except that the red spots correspond to different DNA sequences as indicated in the DNA motif. The results show that the HNF4α double mutant D69E/R76K no longer binds CAAAGGTCA or CAAAGTTCA.

(D) Scatter plots as in (C) except comparing HNF4αWT and mutants to COUPTF2.

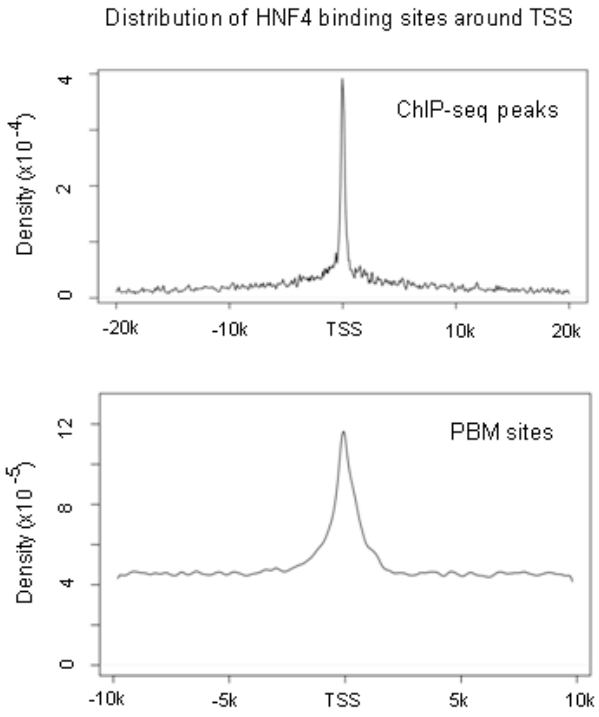


Figure S9. HNF4α binding sites are symmetrically distributed around the start site of transcription in human genes.

Distribution of HNF4α ChIP-seq peaks and HNF4α binding sequences from PBM (PBM sites) are shown in density plots. The x-axis indicates the distance from the transcription start site (TSS, +1). The ChIP-seq data are from HepG2 cells (13). A similar distribution was noted for HNF4α ChIP-seq peaks in mouse liver (14).

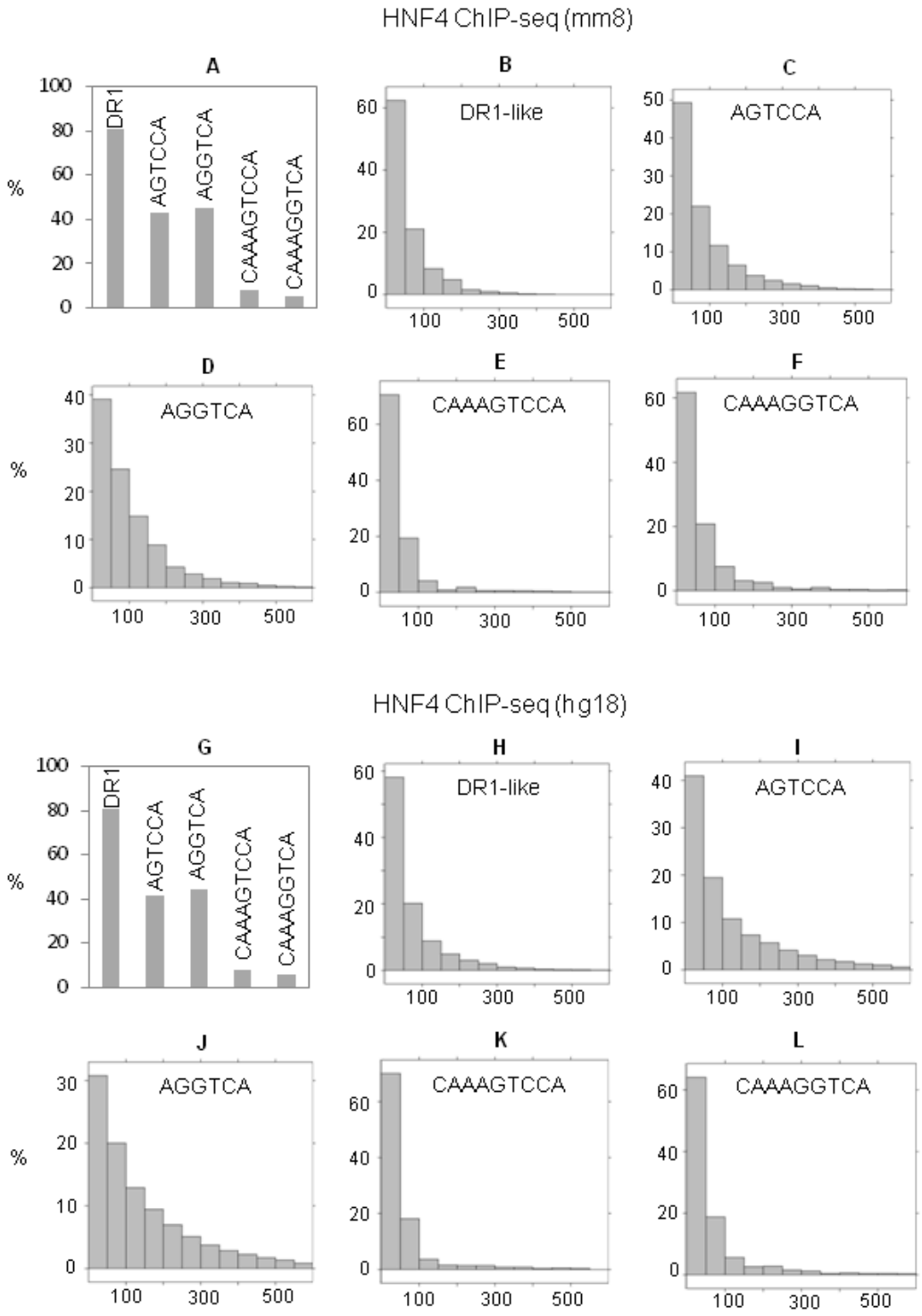


Figure S10. Abundance of HNF4α motifs in ChIP-seq peaks.

(A) Percent of HNF4α ChIP-seq peaks in mouse liver containing different motifs are shown. DR1 represents a group of 13-mer sequences with four or fewer mutated positions in the canonical DR1 motif AGGTCAAAGGTCA. The prevalence of ChIP peaks with the HNF4-specific motif (AGTCCA) is similar to that with the canonical AGGTCA motif (~40%).

(B - F) Distributions of each type of motif in ChIP-seq peaks relative to the peak center. X-axis refers to the distance to the peak center and Y-axis refers to the percentage of total peaks that contain each indicated motif.

(G - L) Same analysis on HNF4α ChIP-seq data from HepG2 cells.

(See Supplementary Table S1 for references to the ChIP-seq datasets.)

Table S1. ChIP-seq datasets for motif enrichment analysis.

Data set	Index	Genome	Cell/Tissue	Peaks	Source	Processing	Reference
HNF4α	1	hg18	HepG2	18994	http://nar.oxfordjournals.org/content/37/22/7498/suppl/DC1	Pre-processed	Wallerman O, et al 2009
	2*	mm8	Liver	12833	Author provided	Pre-processed	Hoffman BG, et al 2010
RXRα	3*	mm9	3T3-L1	5021	GEO: GSE13511(day2)	Pre-processed	Nielsen R, et al 2008
	4	mm9	3T3-L1	8510	GEO: GSE13511(day6)	Pre-processed	Nielsen R, et al 2008
	5	mm9	3T3-L1	2076	GEO: GSM686978	Pre-processed	Siersbæk R, et al 2011
PPARγ	6*	mm9	3T3-L1	7219	GEO: GSE13511(day6)	Pre-processed	Nielsen R, et al 2008
	7	mm9	3T3-L1	7142	http://www.ncbi.nlm.nih.gov/pubmed/20887899	Pre-processed	Mikkelsen TS, et al 2010
	8	hg18	hASCs	39986	http://www.ncbi.nlm.nih.gov/pubmed/20887899	Pre-processed	Mikkelsen TS, et al 2010
	9	mm9	3T3-L1	5493	GEO: GSM686978	Pre-processed	Siersbæk R, et al 2011
PPARδ	10*	hg19	WPMY-1	4543	http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0016344#s4	Pre-processed	Adhikary T, et al 2011
VDR	11*	hg18	GM10855	4019	GEO: GSM558632	Pre-processed	Ramagopalan SV, et al 2010
	12	hg18	GM10855	2864	GEO: GSM558633	Pre-processed	Ramagopalan SV, et al 2010
FXRα	13	mm9	Liver	1656	http://cbcl-1.ics.uci.edu/public_data/FXR/	Reprocessed	Chong HK, et al 2010
	14*	mm9	Liver	6395	http://genome.ucsc.edu/goldenPath/customTracks/custTracks.html#Mouse	Pre-processed	Thomas AM, et al 2010
	15	mm9	Intestine	5367	http://genome.ucsc.edu/goldenPath/customTracks/custTracks.html#Mouse	Pre-processed	Thomas AM, et al 2010
PXR	16*	mm9	Liver	3699	http://nar.oxfordjournals.org/content/38/22/7943/suppl/DC1 (PCN-treated)	Pre-processed	Cui JY, et al 2010
	17	mm9	Liver	2181	http://nar.oxfordjournals.org/content/38/22/7943/suppl/DC1 (non-treated)	Pre-processed	Cui JY, et al 2010
ERRβ	18*	mm8	ES	4801	GEO: GSE11431	Reprocessed	Chen, et al 2008
RARα	19*	Hg18	MCF7	11160	Author provided	Pre-processed	Ross-Innes, et al 2010
ERα	20*	hg18	HepG2	10000	ArrayExpress: E-TABM-828	Pre-processed	Schmidt D, et al 2010
LXRb	21*	mm8	RAW264.7	664	GEO: GSE21512	Pre-processed	Heinz S, et al 2010
LRH-1	22*	mm8	ES	3346	GEO: GSE19019	Pre-processed	Heng D, et al 2010
REV-ERBα	23*	mm8	Liver	30826	GEO: GSE26345	Reprocessed	Feng D, et al 2011
GR	24	mm8	3T3-L1	4008	GEO: GSM544722	Pre-processed	Steger DJ, et al 2010
	25	mm9	3T3-L1	4080	GEO: GSM686976	Pre-processed	Siersbæk R, et al 2011
	26*	Hg18	A549	4393	Author provided	Pre-processed	Reddy TE, et al 2009

*Datasets used for motif mining and density plots.

Table S2. Previously reported HNF4α response elements.

Gene	Site Sequence	Position	Reference	PubMed ID	RXR*	COUP-TF2*
Acaa1	GGTTCAaAGGTCT	-669/-681	Nicolas-Frances et al. 2000	10708554	y	
ACADM	CGGGTAaAGGTGA	-323/-311	Carter et al. 1993	8314750	y	
AGT	AGGGCAgAGGGCA	-395/-407	Yanai et al. 1999	10574924		y
AGT	GGGGCCaAGGTTC	-260/-272	Yanai et al. 1999	10574924		y
APOA1	AGTTCAaGGATCA	-120/-132	Chan et al. 1993	8464705		y
APOA2	AGGGTAaAGGTTG	-721/-733	Ladias et al. 1992	1639815		y
APOB	GGTCCAaAGGGCG	-66/-78	Metzger et al. 1993	8344962		y
APOC2	AGGCCAaAGTCCT	-81/-93	Kardassis et al. 1998	9651383	n	n
APOC3	GGTCCAgAGGGCA	-731/-719	Kardassis et al. 1997	9012660	y	y
Cyp3A1	GTACCAaAGTCCA	-103/-91	Ogino et al. 1999	9917326		n
Cyp7a1	TGAACTAAGTCCA	-146/-134	Crestani et al. 1998	9799805	n	n
Cyp7a1	TGAACTAAGTCCA	-146/-134	Stroup et al. 2000	10627496	n	n
CYP8B1	AGGGCAaGGTCCA	208/220	Zhang and Chiang 2001	11535594	n	
F9	CTAGCAaAGGTTA	18/30	Naka and Brownlee 1996	8562402		y
F9	AGTGGTaAGGTCG	6/-6	Naka and Brownlee 1996	8562402		y
Fabp2	AGTTCAaAGTTCA	-84/-72	Rottman and Gordon 1993	8505324		y
HIV LTR	GGGCCAaGGGTCA	-355/-343	Ladias 1994	8119938	y	y
Hmgcs2	GGGCCAaAGGTCT	-92/-104	Rodriguez et al. 1998	9464279	y	
HNF1	GGGACAaAGTTCA	-237/-225	McNair et al. 2000	11085951	n	y
Hnf1b	AGTCCAaAGGTCA	-258/-246	Power and Cereghini 1996	8622679	y	y
Pck1	CGGCCAaAGGTCA	-439/-451	Hall et al. 1992	1333043	y	
Pck1	GTGGTAaAGGTCT	−451/−439	Scribner et al. 2006	16713227	y	y
PPARA	GGGGCAaAGTTCA	-1493/-1481	Pineda Torra et al. 2002	11981036	y	y
Rbp2	GAGTCAaAGGTCA	-62/-74	Nakshatri and Chambon 1994	8288643	y	y
SERPINC1	AGGTCAaAGGCTG	-73/-85	Fernandez-Rachubinski et al. 1996	8910619	y	
hepatitis B virus enhancer I	GGGGTAaAGGTTC	1151/1139	Garcia et al. 1993	8389913	y	
hepatitis B virus nucleocapsid	AGGTTAaAGGTCT	-4/-16	Raney et al. 1997	8995626	y	y
hepatitis B virus nucleocapsid	AGTCCAAGAGTCCT	-123/-111	Raney et al. 1997	8995626	n	n

* RXR or COUP-TF2 can (y) or cannot (n) activate/repress the gene expression using the same binding sites.

REFERENCES

1. Jiang, G., Nepomuceno, L., Hopkins, K. and Sladek, F.M. (1995) Exclusive homodimerization of the orphan receptor hepatocyte nuclear factor 4 defines a new subclass of nuclear receptors. *Mol Cell Biol*, **15**, 5131-5143.
2. Chartier, F.L., Bossu, J.P., Laudet, V., Fruchart, J.C. and Laine, B. (1994) Cloning and sequencing of cDNAs encoding the human hepatocyte nuclear factor 4 indicate the presence of two isoforms in human liver. *Gene*, **147**, 269-272.
3. Eeckhoute, J., Formstecher, P. and Laine, B. (2001) 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 co-activator. *Mol Endocrinol*, **15**, 1200-1210.
4. Mietus-Snyder, M., Sladek, F.M., Ginsburg, G.S., Kuo, C.F., Ladias, J.A., Darnell, J.E., Jr. and Karathanasis, S.K. (1992) 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**, 1708-1718.
5. Lu, P., Rha, G.B., Melikishvili, M., Wu, G., Adkins, B.C., Fried, M.G. and Chi, Y.I. (2008) Structural basis of natural promoter recognition by a unique nuclear receptor, HNF4alpha. Diabetes gene product. *J Biol Chem*, **283**, 33685-33697.
6. Rastinejad, F., Perlmann, T., Evans, R.M. and Sigler, P.B. (1995) Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature*, **375**, 203-211.
7. Emsley, P., Lohkamp, B., Scott, W.G. and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*, **66**, 486-501.
8. Ji, H. and Wong, W.H. (2005) TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, **21**, 3629-3636.
9. Crestani, M., Sadeghpour, A., Stroup, D., Galli, G. and Chiang, J.Y. (1998) Transcriptional activation of the cholesterol 7alpha-hydroxylase gene (CYP7A) by nuclear hormone receptors. *J Lipid Res*, **39**, 2192-2200.
10. Chong, H.K., Infante, A.M., Seo, Y.K., Jeon, T.I., Zhang, Y., Edwards, P.A., Xie, X. and Osborne, T.F. (2010) Genome-wide interrogation of hepatic FXR reveals an asymmetric IR-1 motif and synergy with LRH-1. *Nucleic Acids Research*, **38**, 6007-6017.
11. Thomas, A.M., Hart, S.N., Kong, B., Fang, J., Zhong, X.-b. and Guo, G.L. (2010) Genome-wide tissue-specific farnesoid X receptor binding in mouse liver and intestine. *Hepatology*, NA-NA.
12. Bolotin, E., Schnabl, J. and Sladek, F. (2009) HNF4A (Homo sapiens). Updated 2010. *Transcription Factor Encyclopedia,* http://www.cisreg.ca/cgi-bin/tfe/home.pl *Accessed [July 14, 2011]*.
13. Wallerman, O., Motallebipour, M., Enroth, S., Patra, K., Bysani, M.S.R., Komorowski, J. and Wadelius, C. (2009) Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Research*, **37**, 7498-7508.
14. Hoffman, B.G., Robertson, G., Zavaglia, B., Beach, M., Cullum, R., Lee, S., Soukhatcheva, G., Li, L., Wederell, E.D., Thiessen, N. *et al.* (2010) Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. *Genome Research*, **20**, 1037-1051.