Published online 13 June 2008

Nucleic Acids Research, 2008, Vol. 36, No. 12 ***4149–4157***
doi:10.1093/nar/gkn366

Cis-regulatory modules in the mammalian liver: composition depends on strength of Foxa2 consensus site

Geetu Tuteja[1,2], Shane T. Jensen[3], Peter White[1] and Klaus H. Kaestner[1,*]

[1]Department of Genetics, [2]Genomics and Computational Biology Graduate Group, University of Pennsylvania School of Medicine and [3]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

Received March 20, 2008; Revised May 21, 2008; Accepted May 22, 2008

ABSTRACT

Foxa2 is a critical transcription factor that controls liver development and plays an important role in hepatic gluconeogensis in adult mice. Here, we use genome-wide location analysis for Foxa2 to identify its targets in the adult liver. We then show by computational analyses that Foxa2 containing *cis*-regulatory modules are not constructed from a random assortment of binding sites for other transcription factors expressed in the liver, but rather that their composition depends on the strength of the Foxa2 consensus site present. Genes containing a *cis*-regulatory module with a medium or weak Foxa2 consensus site are much more liver-specific than the genes with a strong consensus site. We not only provide a better understanding of the mechanisms of Foxa2 regulation but also introduce a novel method for identification of different *cis*-regulatory modules involving a single factor.

INTRODUCTION

Transcriptional regulation in mammals is a complex and highly orchestrated process. One level of control is through the binding of transcription factors to specific sequences of DNA. Most mammalian transcription factors do not act alone, but instead work with other factors to form *cis*-regulatory modules to control gene expression (1). Over the last several years, *cis*-regulatory systems in the liver have been studied in detail (2–12). Some of these studies focused on individual binding and potential interactions between known hepatic regulators, but did not attempt to exploit computational tools to identify additional transcription factors that may be a part of the regulatory modules operative in the liver (2,6–8). Another group of studies utilized tissue-specific gene-expression information, and then identified *cis*-regulatory modules in promoter regions of tissue-specific genes, but did not take into account any *in vivo* binding data (9–12).

One factor that is known to play an important role in regulating gene expression in the liver is Foxa2. Foxa2 is a member of the Foxa subfamily of *Forkhead* transcription factors, characterized by a highly conserved 110 amino acid motif that functions as a DNA-binding domain (13). Gene ablation studies have demonstrated that Foxa2 is a critical factor in the development of the liver, and is also an important regulator of the gluconeogenic program in adult mice (14,15). Genome-wide location analysis has been carried out to identify potential Foxa2 targets in HepG2 hepatoma cells and primary hepatocytes (2,7). These studies confirmed that Foxa2 is commonly bound to promoter regions in which other hepatic transcription factors are also bound, which had also been described in previous studies of promoters and enhancers of individual genes. For example, Foxa2 binds to the Fabpl promoter region, where hepatocyte nuclear factor (HNF) HNF-1, C/EBP-β, GATA-4 and HNF4-α also bind, and also activates transthyretin (TTR) expression by cooperating with other factors in its promoter and enhancer (16,17).

The aim of this study was to identify additional factors that potentially interact with Foxa2. Using genome-wide location analysis combined with computational methods, we identify several potential binding partners of Foxa2 and show that the likelihood of a given factor's consensus sequence to be found near Foxa2 is dependent on the strength of the Foxa2-binding site.

MATERIALS AND METHODS

Chromatin immunoprecipitation (ChIP)

ChIPs were performed as described previously (5).

Mouse livers were minced finely in cold phosphate-buffered saline (PBS) and cross-linked in 1% formaldehyde for 10 min while rotating. Cross-linking was quenched by adding glycine to a final concentration of 0.125 M for 5 min while rotating. The tissue was rinsed in cold PBS and

*To whom correspondence should be addressed. Tel: +1 215 898 8759; Fax: +1 215 573 5892; Email: kaestner@mail.med.upenn.edu

homogenized with a Dounce homogenizer in cold cell lysis buffer (10 mM Tris–Cl, pH 8.0, 10 mM NaCl, 3 mM $MgCl_2$, 0.5% NP-40) and protease inhibitors. Cells were incubated at 4°C for 5 min to release nuclei. Nuclei were centrifuged at 13 000 *g* for 5 min to form a pellet. The pellet was resuspended in nuclear lysis buffer [1% sodium dodecyl sulfate (SDS), 5 mM EDTA, 50 mM Tris–Cl, pH 8.1) and protease inhibitors and sonicated using the Diagenode Bioruptor for 10 min on high, using 30 s intervals. Debris were removed by centrifugation at 13 000 *g* for 10 min, and the supernatant was collected and snap frozen in liquid nitrogen. A 10 µl aliquot was reversed by the addition of NaCl to a final concentration of 192 mM, overnight incubation at 65°C, and purification using a PCR purification kit (Qiagen, CA, USA). The chromatin concentration was determined using a NanoDrop 3.1.0 nucleic acid assay (Agilent Technologies, Santa Clara, CA, USA).

Ten micrograms of chromatin per sample was precleared by adding 90 µl of protein G-agarose in 1 ml of ChIP dilution buffer (0.01% SDS, 1.1% Triton X-100, 167 mM NaCl, 16.7 mM Tris–Cl, pH 8.1) and rotating the sample for 1 h at 4°C. Protein G-agarose was sedimented by centrifugation at 3000 *g* for 30 s. Two micrograms of rabbit anti-Foxa2 serum (provided by J.A. Whitsett), was added to the supernatant and incubated overnight at 4°C. Protein G-agarose was blocked overnight at 4°C with 1 mg/ml bovine serum albumin and 0.1 mg/ml herring sperm DNA in ChIP dilution buffer, added to the chromatin, and rotated for 1 h at 4°C. Following three consecutive washes of 5 min each with TSE I (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–Cl, pH 8.1, 150 mM NaCl), TSE II (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–Cl, pH 8.1, 500 mM NaCl) and ChIP buffer III (0.25 M LiCl, 1% NP-40, 1% deoxycholate, 1 mM EDTA, 10 mM Tris–Cl, pH 8.1), chromatin was eluted by adding 100 µl of freshly made ChIP elution buffer (1% SDS, 0.1 M $NHCO_3$) to the pellet and rotating the sample for 10 min. Elution was repeated with an additional 100 µl of ChIP elution buffer, and the eluates were combined. Cross-linking was reversed by the addition of NaCl to a final concentration of 192 mM and overnight incubation at 65°C.

Ligation-mediated PCR (LM-PCR)

DNA blunting, linker ligation and amplification were carried out following the Agilent Mammalian ChIP-on-chip Protocol (http://www.chem.agilent.com/scripts/LiteraturePDF.asp?iWHID=42637).

Labeling and hybridization to mouse promoter chip BCBC-5A

LM-PCR amplified ChIP DNA was labeled using the BioPrime® Array CGH Genomic Labeling System (Invitrogen Life Technologies, CA, USA) as per manufacturer's instructions. Briefly, 1 µg of DNA was mixed with random primers and denatured at 95°C for 5 min, then cooled briefly on ice. Next, the appropriate Cyanine dUTP fluorescent nucleotides (PerkinElmer Life And Analytical Sciences, Inc., MA, USA) were added, along with the nucleotide mix and Exo Klenow fragment. This was gently mixed and incubated at 37°C for 2 h. The Cy3 and Cy5 labeled samples were purified using the MinElute PCR Purification Kit (Qiagen), and the efficiency of dye incorporation and yield was determined using the NanoDrop® ND-1000 UV-Vis Spectrophotometer. The Cy5 and Cy3 labeled samples were combined and 1 µg of Mouse Cot1 DNA (Invitrogen Life Technologies) was added to each sample and denatured at 95°C for 5 min. The samples were then cooled to 42°C and an equal volume of 2 × hybridization buffer (50% formamide, 10 × SSC and 0.2% SDS) was added, mixed and applied to the array.

The BCBC-5A chip contains over 18 000 proximal and distal promoter regions. Promoter regions were determined from full-length cDNA libraries and Reference Sequences (RefSeqs). Over 12 000 well-characterized genes were chosen and are represented by either a 1- or 2-kb tile, PCR amplified from genomic material. Microarray slides were hybridized overnight, then washed and scanned with Agilent G2565BA Microarray Scanner. Images were analyzed with GenePix 5.0 software (Axon Instruments, Molecular Devices, Union City, CA, USA).

Data processing and analysis

Median foreground intensities were obtained for each spot and imported into the mathematical software package 'R' (http://www.r-project.org/), which is used for all data input, diagnostic plots, normalization and quality checking steps of the analysis process using scripts developed in-house. The ratio of expression for each element on the array was calculated in terms of M[log2(Red/Green)] and A[(log2(Red) + log2(Green))/2)]. The dataset was filtered to remove positive control elements (Cy3 anchors and SpotReport elements) and any elements that had been manually flagged as poor quality. The M-values were then normalized by the print tip loess method using the 'marray' microarray processing package in 'R'. Statistical analysis was performed in 'R' using both the LIMMA and SAM packages. Foxa2 targets were defined as elements that had an expression ratio greater than 1.3 and a false discovery rate (calculated by SAM) of less than 10%.

Identification of enriched binding sites

Position weight matrices (PWMs) from the TRANSFAC database were scanned across target (bound) sequences, and 1000 random sequences from the Promoter Chip BCBC 5A (unbound) sequences, which were used as background. All Promoter Chip tile sequences were padded by 300 bp on each end. Only the top scoring match to the PWM for each sequence was analyzed further. For several score cutoffs, the number of true positives (target sequences with a score above the cutoff), and the number of false positives (background sequences with a score above the cutoff) were calculated. Plotting the true positive fraction versus the false positive fraction produces a receiver operating characteristic curve (ROC curve). ROC curves were generated for all PWMs, and then the area under the curve (AUC) was calculated. The 100 permutations were carried out by combining target sequences and background sequences into one set, and then randomly selecting 107 sequences from the set. This set was treated as

the 'target' set, while the remaining sequences were used as background. The *P*-value was then calculated by counting the number of times the AUC in the random target sets exceeded the real AUC, divided by 100. A PWM was considered to be enriched if the AUC was >0.5, and the *P*-value was ≤0.01.

Quantitative RT–PCR

Real-time PCRs were assembled using SYBR GreenER (Invitrogen). Reactions were performed in triplicate using the Mx3000 PCR System (Stratagene, La Jolla, CA, USA). The enrichment of target genes was calculated using the 28S rRNA locus as a reference for nonspecific DNA, and was calculated by comparing input (sheared genomic DNA) to ChIP material. Primer sequences are provided in Supplementary Table 7.

RESULTS

Genome-wide location analysis

We carried out ChIP with five biological replicates of adult mouse liver using a specific Foxa2 antibody as described previously (18). We validated the specificity of this antibody by confirming that the Foxa2 promoter, which Foxa2 itself is normally bound to, is no longer occupied when ChIP is carried out using the livers of *Foxa2*$^{loxP/loxP}$ *Alfp.Cre* mice, which lack Foxa2 in hepatocytes (2,19,20) (Figure 1A and B). ChIP samples from the wild-type mice were amplified, labeled and hybridized to the Mouse Promoter Chip 5A, which contains over 18 000 promoter and enhancer elements. Using the computational tools and statistical methods described in 'Materials and methods' section, we identified 107 Foxa2 target sites in the liver (Figure 1C, Supplementary Table 1). We used the hypergeometric distribution to show that this set of targets overlaps significantly with the genes identified in a prior location analysis of human primary hepatocytes (2) (Supplementary Table 2).

Scanning of PWMs

To identify potential interacting partners of Foxa2, we used the transcription factor binding site (TFBS) module in Bioperl (21) in order to identify overrepresented *cis*-regulatory elements. After scanning all of the vertebrate PWMs contained in the TRANSFAC database on both Foxa2 target sequences and unbound background sequences,

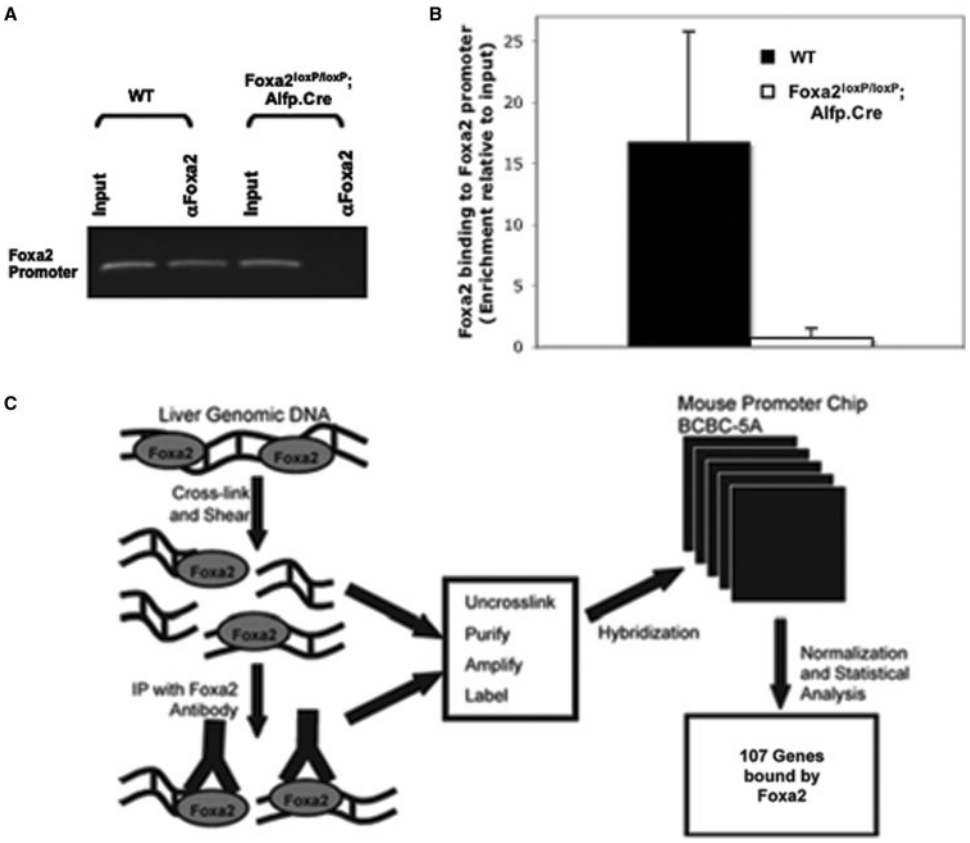

Figure 1. Validation of Foxa2 antibody and results of genome-wide location analysis. (**A** and **B**) Chromatin from livers of wild-type and *Foxa2*$^{loxP/loxP}$ *Alfp.Cre* mice was immunoprecipitated with an anti-Foxa2 antibody. Input chromatin and precipitated DNA were amplified with primers surrounding the Foxa2-binding site in the Foxa2 promoter. Occupancy of the Foxa2 site is not detected in *Foxa2*$^{loxP/loxP}$ *Alfp.Cre* mice both in a qualitative assay (A) and using quantitative real-time PCR (B). In (B), enrichment was calculated relative to the input chromatin and by using the 28S rRNA locus as a control. (**C**) Chromatin was isolated from the liver of five adult mice. Chromatin was cross-linked, sheared and immunoprecipitated with the Foxa2-specific antibody described in A and B. The resulting material was uncross-linked, amplified and labeled. Material that was not immunoprecipitated was also amplified and labeled with a different dye. Both sets were hybridized on the Mouse Promoter Chip BCBC 5A. Statistical analysis resulted in a set of 107 genes that are bound by Foxa2.

we generated ROC curves to quantify enrichment of binding sites for each DNA-binding protein. The AUC calculated from the ROC curve of Foxa2 (Figure 2A) was >0.5, confirming that the PWM for Foxa2 is a reasonable approximation of the preferred *in vivo* contact site. Because 300 of the 524 PWMs scanned had an AUC >0.5, we performed 100 permutations of the data to obtain a reference distribution, which we employed to calculate an approximate *P*-value for each ROC curve. Interestingly, the ROC curve shows that the likelihood of finding the Foxa2 consensus sequence (PWM) in the target sequences is only slightly above that of finding it in the background sequences, no matter what cutoff score for the PWM is chosen. The score distribution for Foxa2 in target sequences and background sequences shows that the consensus sequence is frequently found throughout the genome (Figure 2B). In other words, the Foxa2 consensus sequence does not contain sufficient information to predict real *in vivo* occupancy among the background of thousands of unbound promoters with similar sequences, similar to what has been observed previously for other transcription factors (22,23). This finding suggests that Foxa2 *in vivo* binding might be determined by additional sequences elements.

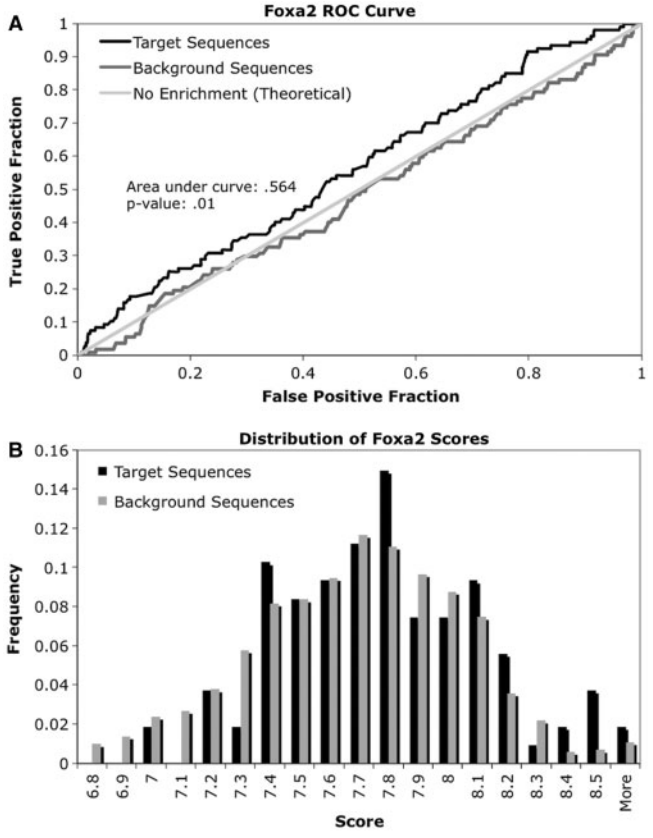

Figure 2. Scanning of Foxa2 PWM from TRANSFAC. (**A**) ROC curve generated from scanning Foxa2 PWM on target (bound in genome-wide location analysis) and background sequences. When scanning PWMs, only the highest scoring match is used for further analysis. The PWM is only slightly enriched in sequences that are bound by Foxa2 (black line). ROC curve generated from scanning a random set of background sequences (dark gray line) is similar to the theoretical line for ROC-based analysis (light gray line). (**B**) Distribution of scores for Foxa2 PWM found in target and background sequences. The Foxa2 consensus is easily found in random sequences, and the match to the consensus sequence does not have to be strong for the factor to bind DNA.

Foxa2 consensus strength and enrichment of other transcription factor binding sites

Enhancers are often composed of binding sites for multiple transcription factors. In the liver, these might include sites for the HNFs, nuclear hormone receptors and others (2,3,9,10). Thus, 'liver-modules' can be made up of binding sites for HNF6, HNF4-α, HNF-1α, Foxa2 etc. In principle, a given enhancer strength of a liver-module could be achieved by combinations of various strong and weak binding sites for multiple factors. We have shown that Foxa2 can bind to sequences that either strongly match the existing consensus or are a weak match. Therefore, we investigated whether the role of additional factors between these strong and weak cases is different. To this end, we compared the enrichment of *cis*-regulatory elements in sequences that have a weak match to the Foxa2 consensus to those sequences that have a strong match by first ordering the Foxa2 target sequences by decreasing Foxa2 PWM score. Starting with the first 10 sequences, which have the highest Foxa2 PWM score, we calculated the AUC for other transcription factor PWMs of interest, focusing on those PWMs that had an AUC >0.5, and a *P*-value ≤0.01 when all target sequences were scanned (Supplementary Table 3). We then iteratively added in the remaining Foxa2 target sequences, one by one, while recalculating the AUC after each addition (Figure 3A). As shown in Figure 3B, the AUC of Foxa2 steadily decreases as sequences with weaker binding sites are added, as expected. Plotting the AUC for other transcription factors against the number of sequences gives us the 'AUC path'. The AUC paths for each of the other transcription factors indicates which TF-binding sites are either enriched more or less as the strength of the Foxa2-binding site decreases. If a particular factor is not dependent on the match to the Foxa2 PWM, its AUC path should remain constant as Foxa2 target sequences are added (Figure 3B). In order to identify factors that show a dependence to the match to the Foxa2 PWM, we calculated the area between the AUC path of each other factor and the AUC path for Foxa2 itself (Figure 3C). The binding sites that are more enriched in the presence of a close match to the Foxa2 consensus have the smallest area between the AUC paths, whereas the factors that are less enriched in the presence of a strong Foxa2 site have the largest area between AUC paths. Strikingly, several of the PWMs analyzed show a relationship between their own enrichment and the PWM score for Foxa2, suggesting that in the liver, Foxa2 containing *cis*-regulatory modules are not assembled at random (Supplementary Table 4). The binding sites for two factors, HNF-1α (Transfac ID: M00132, HNF1_01), which has an important role in hepatic and intestinal gene regulation, and Jun (Transfac ID: M00036, VJUN_01), which is involved in hepatocyte proliferation, showed the strongest dependence on the Foxa2 PWM score (Figure 3C) (24,25). The enrichment of the HNF-1α

PWM decreases as the Foxa2 PWM score decreases, while the enrichment of the Jun PWM increases (Figure 4A and B). To evaluate the significance of this dependence, we permuted the sequence order 100 times, so that it was no longer indicative of the match to the Foxa2 consensus, and repeated the method for calculating the AUC path of the other TFs and the area between AUC paths described earlier. Calculating the area between AUC paths for Foxa2 and all of the permutations for HNF-1α and Jun shows that the observed value obtained when using the actual ordering of sequences by Foxa2 binding site strength is not random (Figure 4C and D). To further validate the dependence of HNF-1α and Jun PWM enrichment on Foxa2 PWM score, we selected background (unbound) sequences that contain the same distribution of Foxa2 PWM scores as the target sequences. As expected, the AUC path for HNF-1α and Jun remains constant as sequences are added, confirming that the dependence of PWM enrichment we see in our target set is not simply the result of sequence composition bias (Figure 4E and F). We also permuted the background sequence order to show that unlike what is seen in the target sequences, when background sequences are ordered by decreasing Foxa2 score the AUC path does not lie on the edges of the permuted AUC paths (Figure 4E and F). While the AUC for Jun calculated using background sequences has an overall similar AUC to that calculated using target sequences, the AUC path is not steadily increasing in the background sequences as it is in the true target sequences (Figure 4F). However, we note that the relationship between the Foxa2 PWM score and the Jun PWM enrichment among the Foxa2 targets determined from genome-wide location analysis is not as strong as the relationship seen with HNF-1α. Interestingly, the de-enrichment of some factors is most dependent on Foxa2 sequences of medium strength (Figure 4G). One such factor is PPAR-γ (Transfac ID: M00512, PPARG_01), which is involved in lipid metabolism and differentiation of adipocytes (26,27). As shown in Figure 4H, when sequences that have the strongest Foxa2 binding sites are removed from analysis, the PPAR-γ enrichment follows a pattern similar to Jun.

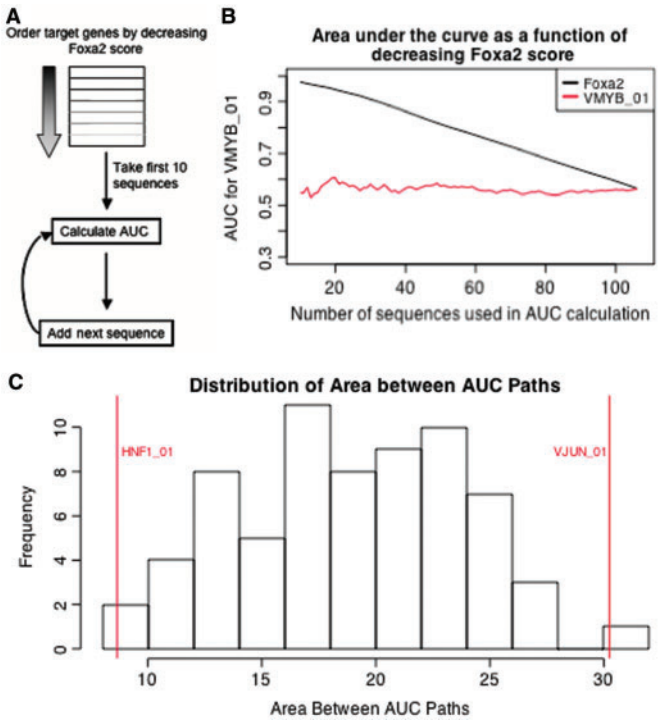

Figure 3. PWMs of other transcription factors are dependent on Foxa2-consensus score. (**A**) Method for calculating AUC path, to identify factors dependent on the Foxa2-consensus score. Sequences were ordered by decreasing Foxa2 score, and after adding one sequence at a time, the AUC for all TRANSFAC PWMs with AUC >0.5 and *P*-value ≤0.01 in the 107 Foxa2-target sequences was calculated. (**B**) Plotting the AUC against number sequences used in the calculation gives the AUC path for a factor. The AUC decreases for Foxa2 as more sequences are added (black line). The AUC for VMYB_01 shows no dependence on the Foxa2-binding site score (red line). (**C**) The area between the AUC path for every factor used in (A) and the AUC path for Foxa2 is plotted. HNF1_01 and VJUN_01 show the smallest and largest area between curves.

Identifying additional TFBS's related to Foxa2 consensus strength

In the previous section, we only investigated the relationship of PWM scores for Foxa2 and the 70 factors that had an AUC of >0.5, and a *P*-value ≤0.01 when using all 107 Foxa2 targets. It is possible that additional binding sites are dependent on the match to the Foxa2 consensus, but did not have a high AUC when all Foxa2 sites were considered previously. To identify these other factors, we split the Foxa2 target sequences into two groups—those that have a strong Foxa2 consensus, and those that have a medium to weak Foxa2 consensus. We explored four possible splits of the data, where the group that was considered to have strong match to the Foxa2 consensus consisted of the top 30, 40, 50 or 60 sequences, and the group with medium/weak binding sites consisted of the remaining sequences for each split. Because we have already demonstrated that HNF-1α is strongly associated with strong Foxa2 binding sites, we chose the grouping of sequences that gave the most dramatic difference in the AUC for HNF-1α, which gave us a grouping of the first 40 sequences as the strong Foxa2-binding sites, and the remaining 67 sequences as the medium/weak Foxa2 binding sites (Supplementary Table 5). These groups of sequences contained a similar overall CG bias and percentage of CpG islands, which was determined using a CpG island searcher, CpGIE (28). Of the sequences with a medium/weak match to the Foxa2 consensus, the average CG bias was 0.021% and 28% of the sequences contained one CpG island. Of the sequences with a strong match to the Foxa2 consensus, the average CG bias was 0.022% and 22.5% contained one CpG island. The difference in CpG island content in the two groups was not significant, as determined using the test of equal proportions. We then scanned all of the vertebrate PWMs in TRANSFAC on both of these sets, again using 'unbound' sequences as background, and found that a different group of transcription factors were enriched in each set. Foxa2 targets closely matching the Foxa2 PWM preferentially contained binding sites for factors such as HNF-1α, HNF6 and CEBP, while the targets matching the Foxa2 consensus only weakly were associated

with PPAR-γ, HNF4-α CREB, Jun and also USF, which regulates genes involved in glucose and lipid metabolism (Supplementary Table 6) (29).

Confirmation of medium/weak Foxa2 targets

Since the match to the Foxa2 consensus sequence in our medium/weak group was relatively poor, we wanted to ensure that these sites were not false positives from our genome-wide location analysis. Therefore, we designed primers around the computationally predicted binding sites for five of these target genes, including Serpinf2, which has the weakest scoring Foxa2 site, and showed by quantitative RT–PCR that these targets were indeed enriched in Foxa2 ChIP samples (Figure 5A). Additionally, we show that neither the predicted binding site score nor the GOMER score, which was calculated using the model that takes potential homotypic interactions into account (30), are related to the fold-change determined from genome-wide location analysis, indicating that the fold-change cutoff chosen does not have any impact on the binding site strength (Figure 5B and C).

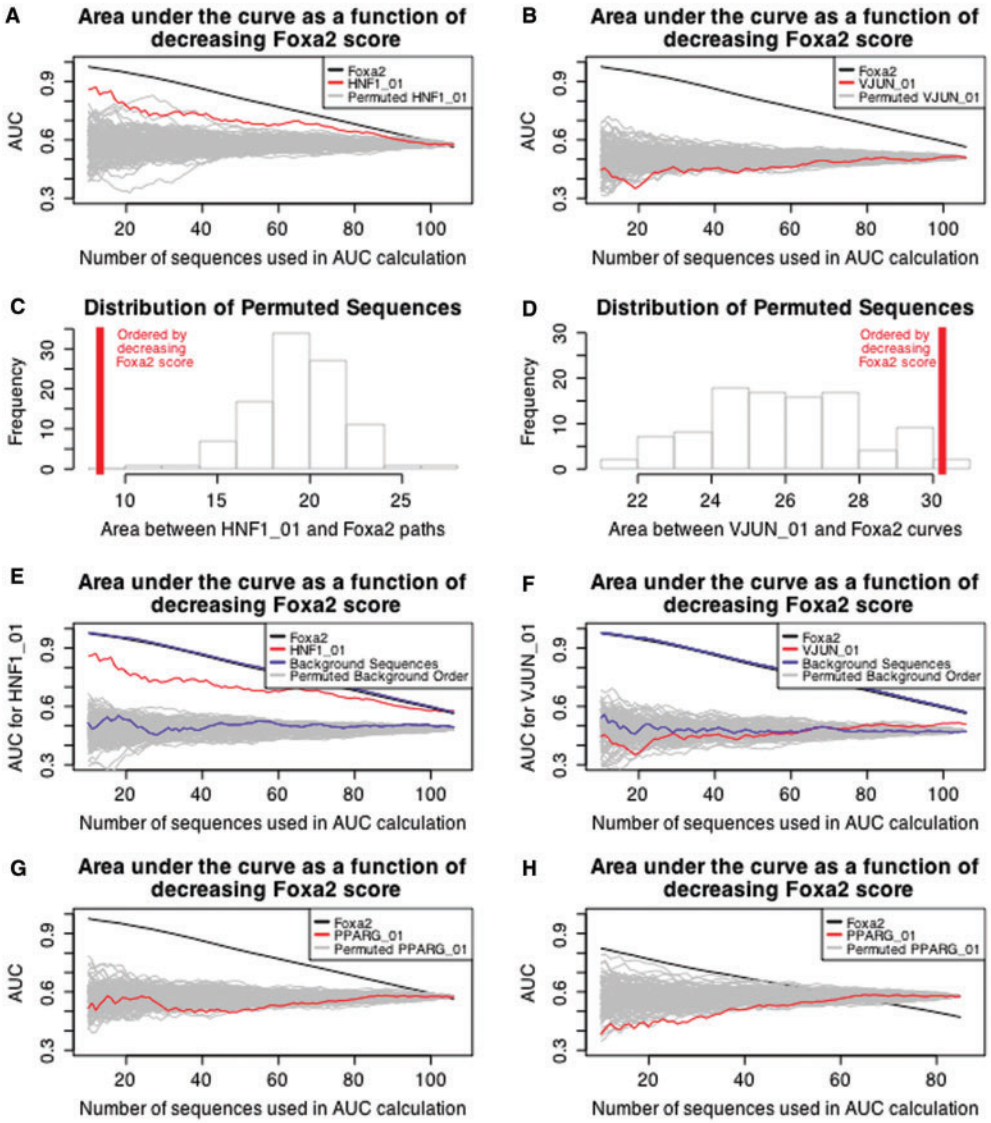

Figure 4. PWMs that show a dependence on Foxa2-binding site score. (**A** and **B**) AUC path for HNF1_01 (A, red) shows that as sequences with decreasing Foxa2 score are added, enrichment for the factor decreases. AUC path for VJUN_01 (B, red) shows that as sequences with decreasing Foxa2 score are added, enrichment for the factor increases. Permuted sequences (not depended on Foxa2 score) are in gray. (**C** and **D**) Distribution of area between paths for permuted sequences when scanning for HNF1_01 (C), VJUN_01(D) and Foxa2. Area between curves for the factors and Foxa2 when sequences are ordered by Foxa2 consensus site strength is in red. (**E** and **F**) The 107 background sequences, which contain Foxa2-binding sites that match the score distribution of binding sites in the Foxa2 targets, were ordered by decreasing Foxa2 PWM score. Permuted sequence order for the background sequences are in gray. AUC paths for HNF1_01 (E) and VJUN_01 (F) were calculated using these sequences (blue line), and no longer show a dependence on Foxa2 PWM score, as they did when the true target sequences are used (red). (**G**) AUC path for PPARG_01 (red) shows this factor is most enriched in the sequences with very strong or very weak Foxa2-binding sites. Permuted sequences (not depended on Foxa2 score) are in gray. (**H**) AUC path for PPARG_01 (red) when the sequences with strongest Foxa2 sites are removed from analysis. Now the factor shows a trend similar to VJUN_01. As expected, the overall enrichment of Foxa2 is lower when the strongest sequences are not included in analysis. Permuted sequences (not depended on Foxa2 score) are in gray.

Varying expression patterns between groups with strong and medium/weak match to the Foxa2 consensus

Because there are different sets of TFBS enriched in sequences with strong Foxa2 PWM match versus those with medium/weak Foxa2 PWM scores, we investigated potential differences in expression patterns between the genes in the two groups. To study gene-expression patterns, we used the Novartis mouse dataset, which profiles gene expression across 91 tissues (31). First, we identified which tissue had the highest expression for each target gene in both the strong and medium/weak Foxa2 consensus groups. In the group with a medium/weak Foxa2 consensus, 34% of the genes are most abundantly expressed in the liver, whereas in the group with a strong Foxa2 consensus, only 14% of the genes are activated most strongly in the liver (Figure 6A and B).

We next investigated the tissue specificity of those genes that are highly expressed in the liver, by first calculating the median expression of each gene across all tissues. We focused on genes that had high expression in the liver, with an expression value more than three times the median. Almost half of the genes in both groups (37% of the genes that have a strong Foxa2 consensus and 45% of the genes that have a medium/weak Foxa2 consensus) were determined to have high expression in the liver. Of these genes, we determined the number of other tissues that also have expression greater than three times the median expression (Figure 6C and D). It is clear from Figure 6C and D that the genes which have high expression in the liver and a medium/weak Foxa2 site are more tissue specific, while the genes that have high expression in the liver and a strong Foxa2 site have high expression in several other tissues.

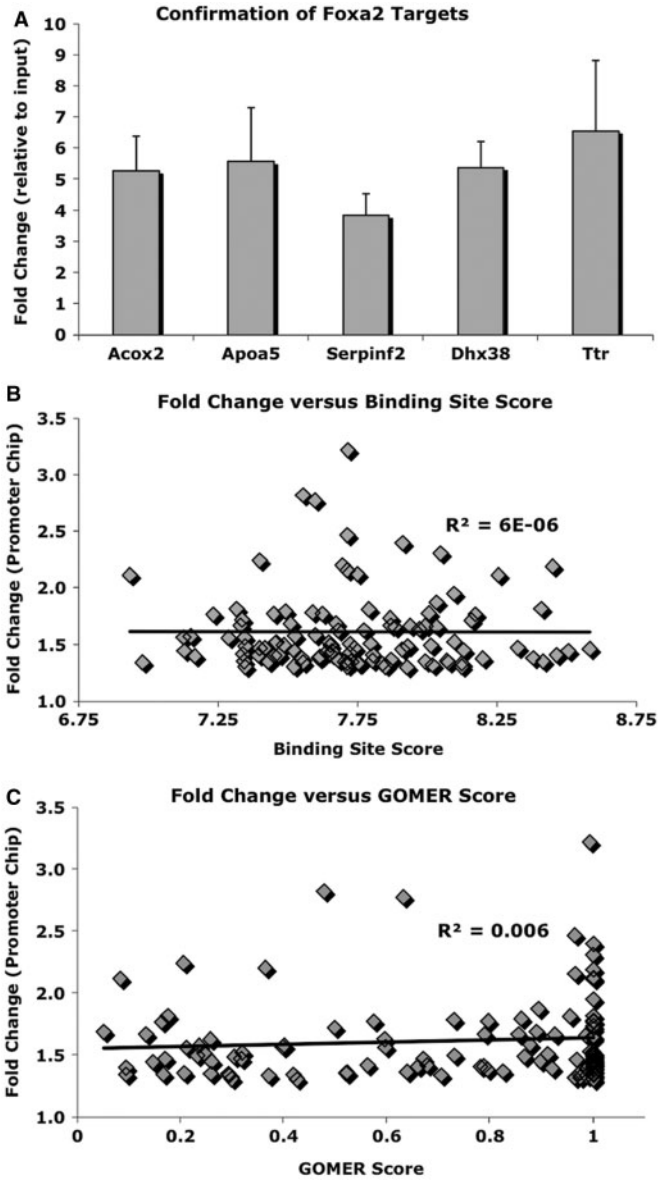

Figure 5. Confirmation of medium/weak Foxa2-target sites. (**A**) Confirmation of medium/weak Foxa2 targets by quantitative RT–PCR. Included in this list is Serpinf2, which had the lowest scoring Foxa2-binding site. (**B**) This plot shows that there is no relationship between fold-change, as determined from the genome-wide location analysis, and binding site score, as determined by scanning the Foxa2 PWM. (**C**) GOMER software was used to calculate a score using the cooperative interaction model to account for homotypic interactions. For a variety of distance parameters, there was no relationship between fold-change, as determined from the genome-wide location analysis and GOMER score (data not shown). Shown here are the GOMER scores using a maximum distance of 50 and minimum distance of 10 for homotypic interactions.

DISCUSSION

We have used a novel methodology to demonstrate the dependence between the strength of a mammalian transcription factor-binding site as determined by its PWM score and the enrichment of other binding sites in the same promoter or enhancer region. We used Foxa2 target regions, as determined by genome-wide location analysis, to show that Foxa2 has the ability to bind DNA even when the sequence to which it is binding is not a strong match to the known consensus. To ensure that the sequences containing weak binding sites were true targets of Foxa2, we confirmed several with quantitative RT–PCR. Additionally, we have shown that when there is a strong match to the Foxa2 consensus, a different set of binding sites for other transcription factors is enriched when compared to those genes that have a medium or weak Foxa2 consensus in the promoter region. The idea that Foxa2 has the ability to bind variations of the consensus sequence with different affinities has been shown previously, using gel shift assays, however, these *in vitro* studies could not explain why Foxa2 does not bind to all 'weak' sites in the genome (32).

The structure of Foxa2 resembles that of histone H5, as shown by X-ray crystallography (33). Foxa2 has been shown to act as a 'pioneer' factor in the case of the albumin enhancer, where Foxa proteins have the ability to bind compacted chromatin and make it more accessible for other transcription factors to bind (34–36). The albumin enhancer contains a site with a strong match to the Foxa2 consensus site, and therefore the question remains whether Foxa2 has the ability to bind compacted chromatin everywhere, or only when strong binding sites

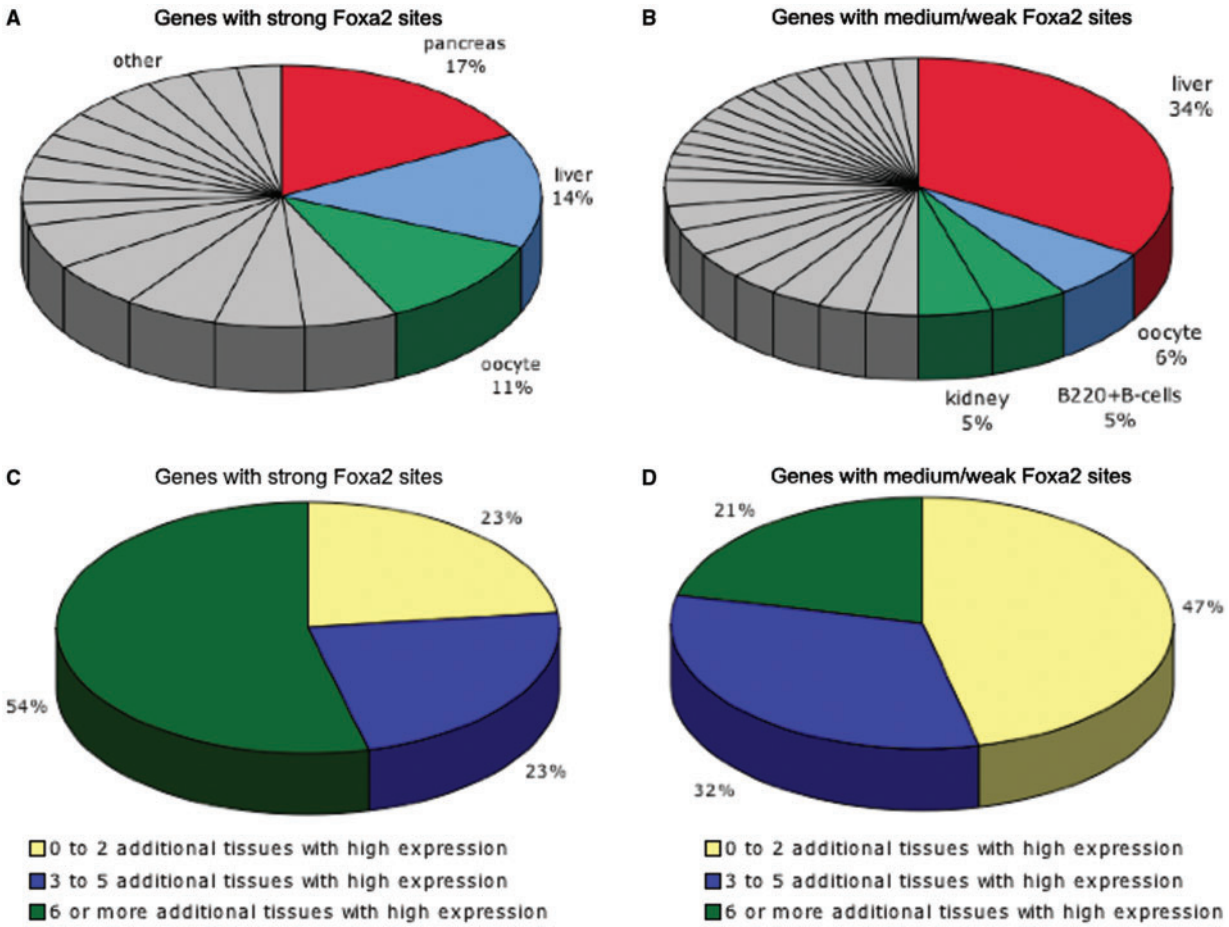

Figure 6. Tissue specificity of Foxa2 targets. (**A** and **B**) Genes with a medium/weak Foxa2-consensus site tend to have highest expression in the liver than any other tissue, whereas genes with a strong Foxa2 consensus are not more highly expressed in a single tissue. The tissues with the highest percentages of genes are labeled. (**C** and **D**) Of the genes highly expressed in the liver, those with a medium/weak Foxa2 consensus are more tissue specific than genes with a strong Foxa2 consensus.

are present. The discovery of a difference in binding site enrichment depending on the strength of the Foxa2 binding sites could indicate that Foxa2 is acting as a pioneer factor only when a strong site is present, but has a different mechanism for binding target genes when a weak site is encountered.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Sridhar Hannenhalli for valuable discussion and advice on this article, and John Brestelli for depositing microarray data into ArrayExpress (accession number E-MTAB-32). This work was supported by National Institutes of Health, Training Grant in Computational Genomics (5-T32-HG000046-09 to G.T.), and National Institutes of Health (2-PO1-DK49210 to K.H.K.). Funding to pay the Open Access publication charges for this article was provided by National Institute of Diabetes and Digestive and Kidney Diseases.

Conflict of interest statement. None declared.

REFERENCES

1. Arnone,M.I. and Davidson,E.H. (1997) The hardwiring of development: organization and function of genomics regulatory systems. *Development*, **124**, 1851–1864.
2. Odom,D.T., Dowell,R.D., Jacobsen,E.S., Nekludova,L., Rolfe,P.A., Danford,T.W., Gifford,D.K., Fraenkel,E., Bell,G.I. and Young,R.A. (2006) Core transcriptional regulatory circuitry in human hepatocytes. *Mol. Syst. Biol.*, **2**, 1–5.
3. Friedman,J., Larris,B., Le,P., Peiris,T., Arsenlis,A., Schug,J., Tobias,J., Kaestner,K. and Greenbaum,L. (2004) Orthogonal analysis of C/EBPbeta targets in vivo during liver proliferation. *PNAS*, **101**, 12986–12991.
4. Le,P.P., Friedman,J., Schug,J., Brestelli,J., Parker,J., Bochkis,I. and Kaestner,K. (2005) Glucocorticoid receptor-dependent gene regulatory networks. *PLoS Genet.*, **2**, 159–170.
5. Rubins,N.E., Friedman,J.R., Le,P.P., Zhang,L., Brestelli,J. and Kaestner,K.H. (2005) Transcriptional networks in the liver: hepatocyte nuclear factor 6 function is largely independent of Foxa2. *Mol. Cell Biol.*, **25**, 7069–7077.
6. Kyrmizi,I., Hatzis,P., Katrakili,N., Tronche,F., Gonzalez,F.J. and Talianidis,I. (2006) Plasticity and expanding complexity of the hepatic transcription factor network during liver development. *Genes Dev.*, **20**, 2293–2305.
7. Rada-Iglesias,A., Wallerman,O., Koch,C., Ameur,A., Enroth,S., Clelland,G., Wester,K., Wilcox,S., Dovey,O.M., Ellis,P.D. *et al.*

(2005) Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Hum. Mol. Genet.*, **14**, 3435.
8. Odom,D.T., Zizlsperger,N., Gordon,D.B., Bell,G.W., Rinaldi,N.J., Murray,H.L., Volkert,T.L., Schreiber,J.R., Rolfe,P.A., Gifford,D.K. *et al.* (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.
9. Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, 1559–1566
10. Yu,X., Lin,J., Zack,D.J. and Qian,J. (2006) Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.*, **34**, 4925–4936.
11. Smith,A.D., Sumazin,P., Xuan,Z. and Zhang,M.Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *PNAS*, **103**, 6275–6280.
12. Smith,A.D., Sumazin,P. and Zhang,M.Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *PNAS*, **102**, 1560–1565.
13. Friedman,J.R. and Kaestner,K.H. (2006) The Foxa family of transcription factors in development and metabolism. *Cell Mol. Life Sci.*, **63**, 2137–2328.
14. Zhang,L., Rubins,N.E., Ahima,R.S., Greenbaum,L.E. and Kaestner,K.H. (2005) Foxa2 integrates the transcriptional response of the hepatocyte to fasting. *Cell Metab.*, **2**, 141–146.
15. Lee,C.S., Friedman,J.R., Fulmer,J.T. and Kaestner,K.H. (2005) The initiation of liver development is dependent on Foxa transcription factors. *Nature*, **435**, 944–947.
16. Divine,J.K., McCaul,S.P. and Simon,T.C. (2003) HNF-1alpha and endodermal transcription factors cooperatively activate Fabpl: MODY3 mutations abrogate cooperativity. *Am. J. Physiol. Gastrointest. Liver Physiol.*, **285**, 62–72.
17. Costa,R.H. and Grayson,D.R. (1991) Site-directed mutagenesis of hepatocyte nuclear factor (HNF) binding sites in the mouse transthyretin (TTR) promoter reveal synergistic interactions with its enhancer region. *Nucleic Acids Res.*, **19**, 4139–4145.
18. Besnard,V., Wert,S.E., Hull,W.M. and Whitsett,J.A. (2004) Immunohistochemical localization of Foxa1 and Foxa2 in mouse embryos and adult tissues. *Gene Expr. Patterns*, **5**, 193–208.
19. Pani,L., Quian,X.B., Clevidence,D. and Costa,R.H. (1992) Restricted promoter activity of the liver transcription factor hepatocyte nuclear factor 3 beta involves a cell-specific factor and positive autoactivation. *Mol. Cell Biol.*, **12**, 552–562.
20. Sund,N.J., Ang,S.-L., Sackett,S.D., Shen,W., Daigle,N., Magnuson,M.A. and Kaestner,K.H. (2000) Hepatocyte nuclear factor 3β (Foxa2) is dispensable for maintaining the differentiated state of the adult hepatocyte. *Mol. Cell Biol.*, **20**, 5175–5183.
21. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
22. Elnitski,L., Jin,V.X., Farnham,P.J. and Jones,S.J.M. (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.*, **16**, 1455–1464.
23. Liu,X., Lee,C.-K., Granek,J.A., Clarke,N.D. and Lieb,J.D. (2006) Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res.*, **16**, 1517–1528.
24. Behrens,A., Sibilia,M., David,J.-P., Möhle-Steinlein,U., Tronche,F., Schütz,G. and Wagner,E.F. (2002) Impaired postnatal hepatocyte proliferation and liver regeneration in mice lacking c-jun in the liver. *EMBO J.*, **21**, 1782–1790.
25. Tronche,F. and Yaniv,M. (1992) HNF1, a homeoprotein member of the hepatic transcription regulatory network. *BioEssays*, **14**, 579–587.
26. Tontonoz,P., Nagy,L., Alvarez,J.G.A., Thomazy,V.A. and Evans,R.M. (1998) PPAR© promotes monocyte/macrophage differentiation and uptake of oxidized LDL. *Cell*, **93**, 241–252.
27. Chawla,A., Schwarz,E.J., Dimaculangan,D.D. and Lazar,M.A. (1994) Peroxisome proliferator-activated receptor (PPAR) ©: adipose-predominant expression and induction early in adipocyte differentiation. *Endocrinology*, **135**, 798–800.
28. Wang,Y. and Leung,F. (2004) An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics*, **20**, 1170–1177.
29. Naukkarinen,J., Gentile,M., Soro-Paavonen,A., Saarela,J., Koistinen,H.A., Pajukanta,P., Taskinen,M.-R. and Peltonen,L. (2005) USF1 and dyslipidemias: converging evidence for a functional intronic variant. *Hum. Mol. Genet.*, **14**, 2595–2605.
30. Granek,J.A. and Clarke,N.D. (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.*, **6**, R87.
31. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *PNAS*, **99**, 4465–4470.
32. Overdier,D.G., Porcella,A. and Costa,R.H. (1994) The DNA-binding specificity of the hepatocyte nuclear factor 3/forkhead domain is influenced by amino-acid residues adjacent to the recognition helix. *Mol. Cell Biol.*, **4**, 2755–2766.
33. Clark,K.L., Halay,E.D., Lai,E. and Burley,S.K. (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature*, **364**, 412–420.
34. Chaya,D., Hayamizu,T., Bustin,M. and Zaret,K. (2001) Transcription factor FoxA (HNF3) on a nucleosome at an enhancer complex in liver chromatin. *J. Biol. Chem.*, **276**, 44385–44389.
35. Cirillo,L.A., Lin,F.R., Cuesta,I., Friedman,D., Jarnik,M. and Zaret,K.S. (2002) Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell*, **9**, 279–289.
36. Shim,E., Woodcock,C. and Zaret,K. (1998) Nucleosome positioning by the winged helix transcription factor HNF3. *Genes Dev.*, **12**, 5–10.