Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays

Alvaro Rada-Iglesias1, Ola Wallerman1, Christoph Koch3, Adam Ameur2, Stefan Enroth2, Gayle Clelland3, Kenneth Wester1, Sarah Wilcox3, Oliver M. Dovey3, Peter D. Ellis3, Vicki L. Wraith3, Keith James3, Rob Andrews3, Cordelia Langford3, Pawandeep Dhami3, Nigel Carter3, David Vetrie3, Fredrik Pontén1, Jan Komorowski2, Ian Dunham3 and Claes Wadelius1,*

1Department of Genetics and Pathology, Rudbeck Laboratory and 2Linnaeus Centre for Bioinformatics, Uppsala University, SE-75185 Uppsala, Sweden and 3Wellcome Trust Sanger Institute, Cambridge, UK

Received July 4, 2005; Revised August 19, 2005; Accepted September 28, 2005

We present a detailed in vivo characterization of hepatocyte transcriptional regulation in HepG2 cells, using chromatin immunoprecipitation and detection on PCR fragment-based genomic tiling path arrays covering the encyclopedia of DNA element (ENCODE) regions. Our data suggest that HNF-4α and HNF-3β, which were commonly bound to distal regulatory elements, may cooperate in the regulation of a large fraction of the liver transcriptome and that both HNF-4α and USF1 may promote H3 acetylation to many of their targets. Importantly, bioinformatic analysis of the sequences bound by each transcription factor (TF) shows an over-representation of motifs highly similar to the in vitro established consensus sequences. On the basis of these data, we have inferred tentative binding sites at base pair resolution. Some of these sites have been previously found by in vitro analysis and some were verified in vitro in this study. Our data suggests that a similar approach could be used for the in vivo characterization of all predicted/uncharacterized TF and that the analysis could be scaled to the whole genome.

INTRODUCTION

Transcriptional control is achieved through a complex interplay between cis-acting regulatory DNA elements (promoters, enhancers, locus control regions, etc.) and trans-acting proteins. The assembly of these proteins to their binding sites displays both synergy and cooperativity, which increases the specificity and flexibility of the process (1). However, at a genomic level, we have very poor knowledge of how transcription regulation is achieved, and most of it comes from the analysis of a limited number of regulatory elements.

Recent experiments have used chromatin immunoprecipitation and detection on genomic microarrays (ChIP-chip) to investigate transcriptional control processes at a large scale. This has revealed that transcription factors (TFs) bind more targets than previously suspected. In some cases, promoter or CpG island arrays have been used, resulting in the identification of numerous genes that could be under the control of particular TFs (2–4). However, this approach cannot detect cis-regulatory elements, which are sometimes located several kilobase from transcription start sites (TSS) (5). Other studies have employed high-resolution tiling path arrays, mainly focusing on chromosomes 21 and 22 (6,7). In these cases, single or not clearly related TFs have been analysed, which do not allow the investigation of transcriptional networks.

This study has two main objectives. First, to increase the knowledge of the complex network of TFs and histone modifications that act on a gene, by deciphering the interactions and...
connections between a set of TFs and histone H3 acetylation. Secondly, to identify the actual base pairs that these TFs are interacting with to exert their regulatory effects. Our aim is to achieve these goals in vivo at a genomic scale, using the human hepatocytes as a model and by studying disease-associated TFs.

Hepatocyte differentiation and metabolism are controlled by ubiquitous and liver-specific TFs. HNF-4α belongs to the nuclear receptor family and is considered to be the major regulator of the hepatocyte phenotype (8). Furthermore, HNF-4α has been associated with both an autosomal dominant form of diabetes, MODY1 (9), and the common form of type 2 diabetes (10). HNF-3β (FOXA2) belongs to the forkhead-domain family and plays a pioneering role in establishing a liver transcriptional hierarchy, e.g. through activation of both HNF-4α and HNF-1α (11). In the adult liver, HNF-3β regulates lipid metabolism and ketogenesis during fasting and diabetes, and its subcellular localization is regulated by insulin (12). The ubiquitous TF USF1 was recently implicated in familial combined hyperlipidaemia, which is characterized by elevated levels of either total serum cholesterol or triglycerides or both (13), making the identification of novel USF1 targets a critical issue. TFs need other proteins known as coactivators (14) to activate transcription, many of which possess histone acetyl transferase (HAT) activity. One of their best characterized targets, histone H3, has been found to be acetylated in lysines 9 and 14 near transcription starting sites in active genes (15,16).

The encyclopedia of DNA elements (ENCODE) project was initiated to evaluate strategies for identifying all functional elements in the human genome (17). To this end, a PCR-based tiling path array covering 1% of the genome was constructed. Using HepG2 cells, we present an exhaustive characterization of hepatocyte transcriptional modules. The in vivo binding sites of HNF-4α, HNF-3β and USF1 and the distribution of acetylated H3 in lysines 9 and 14 (AcH3) were interrogated by ChIP-chip. Some of our major observations were the correlation of HNF-3β and HNF-4α binding sites, indicating cooperativity between these proteins, and the identification of several potential enhancers located far from annotated genes. Although scarce, we found that most of USF1 bindings occurred at proximal promoters, which were usually acetylated on H3. H3 acetylation was generally found near 5’ end of genes, in agreement with recent observations. Most importantly, analysis of the sequences from sites bound by HNF-4α, HNF-3β and USF1 showed that we were able to reliably identify consensus motifs similar to those previously reported. We also inferred tentative individual binding sites at a base pair resolution, and some of these sites have been experimentally verified by us and others. This opens the possibility for similar analyses of all other human TFs. The results have important implications for the strategies to construct a tiling path array over the entire human genome.

RESULTS
Assessing quality of antibodies, ChIP protocol and tissue distribution of proteins
The antibodies against HNF-4α, HNF-3β, USF1 and acetylated H3 showed high specificity by western-blot analysis using HepG2 nuclear extracts (Supplementary Material, Fig. S1). The USF1 antibody recognizes a C-terminal epitope, which occurs in two isoforms (31 and 43 kDa) (18). We detected a band of 70–80 kDa that could correspond to heterodimers of the two isoforms (18). A second antibody against the N-terminal part of USF1 detected only the larger isoform, which is mainly nuclear when compared with the shorter isoform, which is more abundant in the cytoplasm. HepG2 cells were treated with sodium butyrate, a histone deacetylase inhibitor that increases levels of histone acetylation. After 2 h of treatment, a clear increase in the band detected by the anti-acetylated H3 antibody was observed. Furthermore, all antibodies provided consistent results by immunohistochemistry in our tissue microarrays (TMAs) experiments, where the expression pattern of the proteins was analysed in 48 healthy tissues, 20 tumours and 50 cell lines. As expected, ACh3 showed a strong nuclear signal in all cells. The larger USF1 isoform gave a medium-to-strong nuclear staining in most tissues, whereas the antibody detecting both the large and small isoforms gave a signal of varying intensity from both the cytoplasm and nucleus in most tissues. The HNF proteins displayed a more restricted expression pattern and were mostly detected in abdominal organs. HNF-3β had a nuclear localization with mainly median intensity. HNF-4α showed nuclear and some cytoplasmic staining with the strongest signals in liver, pancreas, kidney, stomach and intestine, which could be of importance for future studies of type 2 diabetes (Supplementary Material, Fig. S2 and Table S1).

Several well-characterized enhancers known to bind HNF-4α and/or HNF-3β in HepG2 cells were selected (19–21), and ChIP DNAs were evaluated by PCR (Supplementary Material, Fig. S3a). Clear enrichments were observed in a number of genes for both proteins compared with a negative control, whereas the HNF-1α promoter was only bound by HNF-4α. The ACh3 antibody has been extensively employed for ChIP, and we verified the enrichment in the promoter of HNF-1α. Finally, the USF1 antibody was not pre-evaluated by ChIP, in order to represent a scenario where no previous knowledge is available for a certain TF. This approach was chosen to investigate the potential for using ChIP-chip for identifying in vivo targets of uncharacterized/predicted TFs, and if possible, to establish consensus binding sequences on the basis of in vivo experiments.

Genome-wide localization results
We conducted three independent biological replicates for each of the proteins studied by ChIP-chip. Furthermore, three ChIPs without antibody were analysed, as negative controls. No amplification of the ChIP or input DNA was performed before labelling, to avoid possible bias introduced by this procedure. The reproducibility of our biological replicates was verified by principal component analysis (PCA) (Supplementary Material, Fig. S4).

The sonicated enriched DNA can often hybridize to neighbouring spots in a tiling path array, resulting in positive signals from consecutive spots (Fig. 1). For TFs, the spot with the highest ChIP DNA/input ratio in such blocks was defined as a unique enriched spot (UESs). For ACh3, all enriched spots were generally considered (Materials and Methods).
Genomic regions bound by AcH3 was the most common finding in our experiments (513 enriched spots), followed by binding of HNF-4α (194 UESs) and HNF-3β (154 UESs). Significantly lower numbers were identified for USF1 (31 UESs). When overlaps between the different sets were calculated, a clear co-occurrence in binding between HNF-3β and HNF-4α was observed (40% for HNF-3β, 31% for HNF-4α; P-value: 6.716E−280). Furthermore, most of USF1 sites were heavily acetylated (58%; P-value: 3.523E−220) (Table 1). All UESs, including those for AcH3, were mapped to the closest known gene and can be visualized with the UCSC genome browser using Supplementary Material, File 1.

Data validation

In order to verify the robustness of our genome-wide findings, we selected 8–14 identified targets from each set of ChIP-chip experiments and analysed new independent ChIPs by PCR. We confirmed 11 of 12 UESs for HNF4α, 12 of 14 for HNF3β, eight of eight for USF1 and eight of eight for AcH3 (Supplementary Material, Fig. S3b).

One of the regions included in the ENCODE arrays contains the apolipoprotein C3/A4/A1 cluster, which has been extensively studied in liver and HepG2 cells. Several promoters and enhancers in this region have been characterized in HepG2 cells, mainly through in vitro approaches (22).

We were able to identify all previously known regulatory elements, with the exception of the APOC3 enhancer located 0.8 kb upstream of this gene, because this element was covered by a low quality spot. Importantly, we identified new potential regulatory elements, which usually correspond to regions showing some level of evolutionary conservation. Especially interesting was the identification of new binding sites for HNFs at the proximal promoter of APOA5 and in the intergenic region between APOA5 and APOA4. The latter element could be equally important in APOA4 and
APOA5 transcriptional control. Intriguingly, the 3’ end of the APOC3 gene was found to be occupied by both HNF-4α and HNF-3β (Fig. 2).

Odom et al. (4) identified gene targets of HNF-1α, HNF-4α and HNF-6 in human liver and pancreas, using ChIP-chip with an array comprising 13,000 promoters. They discovered that HNF-4α regulated more genes than previously expected, by binding 12% of the promoters. In agreement with this observation, we have identified a large number of HNF-4α binding sites. However, most of our bindings occur far from any TSS, and only 3% of the Refseq promoters are being directly bound. Of the promoters included in both arrays, we detected 8/11 genes assigned a P-value of < 0.05 in the Odom study. Some of the variation can be probably explained by differences in cells studied, laboratory protocols or statistical evaluation.

The modification of histone H3 through acetylation occurs near a gene’s TSS and is positively associated with gene activity in model organisms (15). Bernstein et al. (16) confirmed these findings in humans, using the same antibody and cells as in our study, and high-resolution tiling path oligonucleotide arrays of chromosomes 21 and 22. Three of the ENCODE regions are located on these chromosomes, and there is a convincing concordance in the results, with 87% of the acetylated sites passing Bernstein’s highest cut-off located within 5 kb of an entry in our data set. Only five of the acetylated sites in our study are > 5 kb from any entry in the Bernstein study. However, Bernstein et al. found three times as many as acetylated spots using the lower cut-off, which could be due to their higher resolution or a higher rate of false positives (Fig. 3).

Identification of consensus sequence and tentative binding sites

Our effort to identify motifs representing consensus binding sequences for a certain TF started by defining a strict set of bona fide enriched spots. We hypothesized that each TF
binding could result in enrichment of several neighbouring spots, as shown in Figure 1. The definition of UES was found to be a critical issue to achieve high quality consensus sequences. Using a motif-finding program (BioProspector) (23) on the sequences of our UES, we reproducibly found a consensus highly similar to the previously established consensus binding sequences for USF1, HNF-4α and HNF-3β (Fig. 4A). The robustness of our consensus sequences is shown by the high similarity between the best motifs obtained in 10 independent BioProspector runs (Fig. 4B). We calculated the probability of obtaining the consensus sequence by chance and found that this was highly unlikely (Fig. 4C). Finally, when similar analysis was performed in randomly selected spot sequences, a motif contained in Alu repeats was consistently obtained. Thus, our data indicate that in vivo ChIP-chip experiments are able to detect consensus sequences similar to those found in the TRANSFAC database (24). To eliminate false positives, we required that the same base pairs should be used to create the consensus in at least five of the 10 iterations using BioProspector. Using these criteria, we inferred the following number of tentative binding sites (TBSs) for the TFs: 159 for HNF-4α, 132 for HNF-3β and 36 for USF1. Our large number of in vivo generated TBS should be compared to the TRANSFAC motifs, on the basis of 32, 24 and 81 in vitro observations separately. The genome-wide mapped locations of the TBS at base pair resolution are presented in Supplementary Material, Table S2.

We were able to confirm several previously established HNF-4α binding sites at base pair resolution, most of them located in the apolipoprotein C3/A4/A1 cluster. We identified TBS in the APOC3 promoter, the APOC3/APOA4 intergenic enhancer and the F10 promoter, which were exactly as defined in the previous studies. Furthermore, we detected one out of two binding sites in the APOA4 promoter, but did not find the reported HNF-3β binding site in the APOA1 promoter. This promoter has been suggested to contain two HNF-4α binding sites, which are relatively different from the established consensus and they were not identified in our study. Instead, we found a TBS in the first intron, close to the TSS (Supplementary Material, Table S3).

Furthermore, electrophoresis mobility shift assay (EMSA) experiments were performed using oligonucleotides designed from TBSs identified by BioProspector. For each of the three proteins investigated, we could confirm that they bound these...
In vivo ChIP-chip experiments can find the consensus binding sequence of a particular TF. (A) Predicted motifs compared with TRANSFAC weight matrices. Logos were created from TBSs for HNF-4α (159 sites), HNF-3β (132 sites) and USF-1 (36 sites), found by BioProspector in at least five out of 10 iterations, when compared with the TRANSFAC matrices M00134 (created from 32 sequences), M00131 (24 sequences) and M00122 (81 sequences). We found at least one TBS in 31 of 31 USF1 targets. Those numbers were 124/154 for HNF-3β and 152/194 for HNF-4α. (B) Our predicted binding motifs are highly stable and reproducible. Consensus logos for the 10 separate runs of BioProspector on the UES for each TF. Each consensus logo is built from all predicted binding sites for the top-scoring motif detected by BioProspector. The set of TBSs are all sequences that are present in at least five out of 10 runs. (C) Significance of the consensus motifs from Figure 3A. Each result from BioProspector is associated with a motif score that is calculated with respect to a null distribution of the score. The red line shows the null distribution, calculated from 100 Monte Carlo simulations by setting the flag ‘–r’ to 100, and the blue crosses show scores for our consensus matrices in each of the 10 iterations. (D) EMSA experiments were performed with oligonucleotides containing TBSs identified by BioProspector in our ChIP-chip data sets. From left to right, the results for labelled HNF-4α consensus (sequence from spot stSG634982), labelled USF1 (sequence from spot stSG627950) and labelled HNF-3β (sequence from spot stSG611299) are presented. In each case, a negative reaction (neg) without nuclear extracts was performed, together with a positive reaction (pos) containing nuclear extracts. The rest of the reactions contained nuclear extracts, together with 100× excess self-competitor (self.comp) (unlabelled consensus oligonucleotide), 100× excess of unrelated competitor (unr.comp) (unlabelled Sp1 consensus), 1 μg antibody against protein of interest or 1 μg of unrelated antibody (Sp1, Ap2α). Specific shifted bands are indicated with filled arrows and supershifted bands with open arrows.
sequences. This indicates that our identified consensus sequences are highly accurate (Fig. 4D). In conclusion, our data suggest that a similar approach can be used for the in vivo characterization of the perhaps 2000 TFs with unknown DNA binding sequences, providing that specific antibodies are available. The results for USF1 illustrate this clearly, because no ChIPs were performed prior to the genomic analysis.

Recent reports indicate that long-range enhancers and proximal promoters are in close proximity in the cellular context, owing to formation of chromatin loops (25,26). When our HNF-4α binding sequences in proximal promoters, defined as within 5 kb from TSS, were analysed with BioProspector, the HNF-4α consensus sequence was not found among the top motifs (Fig. 5A). There could be several explanations for this observation. First, there might be a fraction of HNF-4α binding sites that are different from the established consensus, even though it is difficult to explain why they should be more frequent in promoters. Secondly, HNF-4α might act as a coactivator, interacting with another TF(s) but not with the DNA. Thirdly, it might be the case that some of the HNF-4α interactions with proximal promoters are indirect through formation of enhancer/promoter loops with HNF-4α binding, occurring mainly in distal regulatory elements (Fig. 5B). The motif presented in Figure 5A could be the combined result of these three alternatives, but we favour the last hypothesis, because we identified many binding sites far from proximal promoters. Our data indicate that some of the promoters positive for HNF-4α in the study by Odom et al. might be enriched on the basis of indirect interactions, and this might explain why they found HNF-4α consensus sequences in only 9% of such promoters. It is likely that positive signals are generated in every ChIP-chip study as a consequence of indirect interaction, but we believe that by identifying TBS, it is possible to identify and distinguish some of them.

**HNF-3β and HNF-4α are major regulators of hepatocyte phenotype**

Out of the spots enriched for HNF-4α and HNF-3β, a fraction was enriched for both proteins. The shared spots had higher enrichments (log₂-ratios) than those that bound only one of the two factors (2.37 versus 1.97 for HNF-3β and 2.28 versus 1.83 for HNF-4α), indicating cooperativity in their bindings. This is also suggested by the observed proximity between TBSs for the two proteins. We determined the distances between the TBS for one protein and the closest TBS for the other, and observed a clear over-representation of HNF-4α
and HNF-3β TBS < 1000 bp apart. A majority of these sites was within the typical size of an enhancer, and a trend towards colocalization within 100 bp can be observed (Fig. 6A).

HNF-4α and HNF-3β showed similar genomic distribution of their binding sites, with many of them occurring at long distances from TSS, both upstream and downstream (Fig. 6B and C), as could be expected for enhancer elements. Among the genes closest to the binding sites for these hepatocyte nuclear factors, there were representatives of various gene ontology (GO) (27) biological function categories, but interestingly genes involved in lipid metabolism and transport were significantly over-represented when compared with a random distribution (Supplementary Material, Fig. S5).

**USF1 binds to proximal promoters and is associated with acetylated H3: a model for investigating uncharacterized TFs**

The total number of enriched spots found in the USF1 experiments was considerably lower than that for the other proteins investigated. However, when an enrichment was detected, it was at similar levels as those observed for HNFs, indicating

![Diagram](https://example.com/diagram.png)
that our antibody against USF1 worked efficiently in ChIP (Supplementary Material, Fig. S3b and c). This suggests that USF1 has a restricted number of targets in HepG2 cells, more similar to proteins like HNF1 and HNF6 (4). The genomic distribution of USF1 targets was different from HNF-4α and HNF-3β, because most of USF1 UESs were located in proximal promoters (Fig. 6B and C). It has been recently suggested that USF1 exerts its trans-activation effects through recruitment of coactivators that possess HAT activity (28). In agreement with this, the overlap between USF1 binding and AcH3 (58%; P-value: 3.523E-20), as well as the mean level of acetylation of USF1 targets (log2-ratio = 1.56), was higher than the overall level of AcH3 (normalized log2-ratio = 0).

As previously stated, we started our ChIP-chip experiments for USF1 without including the previously known positive controls. In spite of this, the newly identified USF1 UESs were verified to be the bona fide targets of this protein, by various methods. First, some of the new targets were confirmed as true positives by PCR analysis of ChIP DNA (Supplementary Material, Fig. S3b). Secondly, the same targets were analysed by PCR analysis of ChIP DNA obtained using a second antibody against USF1, which confirmed all tested enriched spots (Supplementary Material, Fig. S3c). Thirdly, the sequence analysis of USF1 UES resulted in the identification of an over-represented motif highly similar to the previously suggested consensus binding sequence (Fig. 5A). This suggests that the same strategy could be used in the case of completely uncharacterized TFs. Immunohistochemistry can be used on TMA to determine the cell type/tissue expressing the TF, and western blot can characterize the specificity of the antibody. ChIP-chip can then be performed to identify in vivo targets of the protein, as well as provide a predicted consensus binding sequence on the basis of in vivo experiments and TBSs. Analysis of GO of genes near binding sites may give indications of which biological processes the TF is regulating.

Histone H3 acetylation is a histone modification that frequently occurs near TSS and is associated with TFs binding sites

Our genome-wide identification of regions acetylated in histone 3 confirms that this is a common modification. In general, the genomic location of AcH3 displayed a clear preference for regions near TSS (Fig. 6B and C). Regions immediately downstream of TSS displayed the highest levels of AcH3 (Supplementary Material, Fig. S6). Our results are in agreement with recent reports both in humans and other eukaryotic organisms (15,16) but extend the knowledge of this distribution to previously uncharacterized regions. Most of the USF1 targets were acetylated in histone 3 (58%; P-value: 3.523E-20). This was also true for a relatively high number of HNF-4α (31%; P-value: 4.353E-44) and to a lesser degree HNF-3β UES (14%; P-value: 7.551E-10). Interestingly, when the AcH3 log2-ratios for HNF targets were investigated, the highest level was found for unique HNF-4α UES (1.18), followed by shared HNF-4α/HNF-3β UES (0.93), whereas acetylation levels for unique HNF-3β UES (0.50) was almost half when compared with the other

Figure 6. Characteristics of the identified protein binding sites. (A) Distances between HNF-4α TBS and nearest HNF-3β TBS (Supplementary Material, Table S2) were calculated, and the cases where such distance was <5 kb are presented in windows of 100 bp. (B) Distribution of UES compared with the UCSC ‘known genes’. Each spot’s midpoint was used to calculate the distance to the closest gene (5’ or 3’) within the boundaries given; spots >-20 kb from any TSS and >+5 kb from any 3’ end were grouped as intergenic. For AcH3, UESs were considered for calculating distances to known genes. (C) Distribution of the ‘intergenic’ UES from (B) around entries in the GenScan data set. Although not as obvious as for the curated data set, an over-representation of acetylation in proximity of TSS and within the predicted genes can be seen.
two groups. These observations suggest that coactivator recruitment (HAT) is a trans-activating mechanism employed by HNF-4α, but not by HNF-3β.

DISCUSSION

After the completion of the human genome sequence, attention has been directed towards determining the function of non-coding sequences. High levels of evolutionary conservation have been observed in numerous non-coding elements, and it has been proposed that a major function of them may be to regulate gene activity (29). To determine which TFs that bind to each conserved and possible non-conserved element in every tissue is a long-term goal in biology. The ultimate map would contain binding sites at base pair resolution, but currently we are far from this and mostly have to resort to comparison to in vitro characterized consensus sequences. Therefore, in vivo identification of TF targets and refinement of consensus sequences are important aims in the post-sequencing era. The major finding of our work is the development of an in vivo strategy to define consensus motifs for USF1, HNF-4α and HNF-3β. These motifs are highly similar to the previously in vitro established consensus binding sequences, which confirm that most of our UESs are bona fide binding sites for the investigated proteins. In addition, the finding of over-represented motifs opens the possibility to infer the exact base pair where in vivo DNA–protein interactions occur in a particular cis-regulatory element. We present the putative location, at base pair resolution, of suggested DNA–protein interactions for each identified TBS. Some of these predicted binding sites have been experimentally confirmed by us and others. Further, analysis combined with more powerful statistical and bioinformatic approaches will improve these predictions. These novel binding sites and nearby SNPs may be further investigated, e.g. by genetic analysis in diseases affecting glucose, lipid or cholesterol metabolism.

Our data has important implications for strategies to construct high-resolution arrays covering the whole genome. The resolution in a ChIP-chip experiment is determined by the size of the sonicated enriched DNA hybridized to the array and by the size of the array elements (Fig. 1). The DNA in our experiments was sonicated to a range between 500 and 2000 bp, and the average size of the spots in this array is 1100 bp. We have shown that when applying optimized protocols and strict statistical evaluation, we can identify consensus sequences and TBSs at base pair resolution. This means that a tiling path array at 1000 bp resolution may be enough to map many of the binding sites for sequence-specific TFs. Such an array may contain around 2 000 000 elements for the whole human genome. For these purposes, the high-resolution arrays with 51 874 388 probes covering the genome at 46 bp resolution (30) and 74 180 611 probe pairs covering 30% of the genome at 5 bp resolution (31) may not be necessary. Future experiments will determine which arrays give the best optimization between cost and resolution.

The second major goal of our study is to understand how a tissue-specific transcriptional program is constructed, by analysing the interconnections between different TFs and epigenetic modifications. Tissue-specific gene expression is under the control of several ubiquitous and tissue-specific TFs, and liver is the best-characterized mammalian tissue in this aspect. HNF-4α and HNF-3β play major roles in liver function and differentiation (8,12), and their collaborative relationship in the transcriptional control of several liver genes is well established (32). In concordance with these observations, we found that HNF-4α and HNF-3β are common binders across the genome, with a high similarity in their binding patterns and many of their targets located far from known genes. Proximity between their TBSs further suggests regulatory cooperation. Despite these similarities, individual HNF-4α UESs displayed higher AcH3 levels than those occupied by HNF-3β only, indicating that HNF-4α acts more often through recruitment of HATs. These results together with previous knowledge about the two proteins suggest a sequential and cooperative model of transcriptional control. HNF-3β is already detectable in the early gastrula, playing a major role in visceral endoderm differentiation. HNF-3 proteins occupy the albumin enhancer early in development, even before the gene is activated (33). They are capable of binding directly to chromatin, thereby initiating chromatin opening events and are therefore considered ‘pioneer’ factors (34). Applied to liver transcriptional control, we could hypothesize that HNF-3β, which is expressed at a very early stage, creates chromatin marks at cis-regulatory elements, making them accessible to other TFs, sequentially expressed during development. Among these TFs, HNF-4α has been shown to establish a myriad of protein–protein interactions including those with other important TFs (35–39), several coactivators (40,41), some with HAT activity and members of the RNA PolII machinery (42). All these interactions identify HNF-4α as a good example of a protein present in enhancosomes (Supplementary Material, Fig. S7).

In our genome-wide interrogation for USF1 targets, we observed a clear preference of USF1 binding to proximal promoters, and furthermore, a high correlation with AcH3. This is in full agreement with a recent study, where West et al. (28) reported that USF proteins interact with histone modifying enzymes (Set7/9, PCAF, p300/CBP) and promotes H3K4 methylation and AcH3. We found that USF1 binding was much less common than bindings of HNFs, even though the enrichment levels in its targets were comparable to those of HNFs. This might be explained by the fact that our HepG2 cells were cultivated in a constant relatively low glucose (2 g/l) medium. Several reports suggest that USF1 and USF2 are involved in promoting the liver glucose response, in which a change from low- to high-glucose conditions activates certain genes (43). In this respect, further experiments should elucidate how different stimuli, resembling the metabolic stress encountered in certain diseases, e.g. familial combined hyperlipidaemia and type 2 diabetes with the accompanying metabolic syndrome, modulate the hepatocyte transcriptome.

In conclusion, we have investigated some aspects of how liver-specific gene expression is achieved and highlighted some of the mechanisms implicated. Furthermore, our in vivo genome-wide identification and inference of TF binding sites should serve as an example of how ChIP-chip technology can be used in the identification of targets and consensus
binding sequences for uncharacterized TF. This could dramatically increase our understanding of the complex transcriptional networks acting on a cell, where the role of most key players still remains elusive.

**MATERIALS AND METHODS**

**Cell culture and nuclear extracts preparation**

HepG2 cells were grown in RPMI-1640 medium (Sigma-Aldrich), supplemented with 10% FBS (Gibco, Invitrogen), 1% PEST (Gibco, Invitrogen) and 1% glutamine (Gibco, Invitrogen), at 37°C with 5% CO2. For nuclear extract preparation, cells were treated with cell lysis buffer for 10 min on ice. Nuclei were resuspended in 1× RIPA buffer for 10 min on ice. Samples were centrifuged, and supernatants were kept at −70°C until used.

**Western blot and antibodies**

HepG2 nuclear extracts were separated on NuPAGE 4–12% Bis–Tris gel (Invitrogen) and transferred to polyvinylidene fluoride membrane (Amersham Biosciences), which was developed using ECL Advance Western Blotting Detection Kit (see manufacturer’s instructions) (Amersham Biosciences). Antibodies against USF1 (C-20 and H-86) and HNF-3β were purchased from Santa Cruz Biotechnology; antibody against HNF-4α was purchased from Active Motif and antibody against AcH3 was purchased from Upstate Biotechnology.

**Chromatin immunoprecipitation**

HepG2 cells were grown as described earlier. Around 10^8 sub-confluent cells were used per ChIP experiment. Cells were crosslinked with 0.37% formaldehyde for 10 min and resuspended in cell lysis buffer for 10 min on ice. Nuclei were resuspended in 1× RIPA buffer and kept on ice for another 10 min. Chromatin was sonicated to a size of 0.5–2 kb and pre-cleared by incubating with protein G-agarose (Roche) for at least 1 h at 4°C with slow rotation. At this step, a fraction of the pre-cleared chromatin was kept as input DNA and the rest was incubated with 10 μg antibody at 4°C overnight, and 100 μl of protein G-agarose were used for each ChIP reaction. Protein G-agarose was washed four times with 1× RIPA buffer, once with ChIP washing buffer and once with 1× TE buffer. DNA–protein complexes were eluted, treated with RNaseA (Amersham Biosciences) and incubated at 65°C for 6 h in order to reverse crosslinks. Proteins were degraded by Proteinase K (Amersham Biosciences), and DNA was extracted by phenol/chloroform/isoamyl alcohol extraction, purified and resuspended in water.

**Microarray construction: primer design, PCR reactions and arraying**

Each array element was generated by PCR using specific primer pairs tiling through all ENCODE regions. Primers were selected, so that the resulting amplicons were 1–1.5 kb long, minimally overlapping and were allowed to contain repetitive elements. Gaps in the tiling array were filled with relaxed parameters chiefly allowing 180 bp to 1.5 kb and 30–70% GC. All the forward primers are normalized to the same length to format for Illumina synthesis. For each array element, one PCR reaction was performed. As a template, we used mainly genomic DNA (Roche/Sigma). A minor set of amplicons was amplified with BAC, PAC or fosmid DNA. Typically, the failure rate was <20%. Failed PCR reactions were repeated. Spotted buffer was added to the PCR products at a final concentration of 250 mM sodium phosphate, pH 8.5, 0.00025% sarkosyl, followed by spin filtration using 96-well filtration plates (Millipore). These array elements were printed without any further purification onto activated amine-binding slides (Codelink, Amersham), using a Biobotics TAS arrayer with a 48-pin tool. Most array elements are printed once onto each slide (about 19 000 spots/slide), only X-chromosomal regions (ENm006 and ENr324) were printed in duplicate. The final array presents a 75% coverage of the ENCODE regions.

**DNA labelling and microarray hybridization**

The DNA obtained from a single ChIP reaction was labelled with Cy5, and a fraction of the total input was labelled with Cy3 (1/5 of total input DNA for HNF-4α, HNF-3β, USF1 and no-antibody samples and 1/3 of total input for acetylated H3). For labelling reactions, the Bioprime Labelling system (Invitrogen) was used. Labelled DNA was purified using Amersham G50 columns. ChIP/Cy5 and total input/Cy3 DNAs were combined and ethanol precipitated together with human Cot-1 DNA, and the resulting pellet was resuspended in hybridization buffer. The arrays were pre-hybridized with human Cot-1 and salmon-sperm DNA, followed by addition of the hybridization solution containing the labelled DNAs. The arrays were then washed, dried and scanned in a GenePix 4000 B scanner (Axon instruments, Molecular Dynamics).

**Microarray data analysis, motifs discovery and GO categories**

The computational data analysis was divided into three major parts. The first performed in the LCB-Data WareHouse (http://www.lcb.uu.se/lcbdw.php), where data was pre-processed through various spot-filters and normalizations. Visualization through PCA was used to assert the data quality of each replicate before and after pre-processing. In addition, a log-odds (B-score) for differential enrichment with respect to the negative control was calculated using an empirical Bayes method (44). Each spot then becomes associated with four B-scores, which represent the probability of it being enriched by USF1, HNF-3β, HNF-4α and/or AcH3, respectively. Empirically, spots were considered as enriched when B-score is >0 and log2-ratio >1.25.

To determine the overlaps in binding between the different TFs and AcH3, UESs were used for TFs and total enriched spots for AcH3. In all cases, only bindings occurring in the same spot were taken into consideration. The P-values for the overlaps were calculated under the hypothesis that the size of the overlaps is hypergeometrically distributed, which
would be the case if selection of UES was done at random. The background used for each comparison is taken to be the common spots of the two proteins tested before the selection of UES was made.

Motif discovery was done in several steps. The first, written in R (45), consisted of detecting enriched spots by both the log$_2$-ratio and the B-score. Then, a set of UESs among all enriched spots was created by filtering out adjacent spots with lower log$_2$-ratios. For the AcH3 data set, all enriched spots were generally counted, because longer DNA sequence can be bound by this modified histone, and UESs were only considered for calculating distances to the closest genes. To identify the binding sites, the corresponding DNA sequences were analysed using BioProspector. As BioProspector is non-deterministic, we repeated the analysis and kept all binding sites occurring in each top scoring motif to generate a set of candidates. From these, a set of TBS was obtained by selecting those present in at least five out of 10 runs. The motif logos were created using the WebLogo (46) service.

ChIP verification
A new set of ChIPs was performed with (ChIP DNA) and without antibody (no-Ab DNA) to serve as templates for PCR verification of newly identified binding sites. PCRs were performed using the same volume from ChIP DNA, no-Ab DNA and a dilution of the input (generally 1/30). Different numbers of cycles (25–35) were used for the different primers tested in order to determine the conditions where linear amplification occurs. Enrichment was scored visually by comparing the PCR amplification from ChIP DNA to no-Ab and input DNAs.

Electrophoretic mobility shift assay
HepG2 nuclear extracts were prepared as previously described (47). Nuclear extracts were mixed with binding buffer, poly(dl–dC) and $^{32}$P-labelled probe. For competition reactions, a certain excess of unlabelled probe was added to the mixture. For supershift assays, 1 μg of antibody was used.

TMA design and immunohistochemistry
The TMAs were designed as described previously (48). A spectrum of 48 normal tissues, 20 different cancers and 50 cell-lines were sampled. Immunohistochemistry was done according to the instructions from the manufacturer of the EnVision kit® (DAKO Cytomation, Glostrup, Denmark) using an automated immunostaining instrument, Autostainer Plus® (Dako Cytomation).

SUPPLEMENTARY MATERIAL
Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS
We thank Ulf Landegren for critical reading of the manuscript. This work was supported by the Swedish Research Council, the Wellcome Trust, the US National Human Genome Research Institute (grant no. 5 U01 HG003168), the Markus Borgström foundation and Knut and Alice Wallenberg Foundation.

Conflict of Interest statement: The authors declare that they have no competing financial interests.

REFERENCES


