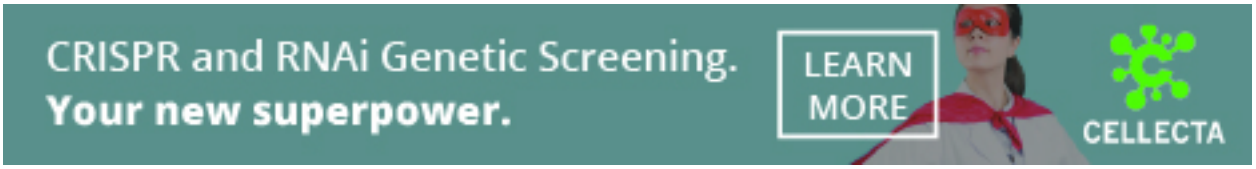
Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders

Alvaro Rada-Iglesias, Adam Ameur, Philipp Kapranov, et al.

Genome Res. 2008 18: 380-392 originally published online January 29, 2008
Access the most recent version at doi:10.1101/gr.6880908



Letter

Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders

Alvaro Rada-Iglesias,[1,5,6] Adam Ameur,[2,5] Philipp Kapranov,[3] Stefan Enroth,[2] Jan Komorowski,[2,4] Thomas R. Gingeras,[3] and Claes Wadelius[1,7]

[1]*Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, SE-75185 Uppsala, Sweden;* [2]*Linnaeus Centre for Bioinformatics, Uppsala University, SE-75185 Uppsala, Sweden;* [3]*Affymetrix, Inc., Santa Clara, California 95051, USA;* [4]*Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, 02-106 Warsaw, Poland*

Transcription factors and histone modifications are crucial regulators of gene expression that mutually influence each other. We present the DNA binding profiles of upstream stimulatory factors 1 and 2 (USF1, USF2) and acetylated histone H3 (H3ac) in a liver cell line for the whole human genome using ChIP-chip at a resolution of 35 base pairs. We determined that these three proteins bind mostly in proximity of protein coding genes transcription start sites (TSSs), and their bindings are positively correlated with gene expression levels. Based on the spatial and functional relationship between USFs and H3ac at protein coding gene promoters, we found similar promoter architecture for known genes and the novel and less-characterized transcripts human mRNAs and spliced ESTs. Furthermore, our analysis revealed a previously underestimated abundance of genes in a bidirectional conformation, where USFs are bound in between TSSs. After taking into account this promoter conformation, the results indicate that H3ac is mainly located downstream of TSS, and it is at this genomic location where it positively correlates with gene expression. Finally, USF1, which is associated to familial combined hyperlipidemia, was found to bind and potentially regulate nuclear mitochondrial genes as well as genes for lipid and cholesterol metabolism, frequently in collaboration with GA binding protein transcription factor alpha (GABPA, nuclear respiratory factor 2 [NRF-2]). This expands our understanding about the transcriptional control of metabolic processes and its alteration in metabolic disorders.

[Supplemental material is available online at www.genome.org. The microarray data from this study have been submitted to ArrayExpress under accession no. E-TABM-314.]

Familial combined hyperlipidemia (FCHL) is the most common lipid metabolism disorder, affecting 1%–2% of the general population, and is observed in 20% of individuals with premature coronary heart disease. Affected cases have high serum cholesterol and/or triglycerides levels (Shoulders and Naoumova 2004). Recently, intronic SNPs in the *USF1* gene were associated with FCHL (Pajukanta et al. 2004; Coon et al. 2005; Huertas-Vazquez et al. 2005), the metabolic syndrome (Ng et al. 2005), and the risk of cardiovascular disease (Komulainen et al. 2006). USF1 and USF2 belong to the basic helix-loop-helix leucine zipper family of transcription factors (TFs). They have a conserved C-terminal domain responsible for dimerization and DNA binding and a more variable N-terminal domain responsible of transactivation. USF1 and USF2 can form both homo- and heterodimers, although the USF1-USF2 heterodimer is the major DNA binding form, which recognizes the canonical E-box sequence CACGTG. Key genes in lipid and carbohydrate metabolism are known to be regulated by USF1 and/or USF2 (Shoulders and Naoumova 2004; Corre and Galibert 2005).

Recent studies have shown that a large portion of the genome is actively transcribed and that the number of promoters exceeds the number of genes (Carninci et al. 2005, 2006; Cheng et al. 2005; Kapranov et al. 2007). Improvements in array technology have allowed genome-wide studies of protein–DNA interactions to be performed (Kim et al. 2005, 2007; Carroll et al. 2006; Lee et al. 2006; Yang et al. 2006). By use of this strategy, all the sites in the genome where a TF bind can be identified, irrespective of whether it is in a known promoter of a protein coding gene (PCG), upstream of a less-characterized transcript, or in an enhancer or other distal regulatory element.

Transcription factors interact with DNA in a chromatin context, where most of the transcription factor binding sequences are occupied by nucleosomes. Through post-translational modifications of histones and/or nucleosome remodeling, some of these sites can become accessible (Li et al. 2007). Once bound to DNA, some transcription factors can recruit proteins that further modify and/or remodel histones (Li et al. 2007). Despite this mutual relationship between TFs and histones, most large-scale in vivo studies in humans have separately investigated these components of transcriptional regulation (Bernstein et al. 2005; Kim et al. 2005, 2007; Carroll et al. 2006; Lee et al. 2006; Yang et

[5]These authors contributed equally to this work.
[6]Present address: Endocrinology Unit, Hospital Clinic, Institut d'Investigacions Biomediques August Pi i Sunyer, Barcelona 08036, Spain.
[7]Corresponding author.
E-mail Claes.Wadelius@genpat.uu.se; fax 46-18-4714808.
Article published online before print. Article and publication date are online at http://www.genome.org/cgi/doi/10.1101/gr.6880908.

18:380–392 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

al. 2006). In a pilot study we mapped the binding sites for USF1 and H3ac in 1% of the genome in the liver cell line HepG2 using ChIP-chip (Rada-Iglesias et al. 2005). We found that most USF1-bound regions were close to PCG TSSs and enriched in H3ac. However, the limited resolution of the arrays did not allow clear structural insights into the identified regulatory elements. Compared with yeast (Yuan et al. 2005; Segal et al. 2006), most ChIP-chip studies in mammalian organisms to date provide a limited structural view of promoters and other regulatory sequences (Ozsolak et al. 2007).

In the present study, we have generated a detailed view of USF1, USF2, and H3ac binding profiles in HepG2 cells by analyzing the whole human genome using oligonucleotide arrays with 35-bp resolution and ChIP DNA fragmented to ~300 bp. The combination of high-resolution data and a large number of bound loci allowed us to understand the spatial and functional relationship of these proteins. Since most binding events occurred in proximity of TSS, we present important features of human promoter architecture, for both novel and well-characterized genes, including a previously underestimated abundance of genes with a bidirectional conformation, and a set of candidate genes for FCHL.

Results

Study design

Extensive experiments were performed in this and previous studies (Rada-Iglesias et al. 2005) in order to establish the specificity and suitability of all the antibodies used for ChIP (Supplemental Fig. S1). ChIP was performed in HepG2 cells in three biological replicates for USF1, USF2, and H3ac. We also performed ChIP using IgG as a negative control (Supplemental Figs. S2, S3). ChIP DNAs and input DNAs were sonicated to an average size of 300 bp to generate sharply defined peaks of enrichment and were then hybridized to a genomic tiling seven-array set representing the nonrepetitive part of the human genome at 35-bp resolution, yielding >700,000,000 data points. We have developed a method to find regions with signals significantly higher than the null distribution, i.e., ChIP using IgG, by applying a statistical modeling approach (Smyth 2004; Rada-Iglesias et al. 2005). The quality of the ChIP-chip experiments was extensively validated, including comparisons to previous data as explained in the Supplemental material (Supplemental Figs. S4–S6; Supplemental Tables S1, S2). A summary of the ChIP-chip results is presented in Table 1. Most of our analyses were performed on the stringent results data sets, which we estimated to contain <1% false positives. A parallel relaxed data set was created to compensate for false negatives in the stringent data set.

USFs preferentially bind to TSS of active genes

Overall, USF1 and USF2 shared most of their binding sites (Table 1; Supplemental Fig. S7), in agreement with previous estimates indicating that USF1-USF2 heterodimers represent 76% of the binding activity in HepG2 cells (Viollet et al. 1996). Furthermore, previous studies have shown that most enriched regions for USF1 (Rada-Iglesias et al. 2005) and H3ac (Bernstein et al. 2005) are close to TSS. We therefore annotated our genome-wide data against a hierarchy of transcript annotations ranging in the degree of experimental support as previously described (Harrow et al. 2006).

Concerning well-annotated PCGs, i.e., known genes/Ensembl/RefSeq, 41% and 50% of USF1 and USF2 targets were within 1 kb of a TSS (Fig. 1A,B). When considering also human mRNAs, spliced ESTs, and human ESTs, the proportion of binding sites within 1 kb of a TSS increased to 75% for USF1 and 81% for USF2, suggesting that these less-characterized transcripts have bona fide promoters. Finally, USFs enriched regions were mapped relative to transfrags generated in HepG2 cells and detected on high-resolution tiling arrays (Kapranov et al. 2007). Seventy-two percent of USF1 and 70% of USF2 bindings not assigned to a TSS were within 1 kb of such transcripts, indicating that they could represent undetected TSSs (Supplemental Fig. S8). Short RNAs (<200 bp) detected on the high-resolution arrays identified a new class of short RNAs called promoter associated short RNAs (PASRs) (Kapranov et al. 2007) generated on both strands around the TSS of well-annotated active genes. The position and frequency profiles of these short RNAs in USF-bound regions further confirmed that most USFs bind close to TSSs (Supplemental Fig. S8).

For H3ac we found that 53% of enriched regions were within 1 kb of a PCG TSS and 91% were within 1 kb of a TSS when human mRNA/spliced EST/human ESTs were also considered (Fig. 1C). When combining the information, we found that a significant fraction of USF1 (45%) and USF2 sites (56%) were acetylated in H3 (Table 1). The colocalization of the combined signals was clearly centered to PCG TSS; e.g., 69% of regions positive for USF1 and H3ac were within 1 kb of PCG TSS and 93% were within 1 kb when all transcript annotations were considered (Fig. 1D).

This suggested that USFs frequently bind to promoters of transcriptionally active genes. Therefore, we collected HepG2 expression data for ~18,000 genes and compared them with our ChIP-chip signals within 1 kb of their TSS (Supplemental Methods). As shown in Figure 1E, PCG, whose promoters were bound by USF1 or USF2, and especially by H3ac, showed significantly higher mRNA expression than the overall levels ($P < 10^{-70}$ in all cases). Finally, we observed that not only are USFs and H3ac frequently bound to promoters of highly expressed genes, but

Table 1. Overall number of bound regions and overlaps between data sets

	Stringent USF1 (2518)	Relaxed USF1 (3771)	Stringent USF2 (1351)	Relaxed USF2 (4206)	Stringent H3ac (10,900)	Relaxed H3ac (17,443)
Stringent USF1			1058 (42%)	1800 (71%)	1145 (45%)	1442 (57%)
Relaxed USF1			1049 (28%)	2107 (56%)	1588 (42%)	2039 (54%)
Stringent USF2	984 (73%)	1029 (76%)			751 (56%)	908 (67%)
Relaxed USF2	1736 (41%)	2094 (50%)			1828 (43%)	2267 (54%)

Regions in the left column are compared with regions on the top row. The window size was 1 kb for comparisons between USFs data sets or 2 kb when comparing USFs with H3ac. Significance of the different overlaps were calculated assuming a hypergeometric distribution of the different data sets and for all overlaps, $P < 1 \times 10^{-308}$.

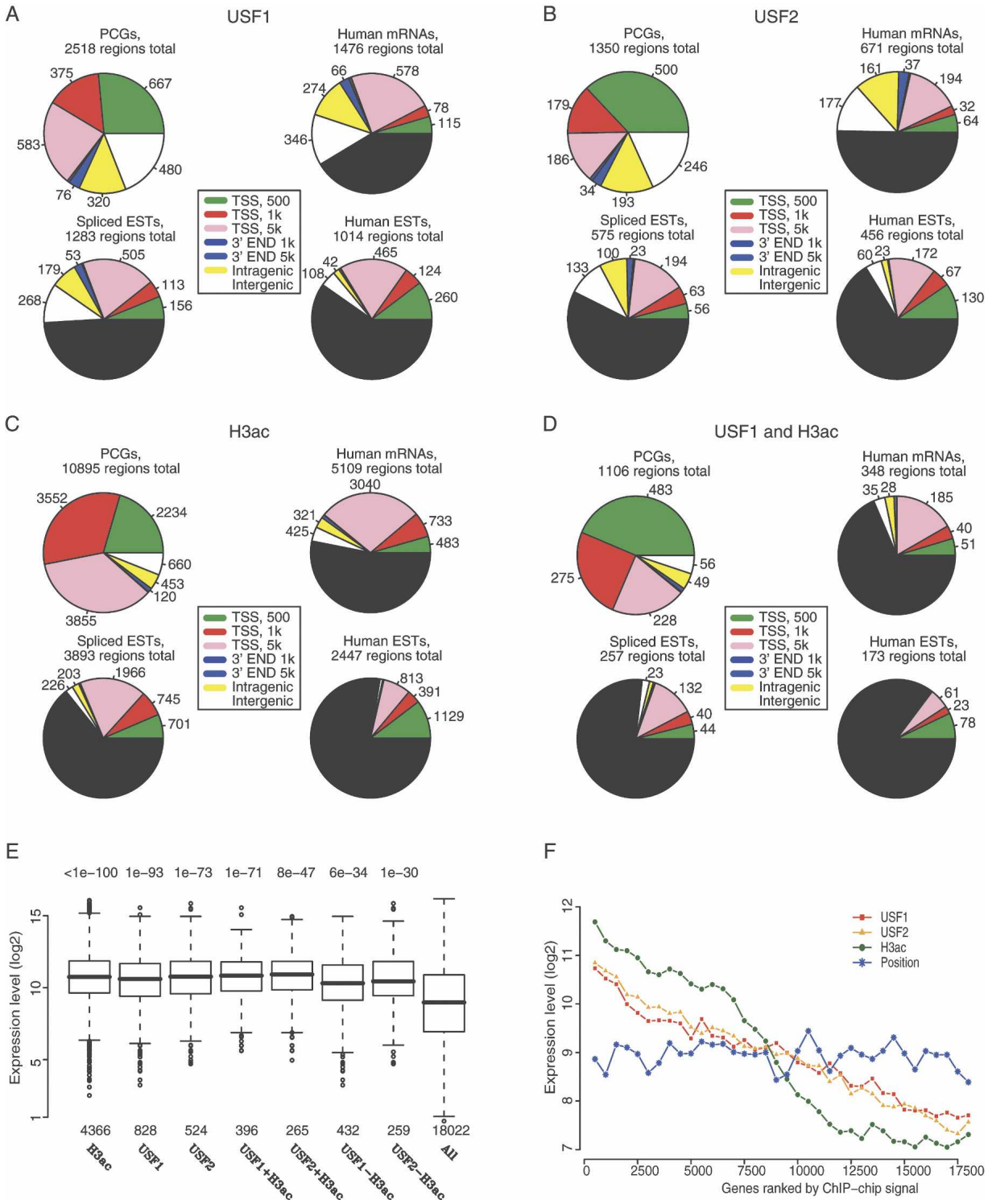


Figure 1. Mapping of USFs and H3ac binding sites to different gene and transcript annotations and correlation with expression levels. Pie charts are presented for USF1 (*A*), USF2 (*B*), H3ac (*C*), and USF1+H3ac (regions enriched both for USF1 and H3ac) (*D*). In each case, mapping of sites to the different genomic location was first established for known genes/Ensembl/RefSeq; subsequently for RNA genes, human mRNAs/spliced ESTs; and lastly for human ESTs. In each annotation, the regions mapping within 500 bp or 1 kb of TSS were not mapped in the following annotation and are indicated by the black portions in the charts. The number of regions corresponding to each genomic location category is presented next to the pie chart fractions. (*E*) The expression level of ~18,000 genes was collected from Gene Expression Omnibus (Supplemental methods). The box plots show the overall expression (in $\log_2$ scale) for all 18,000 genes and for the genes bound by the proteins indicated at the *bottom* of the graph within ±1 kb of their TSS. *P*-values from two-tailed *t*-tests between each group of genes compared with all genes are given above the box plots. (*F*) Genes were ranked in descending order according to the maximal ChIP-chip signals within ±1 kb from TSS for USF1, USF2, or H3ac experiments. As a control, we also ranked genes according to their genomic positions. Within each ranking (*X*-axes), we plotted the mean expression levels (*Y*-axes) of groups of 500 genes.

also their binding signals, as inferred from the ChIP-chip enrichment ratios, are positively correlated with expression levels (Fig. 1F). Knockdown of USF1 by siRNA was performed primarily as control of the specificity of the antibody (Supplemental Fig. S1). Subsequently and as an attempt to demonstrate that USF1 regulates the expression of its bound genes, we analyzed expression of some of the potential USF1 targets genes by RT-PCR after USF1 siRNA. We did not observe any major changes despite a reduction in USF1 binding at promoter regions (Supplemental Fig. S6a,b). These results might be explained by a compensatory effect exerted by USF2 (Supplemental Fig. S6a,b; Supplemental material), the binding of which was not affected by reduced USF1 levels or by the inability to remove all USF1 from the cell (Supplemental Fig. S1). This is not unexpected and has been found also in studies of other TFs (Xu et al. 2007). Interestingly, preliminary results analyzing effects of SNPs that disrupt USF1/USF2 binding suggest functional relevance of USFs binding sites (Supplemental material).

H3ac footprints around different TSS annotations—influence by CpG islands

We wanted to determine the profile of H3ac signal around the TSS of different annotations, and given the strong association between TSS and CpG islands (Carninci et al. 2006; Kim et al. 2005), we were interested to see how that affected the signals. We generated footprints of H3ac signals for PCG, human mRNAs, and spliced ESTs and in each group separated between CpG+ and CpG− TSSs (Fig. 2). For PCGs, we found the highest signal downstream of the TSS peaking around +500/+800, a minor peak symmetrically located upstream and a trough between (Fig. 2A), probably associated with nucleosome depletion and preinitiation complex (PIC) formation (Heintzman et al. 2007). The presence of PASRs generated from both strands around the TSS further supported the association to active promoters (Supplemental Fig. S8; Kapranov et al. 2007). CpG+ genes showed the same pattern, but the CpG− genes displayed a unique and lower downstream peak (Fig. 2A), probably as a result of a lower frequency of bidirectional conformation and slightly lower expression among these genes (Supplemental Fig. S9; Supplemental Table S3).

The CpG+ human mRNAs showed a downstream peak resembling that of PCGs, whereas the CpG− human mRNAs showed a lower and plateau formed enrichment over a couple of Kb (Fig. 2B). The CpG+ spliced ESTs showed a slightly lower peak than the human mRNAs, whereas the CpG− signal was low and covered a couple of Kb around the TSS (Fig. 2C). Our conclusion is that the height and location of H3ac around CpG+ human mRNAs TSS support that a majority of them are active TSSs, and to some degree that also is true for the CpG+ spliced ESTs TSS. The broader signal at CpG− mRNAs and spliced ESTs can arise due to the presence of 5′-truncated cDNA in these libraries and, consequently, less well defined TSS in these annotations.

H3ac signals at TSS of genes with bidirectional conformation

Pairs of PCGs with a bidirectional conformation are divergently transcribed and with their TSS separated by less than 1 kb. This type of conformation, also known as head-to-head, is common in the human genome, and ~10% of PCG are organized in this manner (Trinklein et al. 2004). When considering pairs of genes with one member being a PCG and the other a PCG or human mRNA, the proportion of PCG with bidirectional conformation is 17%, and when including spliced ESTs encoded in the opposite direction, the frequency increased to 31% (Supplemental Table S3). In the subset of PCGs with H3ac within 1 kb of their TSSs, 30% have an adjacent PCG/human mRNA and an additional

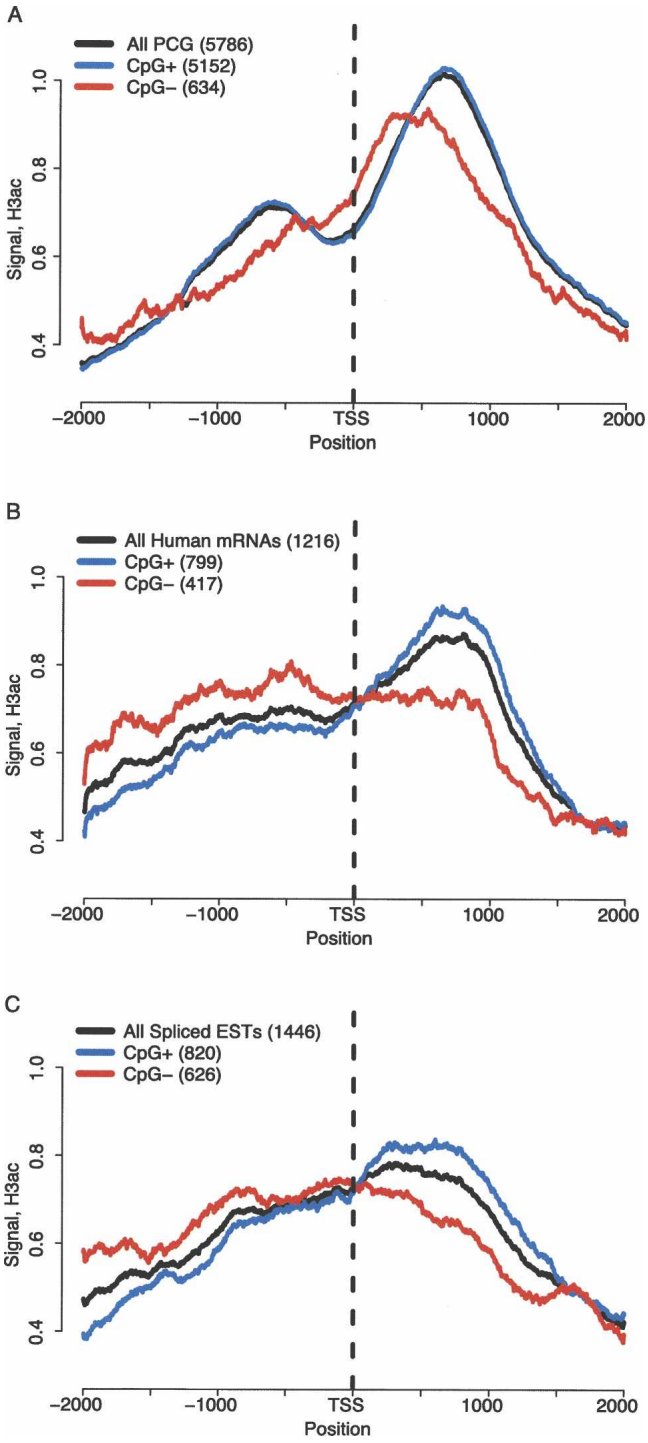


Figure 2. H3ac profile around TSSs. H3ac profiles around TSS of PCGs (*A*), human mRNAs (*B*), and spliced ESTs (*C*) depending on the presence/absence of CpG islands. All genes are bound by H3ac within 1 kb of their TSSs. H3ac binding profiles were created for each group by plotting the H3ac ChIP-chip binding signals (in $\log_2$ scale, *Y*-axes) around the TSS of each transcript, using a window of ±2 kb (*X*-axes, positive numbers are downstream of TSS; negative numbers are upstream of TSS).

17% have a spliced EST encoded in the opposite strand (Supplemental Table S3), giving a total of 47% PCGs in bidirectional conformation.

To see how this high proportion of PCGs in a bidirectional conformation might affect the shape of the H3ac footprints, we first confirmed that 80% of the PCG found in bidirectional conformation were of nonoverlapping type and that in the majority of cases the distance between the two TSSs is 100–300 bp (Fig. 3A), in agreement with previous observations (Trinklein et al. 2004). We then generated H3ac footprints for all PCG and separately for those in a bidirectional conformation and for those without any kind of anti-sense transcript within 1 kb of their TSS, which we called unidirectional PCG. The peak downstream of the TSS is in essence the same in all groups (Fig. 3B). However the bidirectional group gave an upstream peak that was higher and more extended in the upstream direction. Furthermore, if we considered PCGs with a spliced EST as the anti-sense transcript, the upstream peak was also higher than for all PCGs. The unidirectional PCGs displayed an upstream peak that was lower than for all genes and for the bidirectional groups. We also generated footprints specific for bidirectional PCGs part of overlapping or non-overlapping pairs and found that the non-overlapping group explained the overall double symmetric peak pattern with the through at −100/−200 bp (Fig. 3C; Supplemental Fig. S10).

One potential explanation for the higher upstream peak in the bidirectional group could be that these pairs of genes are divergently transcribed, resulting in two distinct H3ac peaks. Another possibility is that only one gene in each case is transcribed, but at higher levels than unidirectional genes, resulting in extended H3ac to upstream locations. However, we did not observe significant expression differences between bidirectional and unidirectional PCG (Supplemental Fig. S11). We wanted to use an independent data set to verify that the prominent double peak pattern is generated downstream of the respective TSS at promoters with bidirectional/divergent conformation and therefore made use of CAGE-tags (cap analysis of gene expression) data, which identify the 5′-ends of capped mRNAs, generated in different cell lines including HepG2 (Carninci et al. 2006). We determined the proportion of bidirectional and unidirectional PCGs positive for H3ac that presented bidirectional CAGE-tags (Methods) within 1 kb of their TSSs (Supplemental Table S4). The proportion of bidirectional CAGE-tags was clearly higher in PCG with bidirectional conformation compared with unidirectional PCG in HepG2 cells (78%, 1291 out of 1661, vs. 25%, 746 out of 2930). We also generated H3ac signal profiles for all PCGs, separating them in groups based on the presence of bidirectional CAGE-tags within 1 kb of their TSS, and the results are very similar to Figure 3B; interestingly, the upstream H3ac signal in the unidirectional group is even less evident (Fig. 3D).

Furthermore, the observation of 25% of unidirectional PCG overlapping with a bidirectional CAGE in HepG2 cells suggests that we might have underestimated the proportion of PCGs positive for H3ac and with bidirectional conformation. In order to investigate this, we separated unidirectional PCG in those overlapping with any bidirectional CAGE-tag, with HepG2 bidirectional CAGE tags, or non-overlapping (unidirectional CAGE), and then we created H3ac signal profiles for each of the groups. Once again, the upstream H3ac signal was higher in the HepG2 bidirectional CAGE-tags group than in the unidirectional CAGE-tags group, where the upstream peak has almost disappeared (Fig. 3E).

Once we understood the origin of the H3ac double peak pattern, we investigated how the H3ac signals are correlated to expression levels, by generating H3ac footprints depending on the expression pattern of the respective genes. Footprints were generated for all PCGs and separately for those belonging to unidirectional and bidirectional groups that were positive for H3ac within 1 kb of their TSS. In all cases, the downstream peak was positively correlated with expression levels (Fig. 4A–C), while the upstream peak was affected only to a small degree in all groups. The lack of correlation between upstream H3ac and expression even for bidirectional PCGs could suggest that pairs of genes in bidirectional conformation, although frequently transcriptionally active (Supplemental Table S4), can be expressed at quite different levels. We investigated the correlation in expression between pairs of PCG in a bidirectional conformation and enriched in H3ac and compared with randomly sampled pairs of PCG, and we found that in the majority of cases, both PCG in a bidirectional pair shared the same transcriptional state, typically highly active, in accordance with the CAGE-tag data, although the correlation of the expression levels was low (Fig. 4D).

The conclusion from this analysis is that H3ac signal is located downstream of the TSSs, and the peak located upstream of the TSS to a large degree is associated to previously unrecognized adjacent transcripts on the opposite strand. By our definition, 53% of PCGs positive for H3ac have unidirectional conformation, and their remaining upstream signal could be associated to still unrecognized transcripts as indicated by CAGE-tag data (Fig. 3E; Carninci et al. 2005; Kapranov et al. 2007), which suggests that the frequency of PCGs in bidirectional conformation could be even higher.

Relationship between USFs and H3ac signals at USF-bound loci

To gain further insight into the relationship between the DNA-binding proteins, combined binding profiles for the both USFs and H3ac were generated for the promoters of PCGs. The peak for USF1 and USF2 signals was found immediately upstream of the TSS at −300/−100 bp, close to the peak of the frequency of occurrence of perfect E-boxes and with the H3ac signal being essentially the same as seen for all genes (Fig. 5A). When separating PCG in CpG+ and CpG−, the USF1 signal was slightly different, but the accompanying H3ac footprint became clearly lower in the CpG− group (Supplemental Fig. S12), corresponding to significantly lower expression of these genes (Supplemental Fig. S9).

Since 25% of USF1-bound PCGs were found in bidirectional conformation (Supplemental Table S3), we created footprints for the unidirectional and bidirectional subsets. The USF1 signal is the same in the two subgroups, but in the bidirectional group, the H3ac upstream peak is higher and shifted further upstream from the TSS. As for H3ac, the majority of USF1-bound bidirectional PCGs involved pairs of non-overlapping transcripts. Therefore we created footprints for H3ac and USF1 centered on the midpoint between the TSSs of these non-overlapping transcripts. The H3ac signal peaks at 500–800 bp downstream of the respective TSS and the USF1 signal are maximal between the TSSs, which is where most E-boxes are found and where the trough in H3ac is located, probably corresponding to a nucleosome free region (NFR) (Fig. 5B).

Properties of USF enriched regions concerning chromatin and sequence

After showing the spatial and functional association between H3ac and USF proteins, we wanted to get a broader picture of

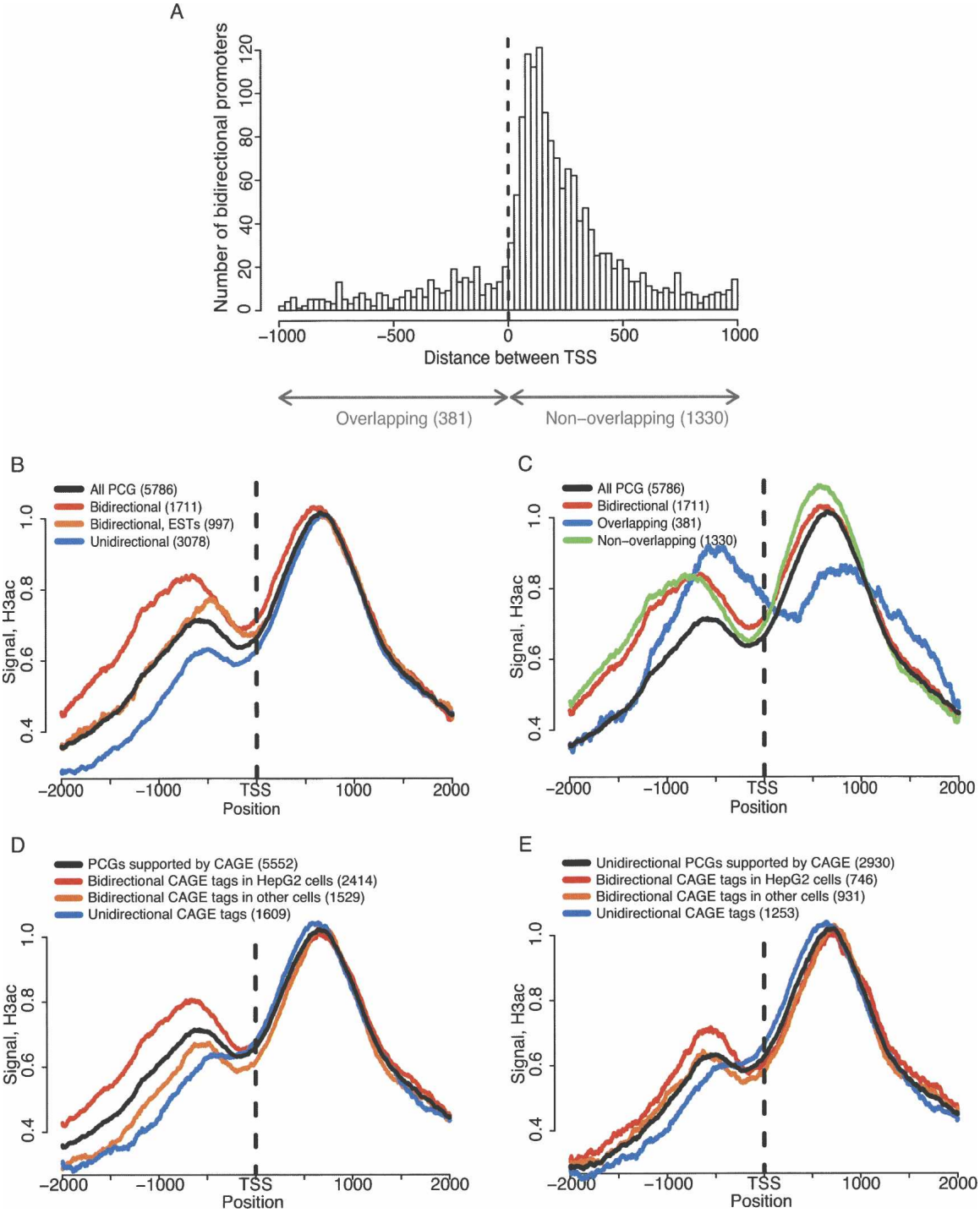


Figure 3. H3ac profiles around TSS are influenced by bidirectional conformation. (*A*) Histogram showing the distances between the two TSSs in each bidirectional pair. PCGs bound by H3ac within 1 kb of their TSS and with at least one anti-sense PCG or human mRNA TSS within 1 kb of their TSS were considered. Negative distances correspond to overlapping TSSs. (*B*) H3ac signals around PCGs separated into bidirectional (at least one anti-sense PCG or human mRNA TSS within 1 kb of their TSS), bidirectional-ESTs (at least one anti-sense spliced EST TSS within 1 kb of their TSS) or unidirectional (no anti-sense PCG, human mRNA, or spliced EST TSS within 1 kb of their TSS). (*C*) H3ac binding signals for all PCGs (black line), all PCGs with bidirectional conformation (red), overlapping bidirectional conformation (blue), and non-overlapping bidirectional conformation (green). (*D*,*E*) H3ac signal profiles were generated for different groups of PCGs (*D*) and unidirectional PCGs (*E*), as indicated by color codes and that were created based on CAGE-tag data.

how other components of chromatin affect the binding of the TFs. We therefore analyzed various components of euchromatin (e.g., H3K4me3, H4ac, RNA polymerase II [POLR2A]), heterochromatin (e.g., H3K27me3), and other factors in 26 regions, 23 positive for USFs and three unbound negative controls, by ChIP and qPCR. Regions and proteins were clustered based on absolute

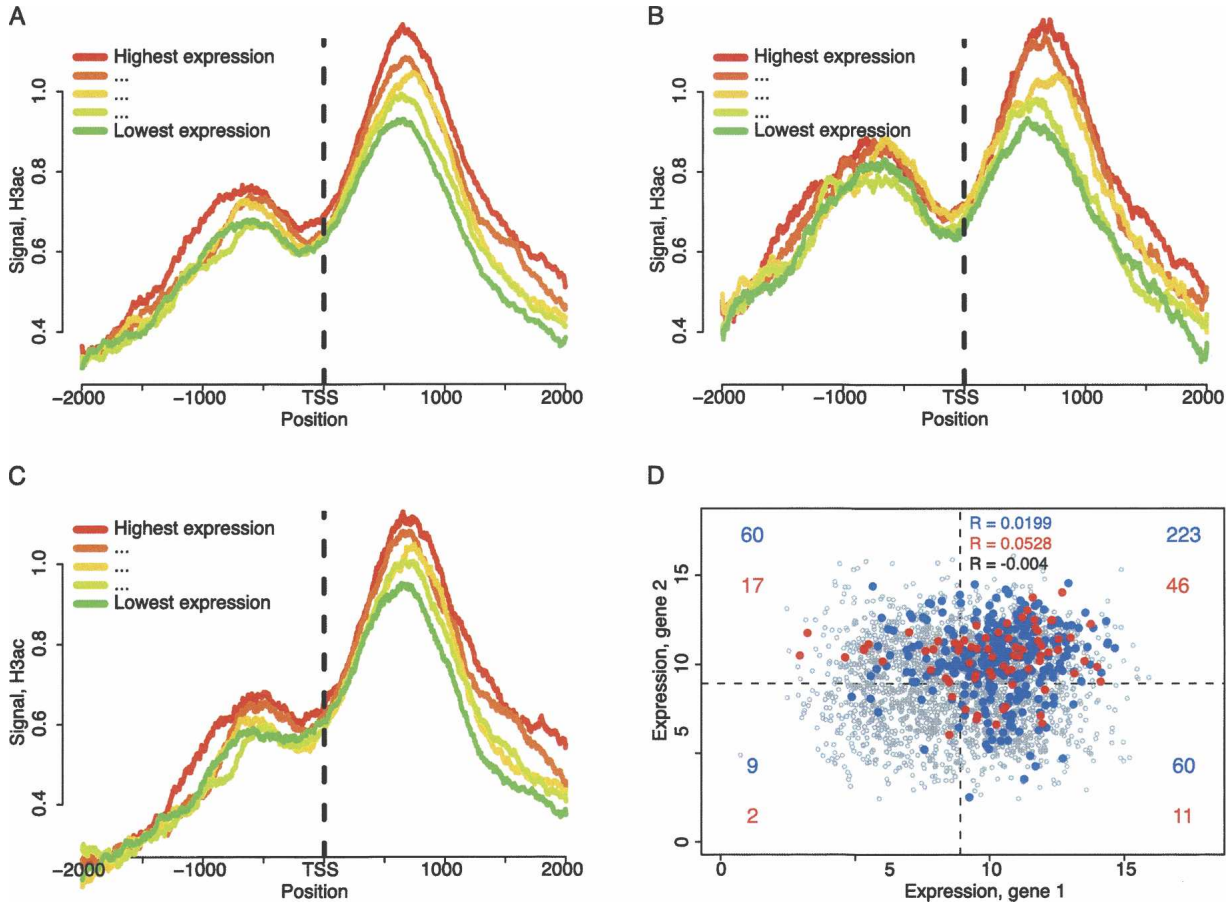


Figure 4. Positive correlation between H3ac downstream of PCG TSS and expression levels. (*A*) H3ac profiles of PCGs divided in five equal numbered groups based on their expression levels. All PCGs bound by H3ac within 1 kb of their TSS and where expression data were available were considered. Similar H3ac profiles when separating the PCGs above in bidirectional (PCG+PCG or human mRNA) (*B*) or unidirectional (*C*). (*D*) PCGs bound by H3ac (blue) or USF1 (red) within 1 kb of their TSS, found in non-overlapping bidirectional conformation and with available expression data for the two PCGs involved in each bidirectional pair were considered; 1500 pairs of PCG were randomly sampled, and their expression levels were similarly represented (gray). The dotted lines indicate the average level of expression for the randomly sampled PCGs. Pearson correlation (*R*) values for USF1, H3ac, and randomly sampled data sets expression levels are in the *upper right* corner. The expression data was $\log_2$ transformed.

qPCR signals, and pairwise correlation coefficients were calculated (Fig. 6A; Supplemental Figs. S13, S14). All but one of the USF-bound regions could be separated from the negative controls. This was not just due to the USFs signals since similar clusters were obtained when the USFs were excluded from the analysis (Supplemental Fig. S13). USFs binding occurred in domains with high levels of one or more of the euchromatic marks and low levels of H3K27me3. They could be further subdivided in two groups (I and II). Group I consisted mainly of regions close to TSS with high levels of H3ac, H3K4me3, and POLR2A. Group II contained most of TSS distal regions analyzed, which contained other marks such as H4ac or HNF4A but lower levels of H3ac and H3K4me3, although higher than unbound regions, suggesting that USFs can participate in distal transcriptional regulation (West et al. 2004). Among this second group, there were a few TSS proximal regions (subgroup II.1) with POLR2A levels similar to group I, but lower H3ac/H3K4me3, which together with the high correlation between USFs and POLR2A bindings (Supplemental Figs. S13–S15), suggest that USFs main function is POLR2A recruitment in a promoter context. This functional role of USFs is supported by the discovery of heterozygous SNPs occurring in USF binding sites, where USFs preferential binding to one of the alleles is accompanied by higher POLR2A binding to that same allele (A. Ameur, A. Rada-Iglesias, J. Komorowski, and C. Wadelius, unpubl.; Supplemental material).

To get further knowledge of the sequence preference for USF1 and USF2, we used ab initio and candidate scanning approaches. All USF1- and USF2-bound sequences in the stringent data set were divided in groups based on $\log_2$ values and analyzed using Bioprospector (Liu et al. 2001). For both USF1 and USF2, clear E-boxes were found in regions with strong enrichment, down to $\log_2$ values of 1.25 (Fig. 6B), extending the established motif from CACGTG to CACGTGAC, as previously suggested (Corre and Galibert 2005). We have verified several regions with $\log_2$ values <1.25 using qPCR (five for USF1 and 19 for USF2, with one and three having matches with CACGTGAC sequences, respectively), so the absence of the established consensus is not due to false positives. In the groups with lower enrichment, we instead found polypyrimidine-rich sequences resembling initiator elements. The positive correlation between E-box sequences and the strength of USF–DNA interactions was also confirmed after counting occurrences of perfect E-box matches among USFs targets (Fig. 6C; Supplemental Fig. S16).

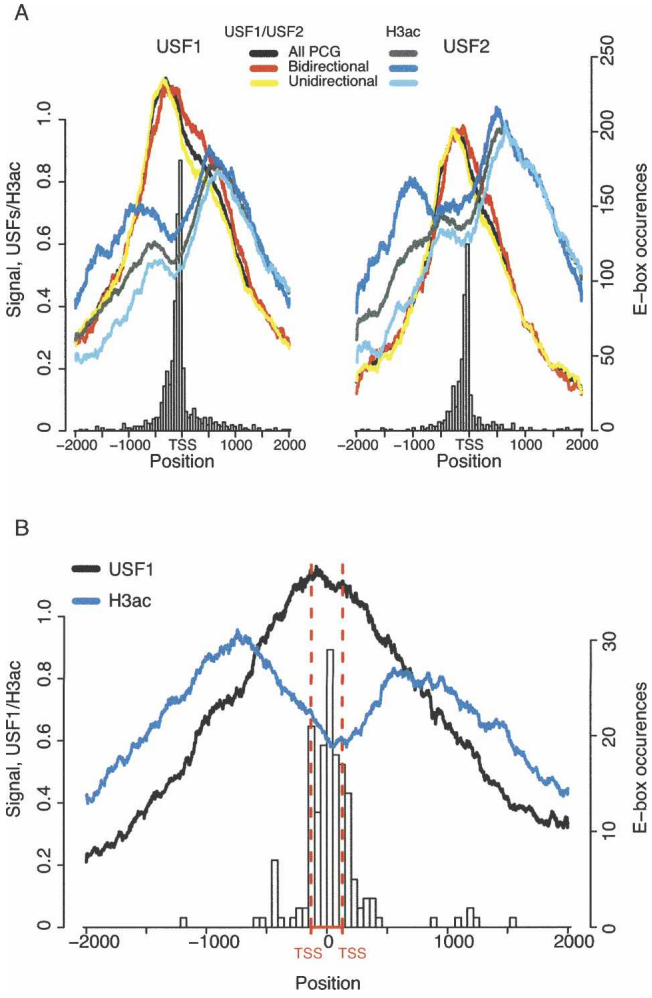


Figure 5. Architecture of USFs bound promoters. (*A*) PCGs bound by USF1 (*left*) or USF2 (*right*) within ±1 kb of their TSS. The ChIP-chip signals (in $\log_2$ scale, *right Y*-axes) of USF1 or USF2 and H3ac were plotted around TSS. Profiles are presented for all USF1- or USF2-bound PCG (all), bidirectional PCGs (at least one anti-sense PCG or human mRNA TSS within 1 kb of their TSS), or unidirectional PCGs (no anti-sense PCG, human mRNA, or spliced EST TSS within 1 kb of their TSS). Histograms of perfect matches to E-box consensus sequence (CACGTGAC, *left Y*-axes) are presented for all USF1- or USF2-bound genes in 50-bp windows. (*B*) All PCGs in a non-overlapping bidirectional conformation bound by USF1 within ±1 kb of their TSSs were selected. The binding profiles (*left Y*-axes) of USF1 and H3ac were plotted around 2 kb of the middle position between the two transcripts in each bidirectional pair. The red dotted lines show the median size of the intergenic region between TSSs in each bidirectional pair. The histogram shows E-box consensus occurrences (*right Y*-axes).

Regions bound by USF2 define at set of distal regulatory elements

USF1 and USF2 shared most of their targets, suggesting that they bind as a heterodimer (Table 1; Supplemental Fig. S7). However, there was clear evidence for a subset of regions bound only by USF2 (Supplemental Fig. S7), and based on USF2/USF1 signal ratios, we identified 240 sites that were unique to USF2 and 627 sites bound by USF1–USF2 heterodimers (Fig. 7A). The USF2-unique sites were less frequently found close to PCG TSS than sites bound by both USFs (45% vs. 62% within 1 kb) (Supplemental Fig. S17). Also, a motif similar to HNF4A and FOXA2 (HNF3B) consensus sequences was found (Fig. 7B) to be most abundant among weak USF2-unique targets (Supplemental Fig. S17) that tend to be far from TSSs (Supplemental Fig. S18). Eight regions with perfect matches to HNFs consensus were investigated by ChIP for HNF4A, FOXA2 (HNF3B), and FOXA1 (HNF3A). All regions were bound by HNF4A, and seven of eight were bound by all proteins (Supplemental Fig. S19), confirming a new type of distal regulatory module involving USF2 and the three HNFs.

Candidate genes for FCHL

Alleles of USF1 are associated with FCHL (Pajukanta et al. 2004), and out of the genes that are potentially regulated by this factor, we wanted to find the most relevant ones. We selected genes with USF1 or USF2 binding within 5 kb or 500 bp of their TSS and performed gene ontology (GO) analysis comparing with the whole genome. This was done for cellular compartment, molecular function, and biological process of GO categories and of the overrepresented groups; the one most relevant to the phenotype contained nuclear mitochondrial genes, especially enzymes involved in energy production and ATP biosynthesis (Supplemental Fig. S20; link to the complete GO analysis can be found in Supplemental Methods). In all three ontologies, mitochondrial terms were overrepresented. Also other genes involved in energy homeostasis are bound by USF1, e.g., the transcription factors *HNF4A* and *PPARG* and the signaling proteins *PTEN* and *SIRT1*. Among the genes bound by USF1, we also searched for functional candidates and found several genes involved in triglyceride, lipoprotein, and cholesterol metabolism and a large group involved in fatty acid beta-oxidation (Supplemental Table S5).

GABPA (NRF-2) is known to regulate mitochondrial genes (Scarpulla 2006), and we found an overrepresentation of its consensus binding sequence in strong and TSS proximal USFs targets (Supplemental Fig. S21). We wanted to see if GABPA coregulated mitochondrial and other types of USF-bound genes and investigated GABPA binding by ChIP in nine USF-bound gene promoters containing perfect GABPA consensus sequences and six USF-bound mitochondrial gene promoters, with no perfect match to GABPA consensus. In five out of nine and four out of six cases, there was evidence of GABPA binding, indicating that USFs frequently cooperate with GABPA in a promoter context (Supplemental Fig. S21).

Discussion

Recent studies in yeast have resolved the structure of RNA polymerase II promoters at a single nucleosome resolution, concluding that active promoters present a NFR of ~200 bp upstream of TSSs flanked by well-positioned nucleosomes (Yuan et al. 2005; Segal et al. 2006). In this context, histone acetylation preferentially occurs downstream of TSSs and spans several hundred base pairs (Liu et al. 2005; Pokholok et al. 2005). In human cells, this basic promoter architecture seems conserved, with the NFR occupied by PIC when genes are actively transcribed or poised for transcription (Kim et al. 2005; Heintzman et al. 2007). However, it is believed that histone acetylation upstream of TSS plays an important role in active promoters, although histone acetylation downstream of TSS is also observed (Heintzman et al. 2007). Our high-resolution analysis shows a double symmetric peak of acetylation 500–800 bp upstream and downstream of the TSS, similarly as previously reported for this and other histone marks (Birney et al. 2007; Guenther et al. 2007; Heintzman et al. 2007).

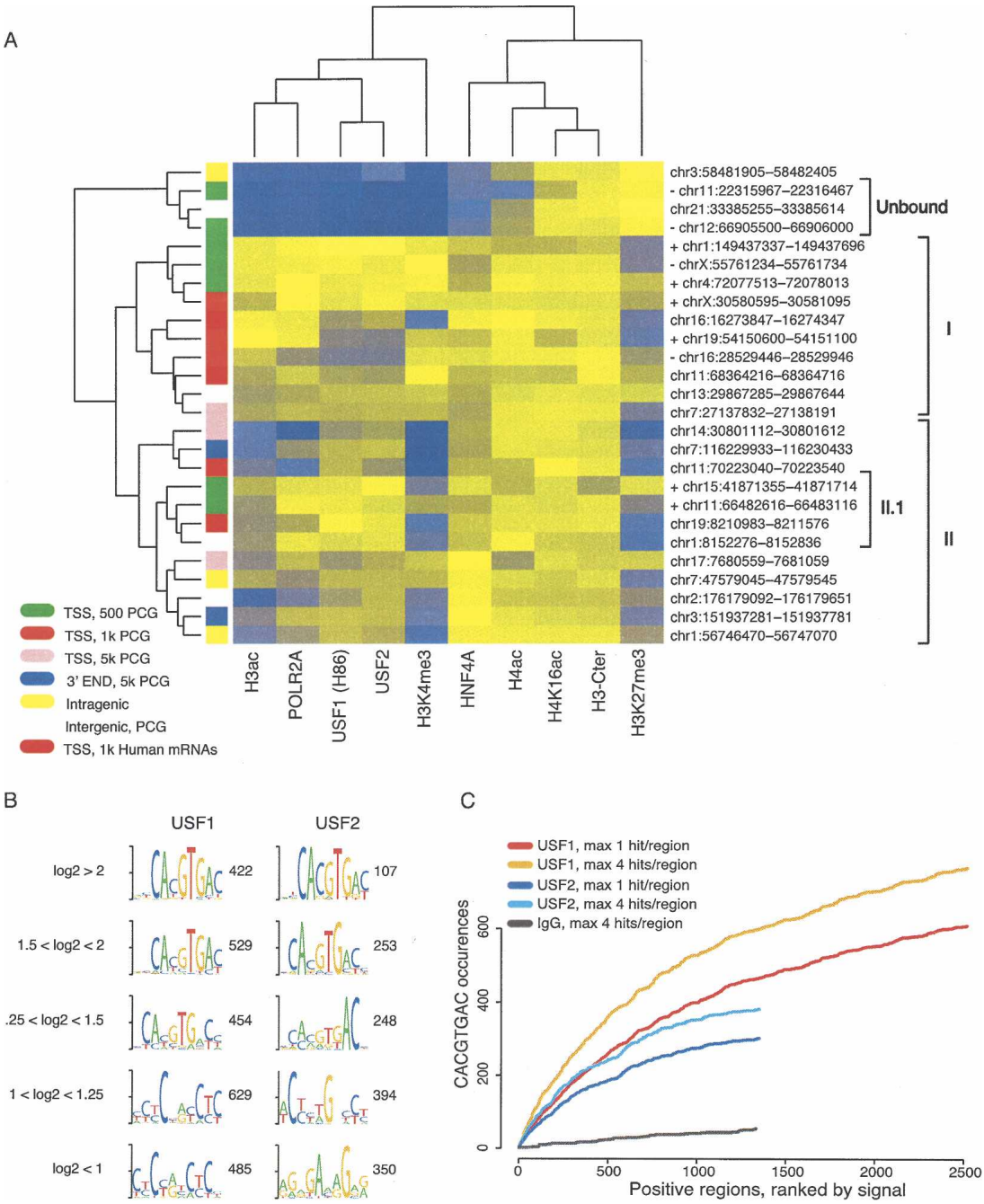


Figure 6. USFs binding determinants. (*A*) Twenty-six selected regions were clustered based on qPCR signals for the indicated transcription factors and histone modifications. The column colors indicate the qPCR signal from high (yellow) to low (blue) values. The genomic coordinates of the analyzed regions are presented to the *right*, while the genomic location of each region is indicated by the color to the *left*. Plus (+) and minus (−) signs next to the genomic coordinates for regions proximal to PCG TSS indicate CpG+ or CpG− regions, respectively. The genomic positions are based on Human Mar. 2006 (hg18) assembly (NCBI Build 36.1). (*B*) The top enriched motifs identified ab initio for USF1 and USF2 in groups based on their $\log_2$ enrichments values. The number of regions in each group is presented. (C) Candidate scanning approaches were used to search for the E-box (i.e., CACGTGAC) among USF1- or USF2-bound regions, ranked in descending order according to their $\log_2$ enrichment values. For both USF1 and USF2, we considered either the total number of E-box occurrences with a maximum of four per region or just one occurrence per region. To see how many occurrences are expected in a background data set, we also generated a similar slope from IgG-positive regions.

However, no functional distinction between the two peaks has been established, which leads to the assumption that both peaks are functionally associated to the nearest gene. When we take into account that at least 30% of H3ac-positive PCG promoters have a head-to-head configuration with a preferred distance of 100–300 bp between TSSs, we see that the H3ac signal upstream of the TSS is drastically decreased at genes with unidirectional conformation, and the preponderant downstream location of

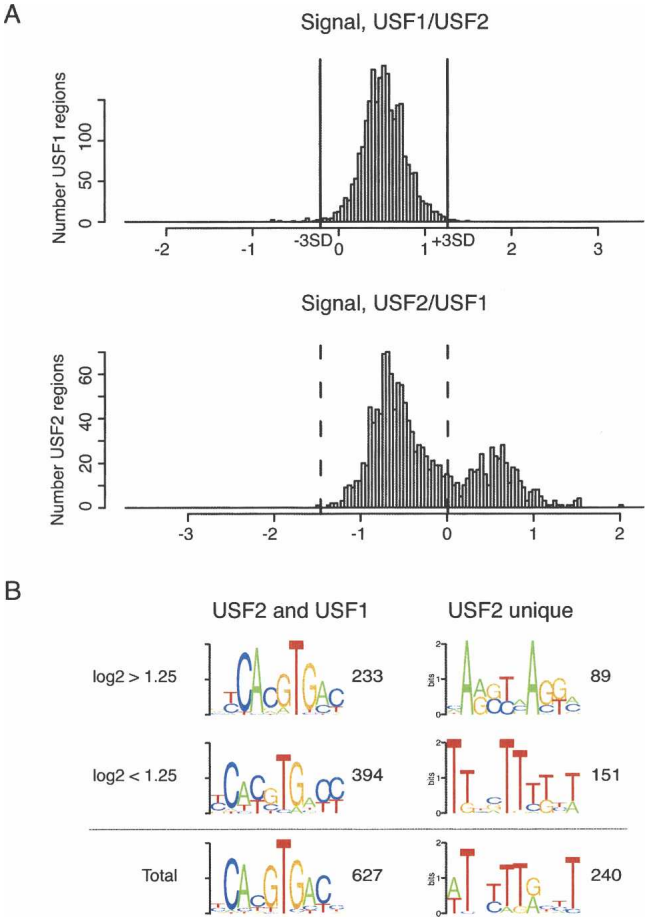


Figure 7. USF2 binds a subset of unique targets. (*A*) For USF1 (*top*) and USF2 (*bottom*) the $\log_2$ ratios between USF1/USF2 and USF2/USF1 signals were calculated (*X*-axes) and shown in histograms. The vertical gray lines demarcate the regions deviating $>\pm 3$ SD from the median of the USF1/USF2 distribution. (*B*) USF2-bound regions were divided based on the $\log_2$ USF2/USF1 signal ratios in *A*, in those with ratios higher than 0.5 (USF2 unique) or lower than –0.5 (USF2–USF1 heterodimers). Enriched motifs were identified ab initio for all USF2-unique and USF2–USF1 regions, and after dividing them in groups based on $\log_2$ enrichment signals.

H3ac is also supported by independent CAGE-tag data. Therefore, a majority of upstream histone acetylation is present downstream of adjacent transcripts on the opposite strand. Furthermore, although positive correlation between H3ac and gene expression has been previously reported, here we demonstrate that it is the downstream H3ac that mainly explains this correlation.

The double peak pattern of H3ac resembles those observed for H3ac and H3K4me3 (Ng et al. 2003; Pokholok et al. 2005) in yeast and for H3.3 (Mito et al. 2005) in *Drosophila*. Localization of H3K4me3 at the 5′ end of genes is dependent on elongation factors and POLR2A ser5 phosphorylation (Ng et al. 2003). Interestingly, H3K4me3 serves as a recruitment mark for HATs and is an optimal substrate for histone acetylation (Pray-Grant et al. 2005; Taverna et al. 2006). Elucidating that the H3ac associated with a particular gene is located largely downstream of its TSS leads us to propose that the observed H3ac patterns could be transcription dependent and follow H3K4me3 patterns. H3ac can be restricted from 3′end gene regions by HDACs coupled to H3K36me dependent on POLR2A ser2 phosphorylation. This H3ac pattern could be part of a short transcriptional epigenetic memory (Ng et al. 2003; Kouskouti and Talianidis 2005) but could also facilitate the transition between transcription initiation and elongation, which is rapid in eukaryotic organisms compared with prokaryotes (Reppas et al. 2006). This model does not contradict the generalized view of HAT recruitment to promoter regions by TFs, since an initial sequence/gene-specific recruitment of HAT complexes can be followed by complex stabilization and catalytic activation in a transcription-dependent manner (Ruthenburg et al. 2007).

Under the described structure for active promoters, the NFR not only accommodates the PIC but transcription factors and their binding sequences seem to preferentially occur within these regions as well (Yuan et al. 2005; Segal et al. 2006; Ozsolak et al. 2007). In agreement with this, USFs bindings and E-boxes are clearly overrepresented within the predicted NFR. This is best exemplified by the subset of genes with bidirectional conformation that have a binding site for USF1 within 1 kb of their TSS and where we see that USF1 binds in between the two TSSs.

The relationship between USF binding and H3ac seems to be mainly restricted to promoter regions of actively transcribed genes. For example, CpG− PCG promoters bound by USFs had lower H3ac and expression levels than did their CpG+ counterparts. These genes might be poised for transcription in some cases (Radonjic et al. 2005) and could be induced under the appropriate conditions, illustrating the tissue-specific expression patterns typical of CpG− genes (Yamashita et al. 2005; Sandelin et al. 2007). GO analysis of USF-bound CpG− genes compared with all genes or all USF-bound genes identified an overrepresentation of genes involved in regulation of NF-κB cascade, which is inducible and that could involve genes in transcriptionally poised states (data not shown). Finally, USF binding patterns best correlated with POLR2A, which, together with previous reports, suggests that the major transactivating mechanisms of USF proteins is the interaction with the TFIID complex and the RNA polymerase II machinery (Meisterernst et al. 1990; Chiang and Roeder 1995), which is also supported by our analysis of regulatory SNPs disrupting USF binding (A. Ameur, A. Rada-Iglesias, J. Komorowski, and C. Wadelius, unpubl.).

Besides functional and structural data, our study also provides a better understanding of the major determinants of USF–DNA interactions. USF binding sites are restricted to chromatin rich in euchromatic marks and depleted of H3K27me3. This agrees with reports indicating that CpG methylation at CACGTG E-box sequences abrogates USF–DNA interactions (Fujii et al. 2006), and given the connections between DNA methylation and H3K27me3 (Vire et al. 2006) may explain tissue-specific gene expression of USF-regulated genes.

Once USF–DNA interaction is allowed by the chromatin context, the strength of such interaction depends on the underlying DNA sequences, with most intense/frequent interactions occurring with canonical E-box sequences that are most common in promoter regions. However, weaker interactions do not to depend on E-boxes, but instead cooperative or indirect binding with other proteins seems more important (Carroll et al. 2006; Yang et al. 2006; Kim et al. 2007). This is exemplified by weaker USF1 binding sites, where we found polypyrimidine-rich sequences resembling initiator elements. USF proteins have been reported to bind to such elements either directly or indirectly through GTF2I (Roy et al. 1991; Du et al. 1993). Also among weak and distal USF2-unique targets, we confirmed an overrepresenta-

tion of HNF4A and FOXA1/FOXA2 sites, where all these proteins seem to bind together.

Finally, our data suggest a previously unknown regulatory role for USFs in mitochondrial activity. Given the major role of mitochondrial function in energy homeostasis and its role in diabetes, metabolic syndrome, or insulin resistance (Auwerx 2006), it is tempting to speculate that alteration of the transcriptional control of mitochondrial activity can be one major hallmark associated with USF variation. As an example, USFs were bound to promoters of *PPARGC1A*, which is a key regulator of mitochondrial biogenesis and respiration (Wu et al. 1999), and PPARGC1A regulator *SIRT1* (Rodgers et al. 2005), suggesting a feed-forward regulatory loop. Furthermore, USFs and GABPA (NRF-2), a major regulator of nuclear mitochondrial genes, co-occupy the promoters of various genes not restricted to mitochondrial function, resembling the scenario suggested for E2F4 and NRF1 proteins (Cam et al. 2004). This might be explained by the fact that GABPA binding, like USF1, seems overrepresented in PCGs with bidirectional conformation (Lin et al. 2007), and interestingly, mitochondrial genes are also one of the most overrepresented functional categories when considering bidirectional PCG (Trinklein et al. 2004).

In conclusion, our characterization of USF binding profiles across the entire human genome in a liver cellular model represents a valuable resource in order to expand our knowledge of the transcriptional control of metabolic processes and its alteration in metabolic disorders.

Methods

ChIP-chip analysis

Chromatin immunoprecipitation was performed as previously described (Rada-Iglesias et al. 2005), but sonication conditions were optimized to obtain smaller fragments (~300 bp) to further improve the resolution of our experiments. The antibodies against USF1 (H-86, sc-8983) and USF2 (C-20, sc-862) were from Santa Cruz Biotechnology; anti-Histone H3 acetyl K9/14 (06–599) and normal rabbit IgG (12–370) were purchased from Upstate. Three completely independent biological replicates were performed for each antibody, obtaining the corresponding input as total genomic DNA reference. DNA amplification, fragmentation, and labeling were performed according to Affymetrix recommendations. Hybridizations were performed using Affymetrix GeneChip Human Tiling 2.0R Array set (seven arrays set). Array data files for USF1, USF2, H3ac, and IgG were normalized against corresponding input arrays using Affymetrix Tilling Array Software (TAS) two-group normalization. The resulting signals for USF1, USF2, and H3ac were then grouped together with IgG signals, and an empirical Bayes algorithm (Smyth 2004) was run on the combined data. In this way, three B-values were calculated for each probe measuring the reproducibility of the data and the probability that the probe contains a significantly higher enrichment for USF1, USF2, and H3ac, respectively, when being compared with IgG. A high B-value indicates that the enrichment at a probe is similar between the three biological replicates of one of the investigated proteins, while the corresponding signals are different on the IgG arrays.

Enriched regions were defined as a window containing at least four (three) probes with (1) $\log_2$-signal at least 3 SD above mean, (2) B-value at least 6 SD (2 SD) above mean, and (3) maximum $\log_2$-signal above 0.8 (0.7). The maximum allowed gap between two probes passing the criteria above is 110 bases, and the distance between two different peaks is required to be at least 400. In parenthesis are the cut-offs used for the relaxed results. The $\log_2$ cut-off was lowered to 0.6 in the USF2 relaxed data set since we have noted that USF2 gives an overall lower signal than USF1 for their common targets, due to higher ChIP background level associated with the USF2 antibody (Supplemental material). Average probe level intensities and identified positive regions for USF1, USF2, and H3ac can be viewed in the UCSC Genome Browser (http://genome.ucsc.edu/).

When comparing two sets of genomic regions from ChIP-chip data sets, first we calculated the center positions in the first set. An overlap was reported if any base within a fixed window around the center position was shared with some genomic region in the second set. The window size was 1 kb or 2 kb in each direction around the center position.

Mapping enriched regions to different gene and transcript annotations

Enriched regions were grouped according to the distance to the closest transcript in a hierarchical manner. Regions were first compared with gene coordinates for PCG. If a TSS was found within 500 bases of the center of the enriched region, then the region was put in the TSS 500-bp category. If not, the region was sequentially matched to the categories TSS 1 kb, TSS 5 kb, 3′ END 1 kb, 3′ END 5 kb, and intragenic. Regions not belonging to any of the classes above were categorized as intergenic. In the next step, all regions at more than 1 kb distance of TSS were matched to the human mRNA coordinates. The remaining regions were then matched to spliced ESTs and finally to the human EST coordinates.

PCGs, human mRNAs, or spliced ESTs with their TSS within at most 1 kb of a ChIP-chip enriched region were grouped into different classes based on the surrounding genomic context.

Bidirectional / unidirectional

We started from all PCG having their TSS within 1 kb of an enriched region. One of those PCGs was defined as bidirectional if one other PCG or human mRNA was found on the opposite strand with at most 1 kb separating the two TSSs. PCGs not defined as bidirectional were classified as "bidirectional ESTs" if instead a spliced EST met the same criterion. PCGs not falling in any of the bidirectional groups were considered unidirectional. Bidirectional PCGs were then further separated into overlapping and non-overlapping, depending on whether any base was shared by the two transcripts or not.

CpG+ / CpG–

Enriched regions were defined as CpG+ if there was an overlapping CpG island in a 1-kb window (how overlaps are calculated is described above) and as CpG− otherwise. TSSs with some CpG+ region within 1 kb were grouped into the CpG+ category, while the others were considered to be CpG−. Coordinates for genes and CpG islands were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/).

CAGE–tags

We downloaded the human CAGE primary database (Kawaji et al. 2006) from the FANTOM3 site (http://fantom.gsc.riken.go.jp) to get an independent bidirectional/unidirectional classification of our PCG TSS. We then removed those TSSs not supported by any CAGE-tag within ±1 kb and partitioned the remaining TSSs into three classes based on CAGE-tag data for the opposite strand in a 1-kb window. In the first group were TSSs matching some bidirectional CAGE-tag from HepG2 cells, and in the second were

those of the remaining TSSs that matched in some other cell type. The third group consisted of those with no bidirectional CAGE-tag at all.

ChIP-chip signal profile around TSSs

Each position in a 2-kb window of a TSS was assigned the mean ChIP-chip signal intensity of all probes in the immediate neighborhood (the probe center at most 20 bases away). The procedure was repeated for all TSSs in a specific group, e.g., enriched by USF1, USF2, H3ac, and so on. The final profile was then calculated as the mean value at each position for all TSSs in the group.

Motif discovery

USF1 and USF2 enriched regions were extended so they contained at least 500 bases. The resulting sequences were divided into different categories depending on ChIP-chip signal strength and genomic localization. Each category was then analyzed individually using 200 BioProspector (Liu et al. 2001) iterations (default settings) to detect overrepresented DNA motifs.

Consensus sequences for some selected TFs (USF1, FOXA2, HNF4A, and GABPA) were matched to all enriched regions. The exact genomic localizations and number of occurrences in each sequence were reported.

Acknowledgments

This project was funded by the Swedish Research Council, Novo Nordisk Foundation, Swedish Diabetes Foundation, the Family Ernfors Fund, The Markus Borgström Foundation, The Beijer Foundation, Federal Funds from the National Cancer Institute, National Institutes of Health (under contract no. N01-CO-12400), the National Human Genome Research Institute, National Institutes of Health (under grant no. U01 HG003147), and Affymetrix, Inc.

References

Auwerx, J. 2006. Improving metabolism by increasing energy expenditure. *Nat. Med.* **12:** 44–45.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas III, E.J., Gingeras, T.R., et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120:** 169–181.

Birney, E.J.A., Stamatoyannopoulos, A., Dutta, R., Guigo, T.R., Gingeras, E.H., Margulies, Z., Weng, M., Snyder, E.T., Dermitzakis, R.E., Thurman, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Cam, H., Balciunaite, E., Blais, A., Spektor, A., Scarpulla, R.C., Young, R., Kluger, Y., and Dynlacht, B.D. 2004. A common set of gene regulatory networks links metabolism and growth inhibition. *Mol. Cell* **16:** 399–411.

Carninci, P.T., Kasukawa, S., Katayama, J., Gough, M.C., Frith, N., Maeda, R., Oyama, T., Ravasi, B., Lenhard, C., Wells, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38:** 626–635.

Carroll, J.S., Meyer, C.A., Song, J., Li, W., Geistlinger, T.R., Eeckhoute, J., Brodsky, A.S., Keeton, E.K., Fertuck, K.C., Hall, G.F., et al. 2006. Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* **38:** 1289–1297.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.

Chiang, C.M. and Roeder, R.G. 1995. Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators. *Science* **267:** 531–536.

Coon, H., Xin, Y., Hopkins, P.N., Cawthon, R.M., Hasstedt, S.J., and Hunt, S.C. 2005. Upstream stimulatory factor 1 associated with familial combined hyperlipidemia, LDL cholesterol, and triglycerides. *Hum. Genet.* **117:** 444–451.

Corre, S. and Galibert, M.D. 2005. Upstream stimulating factors: Highly versatile stress-responsive transcription factors. *Pigment Cell Res.* **18:** 337–348.

Du, H., Roy, A.L., and Roeder, R.G. 1993. Human transcription factor USF stimulates transcription through the initiator elements of the HIV-1 and the Ad-ML promoters. *EMBO J.* **12:** 501–511.

Fujii, G., Nakamura, Y., Tsukamoto, D., Ito, M., Shiba, T., and Takamatsu, N. 2006. CpG methylation at the USF-binding site is important for the liver-specific transcription of the chipmunk HP-27 gene. *Biochem. J.* **395:** 203–209.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130:** 77–88.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol.* **7:** 1–9.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39:** 311–318.

Huertas-Vazquez, A., Aguilar-Salinas, C., Lusis, A.J., Cantor, R.M., Canizales-Quinteros, S., Lee, J.C., Mariana-Nunez, L., Riba-Ramirez, R.M., Jokiaho, A., Tusie-Luna, T., et al. 2005. Familial combined hyperlipidemia in Mexicans: Association with upstream transcription factor 1 and linkage on chromosome 16q24.1. *Arterioscler. Thromb. Vasc. Biol.* **25:** 1985–1991.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermueller, J., Hofacker, I.L., et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316:** 1484–1488.

Kawaji, H., Kasukawa, T., Fukuda, S., Katayama, S., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. 2006. CAGE Basic/Analysis Databases: The CAGE resource for comprehensive promoter analysis. *Nucleic Acids Res.* **34:** D632–D636.

Kim, T.H., Barrera, L.O., Qu, C., Van Calcar, S., Trinklein, N.D., Cooper, S.J., Luna, R.M., Glass, C.K., Rosenfeld, M.G., Myers, R.M., et al. 2005. Direct isolation and identification of promoters in the human genome. *Genome Res.* **15:** 830–839.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128:** 1231–1245.

Komulainen, K., Alanne, M., Auro, K., Kilpikari, R., Pajukanta, P., Saarela, J., Ellonen, P., Salminen, K., Kulathinal, S., Kuulasmaa, K., et al. 2006. Risk alleles of USF1 gene predict cardiovascular disease of women in two prospective studies. *PLoS Genet.* **2:** e96. doi: 10.1371/journal.pgen.0020069.

Kouskouti, A. and Talianidis, I. 2005. Histone modifications defining active genes persist after transcriptional and mitotic inactivation. *EMBO J.* **24:** 347–357.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., et al. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125:** 301–313.

Li, B., Carey, M., and Workman, J.L. 2007. The role of chromatin during transcription. *Cell* **128:** 707–719.

Lin, J.M., Collins, P.J., Trinklein, N.D., Fu, Y., Xi, H., Myers, R.M., and Weng, Z. 2007. Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res.* **17:** 818–827.

Liu, X., Brutlag, D.L., and Liu, J.S. 2001. BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* 127–138.

Liu, C.L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S.L., Friedman, N., and Rando, O.J. 2005. Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol.* **3:** e328. doi: 10.1371/journal.pbio.0030328.

Meisterernst, M., Horikoshi, M., and Roeder, R.G. 1990. Recombinant yeast TFIID, a general transcription factor, mediates activation by the gene-specific factor USF in a chromatin assembly assay. *Proc. Natl. Acad. Sci.* **87:** 9153–9157.

Mito, Y., Henikoff, J.G., and Henikoff, S. 2005. Genome-scale profiling of histone H3.3 replacement patterns. *Nat. Genet.* **37:** 1090–1097.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. 2003. Targeted recruitment of Set1 histone methylase by elongating Pol II provides

a localized mark and memory of recent transcriptional activity. *Mol. Cell* **11:** 709–719.
Ng, M.C., Miyake, K., So, W.Y., Poon, E.W., Lam, V.K., Li, J.K., Cox, N.J., Bell, G.I., and Chan, J.C. 2005. The linkage and association of the gene encoding upstream stimulatory factor 1 with type 2 diabetes and metabolic syndrome in the Chinese population. *Diabetologia* **48:** 2018–2024.
Ozsolak, F., Song, J.S., Liu, X.S., and Fisher, D.E. 2007. High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.* **25:** 244–248.
Pajukanta, P., Lilja, H.E., Sinsheimer, J.S., Cantor, R.M., Lusis, A.J., Gentile, M., Duan, X.J., Soro-Paavonen, A., Naukkarinen, J., Saarela, J., et al. 2004. Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). *Nat. Genet.* **36:** 371–376.
Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E., et al. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122:** 517–527.
Pray-Grant, M.G., Daniel, J.A., Schieltz, D., Yates III, J.R., and Grant, P.A. 2005. Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. *Nature* **433:** 434–438.
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:** 3435–3447.
Radonjic, M., Andrau, J.C., Lijnzaad, P., Kemmeren, P., Kockelkorn, T.T., van Leenen, D., van Berkum, N.L., and Holstege, F.C. 2005. Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit. *Mol. Cell* **18:** 171–183.
Reppas, N.B., Wade, J.T., Church, G.M., and Struhl, K. 2006. The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol. Cell* **24:** 747–757.
Rodgers, J.T., Lerin, C., Haas, W., Gygi, S.P., Spiegelman, B.M., and Puigserver, P. 2005. Nutrient control of glucose homeostasis through a complex of PGC-1alpha and SIRT1. *Nature* **434:** 113–118.
Roy, A.L., Meisterernst, M., Pognonec, P., and Roeder, R.G. 1991. Cooperative interaction of an initiator-binding transcription initiation factor and the helix-loop-helix activator USF. *Nature* **354:** 245–248.
Ruthenburg, A.J., Allis, C.D., and Wysocka, J. 2007. Methylation of lysine 4 on histone H3: Intricacy of writing and reading a single epigenetic mark. *Mol. Cell* **25:** 15–30.
Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. 2007. Mammalian RNA polymerase II core promoters: Insights from genome-wide studies. *Nat. Rev. Genet.* **8:** 424–436.
Scarpulla, R.C. 2006. Nuclear control of respiratory gene expression in mammalian cells. *J. Cell. Biochem.* **97:** 673–683.
Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. 2006. A genomic code for nucleosome positioning. *Nature* **442:** 772–778.
Shoulders, C.C. and Naoumova, R.P. 2004. USF1 implicated in the aetiology of familial combined hyperlipidaemia and the metabolic syndrome. *Trends Mol. Med.* **10:** 362–365.
Smyth, G.K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3:** Article3.
Taverna, S.D., Ilin, S., Rogers, R.S., Tanny, J.C., Lavender, H., Li, H., Baker, L., Boyle, J., Blair, L.P., Chait, B.T., et al. 2006. Yng1 PHD finger binding to H3 trimethylated at K4 promotes NuA3 HAT activity at K14 of H3 and transcription at a subset of targeted ORFs. *Mol. Cell* **24:** 785–796.
Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. 2004. An abundance of bidirectional promoters in the human genome. *Genome Res.* **14:** 62–66.
Viollet, B., Lefrancois-Martinez, A.M., Henrion, A., Kahn, A., Raymondjean, M., and Martinez, A. 1996. Immunochemical characterization and transacting properties of upstream stimulatory factor isoforms. *J. Biol. Chem.* **271:** 1405–1415.
Vire, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.M., et al. 2006. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439:** 871–874.
West, A.G., Huang, S., Gaszner, M., Litt, M.D., and Felsenfeld, G. 2004. Recruitment of histone modifications by USF proteins at a vertebrate barrier element. *Mol. Cell* **16:** 453–463.
Wu, Z., Puigserver, P., Andersson, U., Zhang, C., Adelmant, G., Mootha, V., Troy, A., Cinti, S., Lowell, B., Scarpulla, R.C., et al. 1999. Mechanisms controlling mitochondrial biogenesis and respiration through the thermogenic coactivator PGC-1. *Cell* **98:** 115–124.
Xu, X., Bieda, M., Jin, V.X., Rabinovich, A., Oberley, M.J., Green, R., and Farnham, P.J. 2007. A comprehensive ChIP chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res.* **17:** 1550–1561.
Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. 2005. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* **350:** 129–136.
Yang, A., Zhu, Z., Kapranov, P., McKeon, F., Church, G.M., Gingeras, T.R., and Struhl, K. 2006. Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol. Cell* **24:** 593–602.
Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Rando, O.J. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309:** 626–630.

Received July 4, 2007; accepted in revised form December 11, 2007.