Supplementary figures

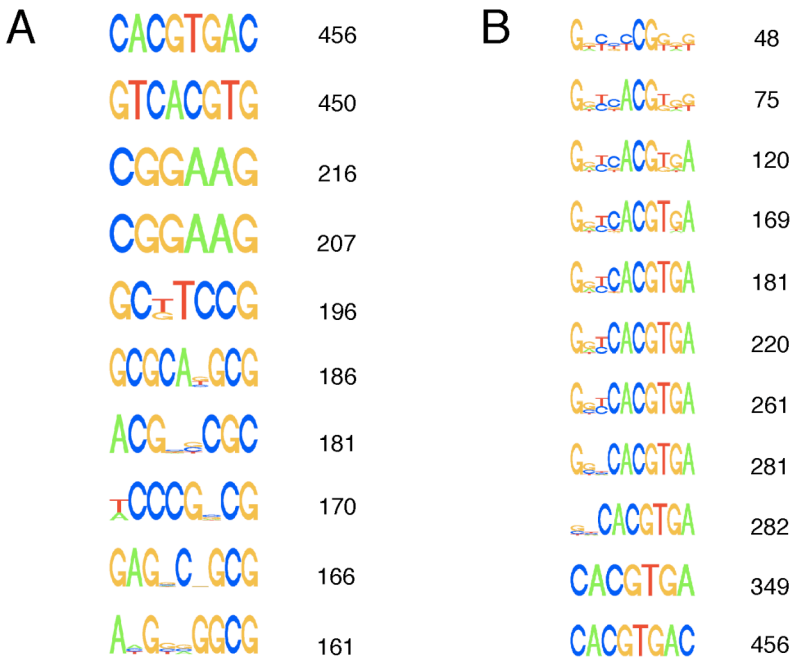

Figure S1. BCRANK results for USF1 whole genome ChIP-chip data on 5211 regions ranked by $\log_2$ signal. a) Top 10 consensus sequences predicted by BCRANK with 25 random restarts. The two highest scoring results are exactly the same consensus sequence, CACGTGAC. Sequences number 3,4 and 5 are similar to the consensus for GABPA, a protein that has previously been suggested to interact with USF1. The same consensus will typically get slightly different scores when found in different restarts, since BCRANK uses random sampling to compute the score (see Supplementary Methods). b) Sequence logos and BCRANK scores for all 12 iterations in the BCRANK search path for the top-scoring consensus. The search starts from GDYBYCTKDK and ends with the locally optimal solution CACGTGAC.

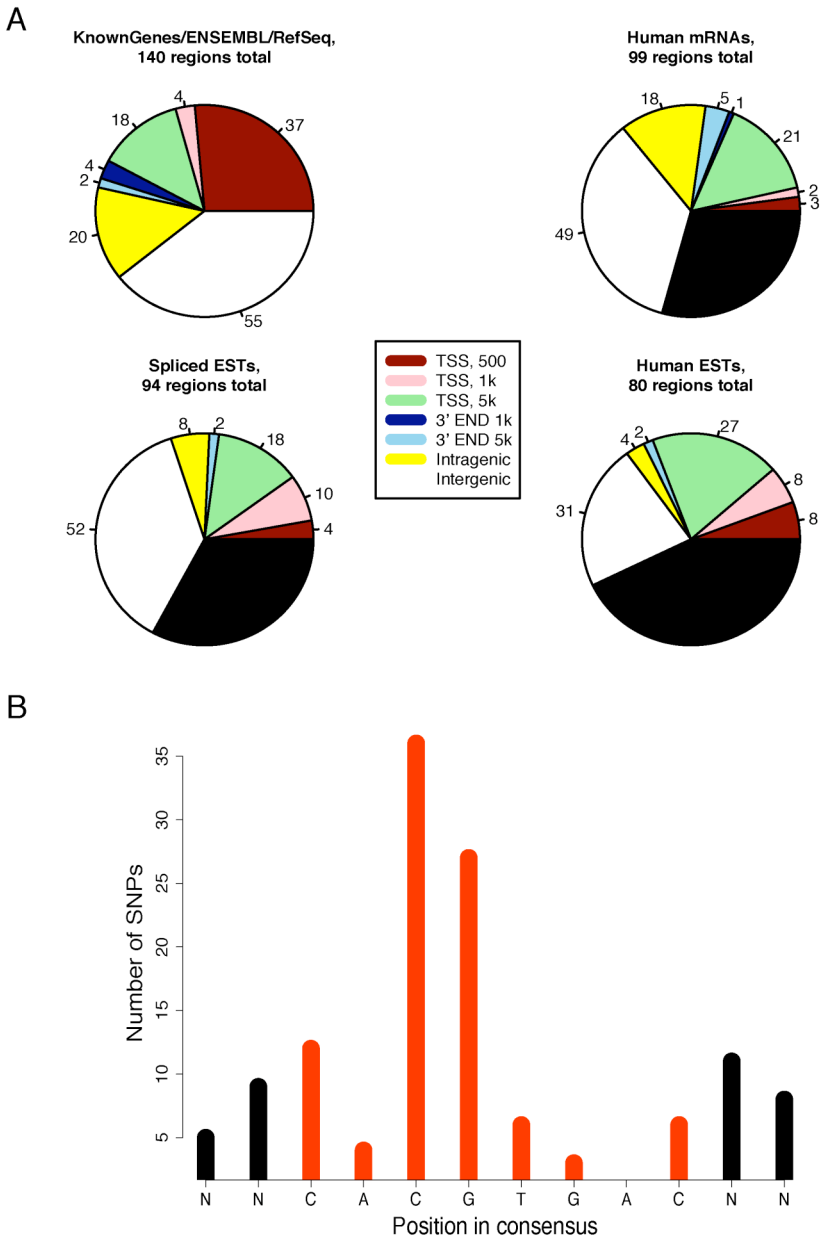

Figure S2. Overall view of 140 SNPs near predicted USF1 binding sites. a) Genomic location of SNPs presented as pie charts. Mapping was first done against Known Genes/ENSEMBL/RefSeq annotations, and subsequently for RNA genes, Human mRNAs/Spliced ESTs and lastly Human ESTs. In each annotation, the regions mapping within 500 bp or 1kb of TSS were not mapped in the following annotation and are indicated by the black portions of the charts. b) Histogram displaying the SNP localizations with respect to the predicted USF1 binding sequence, NNCACGTGACNN. In red are the bases that according to the BCRANK algorithm are important for USF1 binding.

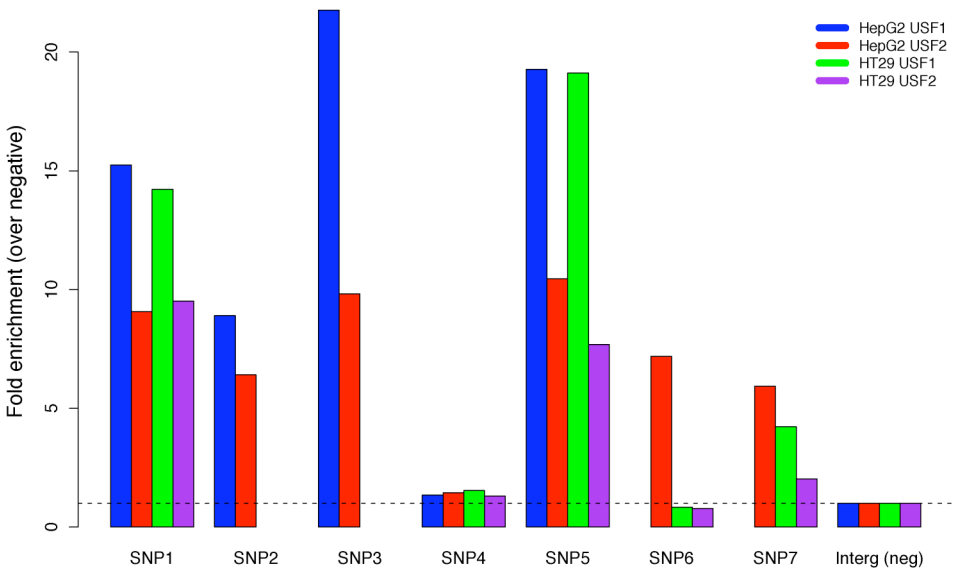

Figure S3: qPCR results showing the enrichment of USF1 and USF2 in regions surrounding SNP1 to SNP7, in HepG2 and HT29 cells. The instances where the signals are absent (for example SNP2 in HT29) indicate that qPCR data is missing. All SNPs are enriched for USF proteins, perhaps with the exception of SNP4 where enrichment is only slightly higher than in the negative region. SNP6 is bound by USFs in HepG2 cells but not in HT29. SNP6 is heterozygous in HepG2, but homozygous for the G allele of C[A/G]CGTGAC in HT29.

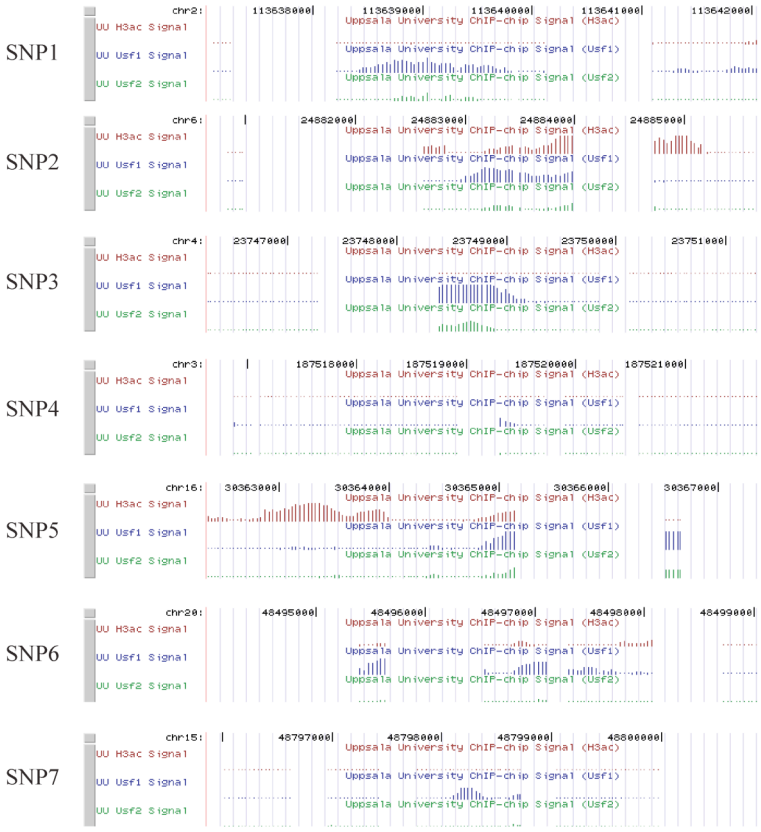

Figure S4: USF1, USF2 and H3ac ChIP-chip signal around SNP1 to SNP7 in HepG2 cells displayed in the UCSC genome browser, in the hg18 (Human Mar. 2006) assembly. Each figure is centered on the SNP. USF1 is enriched near the SNP and in most cases the pattern looks similar for USF2 but with slightly lower signal. This is also shown by qPCR in Figure S2. The H3ac peaks are usually not located at the same positions as the USF peaks.

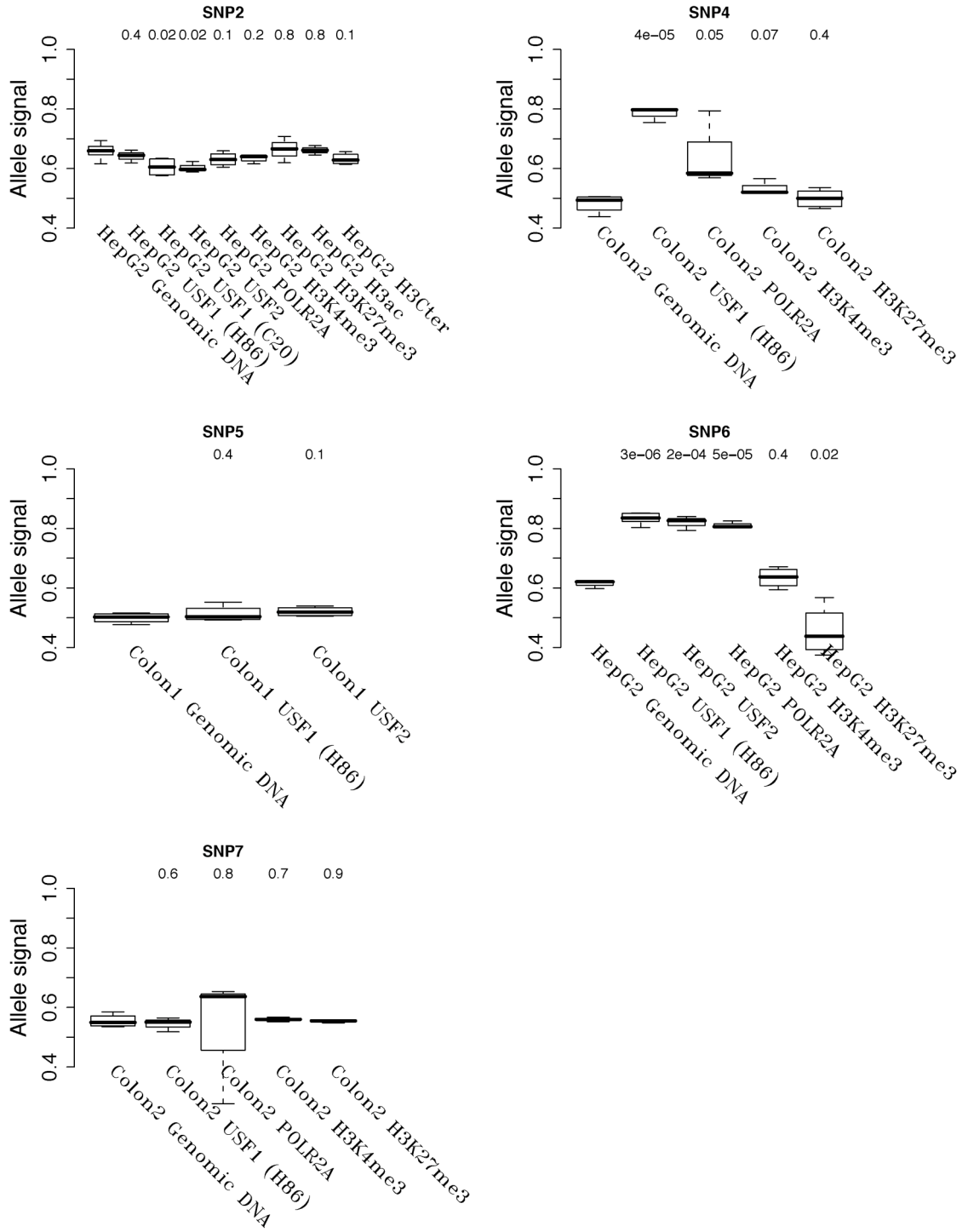

Figure S5: Quantification results for SNP2 and SNPs 4 to 7. At the top of each box are p-values from a t-test, indicating whether allele signal in the ChIP sample is significantly different from that detected in the genomic DNA.

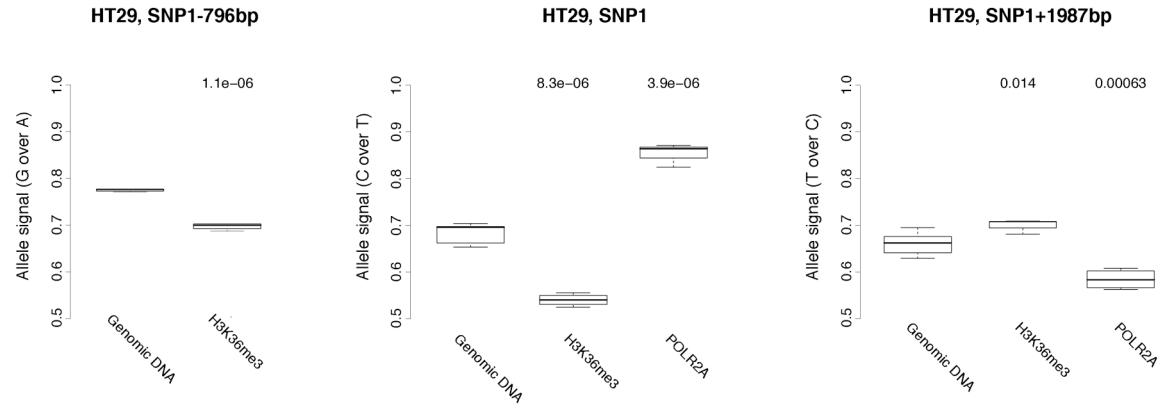

Figure S6: Quantification results for SNP1-796, SNP1 and SNP1+1987 in HT29 cells for POLR2A and HeK36me3. At the top of each box are p-values from a t-test, indicating whether allele signal in the ChIP sample is significantly different from that detected in the genomic DNA. Which of the two alleles that the signals come from is indicated in the y-axes labels.

Supplementary tables

	Top consensus	# sites	E-boxes	In stringent	In relaxed	Second consensus
BCRANK 6	CACGTGAC	1757	1757 (100%)	1194 (68%)	1529 (87%)	CGAACG
BCRANK 8	CACGTGAC	1757	1757 (100%)	1194 (68%)	1529 (87%)	CGGAAG
BCRANK 10	CACGTGAC	1757	1757 (100%)	1194 (68%)	1529 (87%)	CGGAAG
MDscan 6	CACGTG	7724	7724 (100%)	4178 (54%)	6020 (78%)	ACGTGA
MDscan 8	TCACGTGA	5618	1176 (21%)	2145 (38%)	3375 (60%)	CACGTGAC
MDscan 10	TCAAGTGATC	2270	621 (27%)	922 (41%)	1401 (62%)	GGATCACTTG
DRIM 6	ACGTGA	3894	3013 (77%)	2261 (58%)	3112 (80%)	CCGCCC
DRIM 8	CACGTGAC	1757	1757 (100%)	1194 (68%)	1529 (87%)	CGCNNNCGC
DRIM 10	ACGNGAC	2737	1754 (65%)	1594 (58%)	2189 (80%)	CGCNNNCGC

Table S1. Results for BCRANK, MDscan and DRIM on the 5211 ranked USF1 enriched regions. The methods were run with motif length 6, 8 and 10. The number of top sequences to look for candidate motifs was set to 100 for MDscan; otherwise default settings were used both for MDscan and DRIM. For BCRANK, penalty P1 was used. The third column contains the number of binding sites predicted for the top consensus. Column four shows how many of the sites that match an E-box sequence (CACGTG). In columns five and six are the numbers of sites that are within the stringent and relaxed USF1 regions as defined in a previous study. The last column shows the second highest scoring consensus.

Name	SNP label	Sequence	Promoter	Heterozygous in	ChIPs performed
SNP1	rs1867760	AA[T/C]ACGTGACCC	-	HepG2	USF1, H86 (6) + USF1, C20 (4) + USF2 (4) + POLR2A (5) + H3K4me3 (4) - H3K27me3 (4) - H3ac (4) H3Cter (6)
				HT29	USF1, H86 (4) + POLR2A (3) +
SNP1 -91	rs1867761	TAGAG[T/C]GTGGGT	-	HepG2	USF1, H86 (3) + USF2 (3) + POLR2A (3) H3K27me3 (3)
				HT29	USF1, H86 (4) + USF2 (4) +
SNP1 -796	rs724496	GGACT[G/A]GGTAC	-	HT29	H3K36me3 (4) +
SNP1 +1987	rs4849159	ACATG[T/C]GCTCAG	-	HT29	USF1, H86 (4) USF2 (4) POLR2A (4) - H3K36me3 (4)
SNP2	rs2754775	A[C/A]CACGTGACCA	GMNN	HepG2	USF1, H86 (3) USF1, C20 (4) USF2 (3) POLR2A (4) H3K4me3 (4) H3K27me3 (3) H3ac (4) H3Cter (4)
SNP3	rs16875109	CTCA[T/C]GTGACCT	-	Colon1	USF1, H86 (6) + USF2 (7) + POLR2A (3) + H3K4me3 (3) H3K27me3 (3)
SNP4	rs1544702	CTCAC[G/A]TGACAT	-	Colon2	USF1, H86 (3) + POLR2A (3) H3K4me3 (3)

					H3K27me3 (5)
SNP5	rs4787645	AGCACGTGAC[G/A]T	SEPHS2	Colon1	USF1, H86 (4) USF2 (4)
SNP6	rs11696955	GAC[A/G]CGTGACTT	-	HepG2	USF1, H86 (5) + USF2 (3) + POLR2A (3) + H3K4me3 (4) H3K27me3 (4)
SNP6+686	rs16995222	AAGGC[T/C]GACTC	-	HepG2	H3K36me3 (4)
SNP6+1366	rs2869991	ATCTG[T/C]GGAAA	-	HepG2	H3K36me3 (4)
SNP7	rs9920753	TTCACGTG[A/T]CAA	-	Colon2	USF1, H86 (3) POLR2A (3) H3K4me3 (3) H3K27me3 (3)

Table S2. Summary results for 12 heterozygous SNPs. Underlined bases in sequences are predicted USF1 binding sites. Column four indicates the cases where the SNP is in the promoter of a protein coding gene (PCG), at a distance of at most 1kb. The two last columns contain information on the samples where the SNP was heterozygous and the ChIP experiments performed for each heterozygous SNPs. In parenthesis is the number of replicates performed, and the + and – in the last columns mark the ChIP samples that contain a significantly higher or lower signal of the predicted USF1 bound allele when compared to genomic DNA. With the exception of SNP7, all of the SNPs occurring inside the core sequence, CACGTGAC, show a positive effect for USF1. Many of them also show positive effects for USF2 and POLR2A. Two of the SNPs, SNP2 and SNP5, are in promoters of genes but in both cases the SNP is outside of the CACGTGAC core binding sequence and no effects are detected.