Supplementary methods

ChIP

HepG2 cells were cultured in RPMI 1640 medium supplemented with 10% non-inactivated FBS, l-glutamine and PEST (Sigma-Aldrich) at 37°C with 5% CO_2. Antibodies were from Santa Cruz Biotechnology (sc-6554, sc-6556 and sc-13442) and ChIP was performed using 10^7-10^8 cells per reaction. Crosslinking of protein and DNA was done in room temperature for 10 minutes with serum free media containing 0,37 % formaldehyde. Cells were scraped and washed in PBS before treatment of cell lysis buffer with protease inhibitors for 10 min on ice. Pelleted nuclei were resuspeded in RIPA buffer and a BioRuptor (Diagenode) was used to sonicate DNA with 2-3 * 15 cycles of 30 seconds on/off at the highest amplitude. For FOXA2 two biological replicates FOXA2 were done using slightly different protocols. For the first replicate, chromatin from a sub-confluent T-175 flask was sonicated to a broad range of fragments, with a median size close to 500 bp, and ChIP was performed in parallel for FOXA2 and IgG with 10 µg antibody. ChIP DNA was gel fractionated and amplified to create two size libraries corresponding to insert sizes of 50-150 bp and 250-450 bp. For the second replicate, chromatin was sonicated to a median size around 200 bp. DNA obtained from two precipitations using 20 µg of the ChIP-grade antibody (sc-6554x) was pooled. For all ChIPs washing was done four times with RIPA, once with LiCl buffer and once with TE-buffer and DNA-protein complexes were eluted from beads twice with 225 µl 0.1 M $NaHCO_3$ and 1 % SDS and treated with RNAseA at 65°C for 6 hours and Proteinase K at 45°C over night. For the large IgG library one lane was used for sequencing, for all other libraries two lanes were used. For HNF4a the ChIP DNA obtained from two replicates was pooled and concentrated on silica columns by centrifugation using a Quiagen PCR-purification kit before library construction. For GABP, HNF4a and the second FOXA2 replicate one library with insert sizes between 100 and 300 bp was sequenced using the Illumina 1G instrument according to the manufacturers protocol for 36 bp reads. For ChIP-PCR of GABPa an additional ChIP was performed using a monoclonal antibody (sc-28312).

Co-immunoprecipitation

HepG2 cells were washed twice with cold PBS, resuspended in a buffer containing 100mM Tris HCl pH-8.0, 100mM NaCl, 0.2% NP-40, and protease inhibitor cocktail for 30 min on ice and centrifuged in cold to collect nuclei. Pellets were washed once with RIPA and resuspended in the same buffer (1XPBS, 1% NP-40, 0.5% Na-deoxycholate, 0.1% SDS, 0.004% Na-azide, and protease inhibitor cocktail), followed by homogenization in a Dounce homogenizer. Precleared nuclear lysate was incubated with antibody or IgG at a concentration of 1µg/ 500µg of total protein and 50µl Protein G-agarose at +4°C over night. Immune complexes were washed twice with RIPA buffer and eluted in NuPAGE® LDS sample buffer containing reducing agent (Invitrogen) at 70°C for 10 min. Samples were separated on NuPAGE® 4-12% for Western blotting (ECL detection system). A dilution of 1:200 was used for all primary antibodies except the one against HNF4α which was diluted to 1:1000. Secondary antibody dilutions were 1:10000 for anti-goat antibody (Santacruz), 1:5000 for anti-rabbit antibody (Santacruz) and 1:1000 for anti-mouse antibody (ECL detection system). In order to check for equal loading and a negative control for the CoIP experiments one blot was stripped with 1XPBS, 0.1% Tween-20, 0.2% SDS for 30 min at 65°C and incubated with anti GAPDH antibody at a dilution of 1:10000.

FDR calculation

Fragments are supposed to be Poisson-distributed along the mappable genome with a global coverage level λ equal to the product of the number and mean length of mapped extended fragment divided by the mappable genome length. The probability of observing overlapping fragments of certain heights can then be calculated as Poisson probabilities. The mappable genome size was divided by the mean fragment length in order to determine a number of windows, *r*, in which randomly placed fragments could overlap. For each peak-height the expected number of peaks with this height was calculated as *r* multiplied with the Poisson probability. Finally the FDR for each height was defined as the ratio between the sum of the expected number of peaks and the observed number of peaks of that height or higher.

Supplementary figures

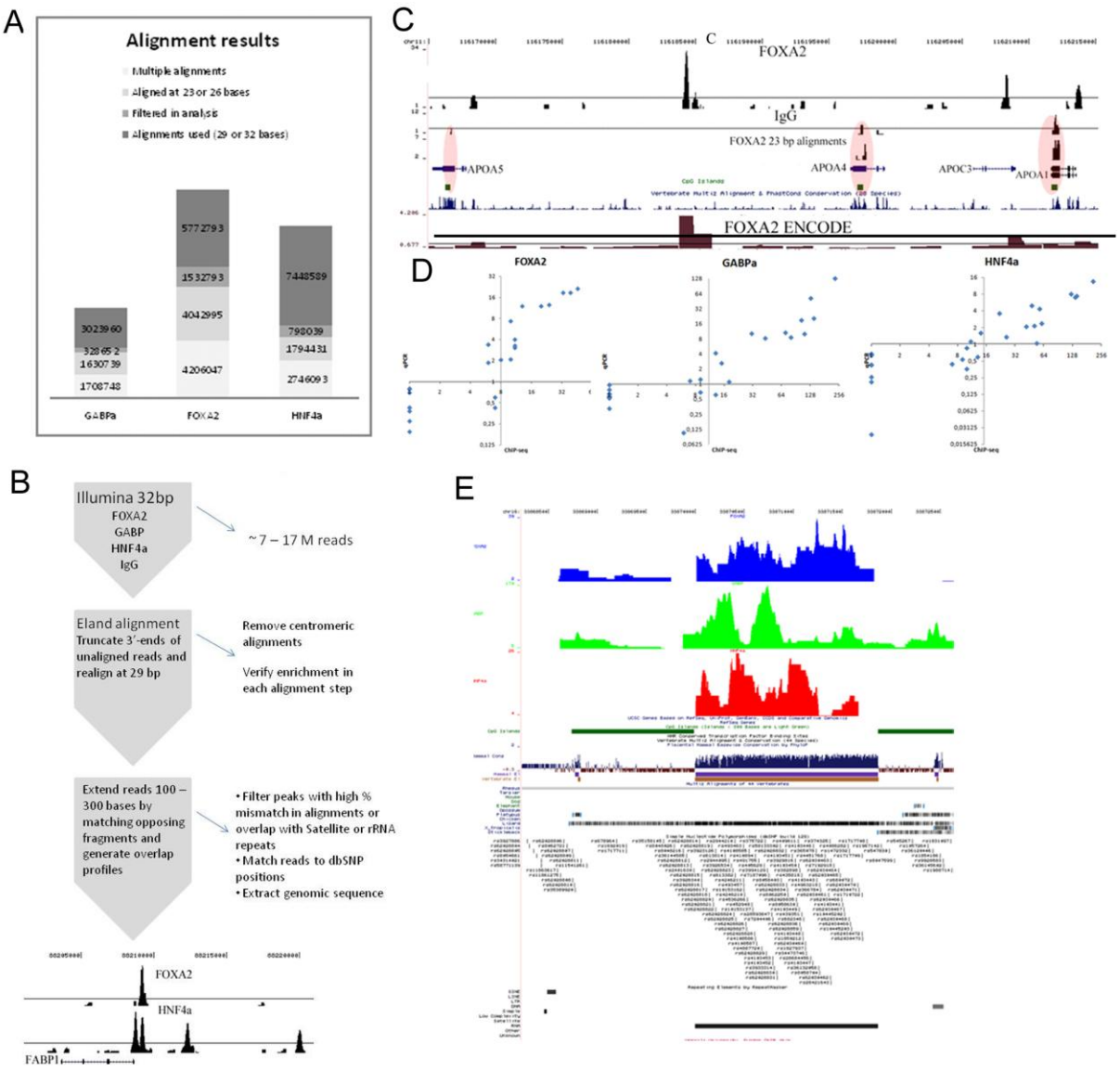

Figure S1. (A) Number of aligned and filtered reads in the different datasets. (B) Alignments were done at 32, 29, 26 and 23 bases using Eland and alignments that were either present in the IgG dataset ot within 1 Mb of centromeric gaps were filtered. Reads were extended to at most 300 bp for GABP and HNF4a and to 450 bp for FOXA2 (first replicate), and peaks overlapping with Satellite or rRNA repeats or with an average mismatch score > 2 were removed. Reads overlapping with dbSNP positions was used to tract genotype information, and 300 bp sequences centered on peak maxima was etracted for motif analysis. (C) Custom track showing results for FOXA2 and IgG ChIP-seq compared to ENCODE ChIP-chip results. All four positive ENCODE results was replicated by ChIP-seq. The FOXA2 23 bp alignment track illustrates that FOXA2 and IgG signals occur at CpG islands when reads are aligned using the first 23 bp only (highlighted regions). (D) Quantitative PCR results for selected regions for HNF4a, FOXA2 and GABP. X- and Y-axis are set to represent the cut-offs used, with peaks positive in both qPCR and ChIP-seq in the upper right quadrant. (E) Signals in rRNA- and centromeric repeats were removed as they give rise to a high number of peaks in all datasets.

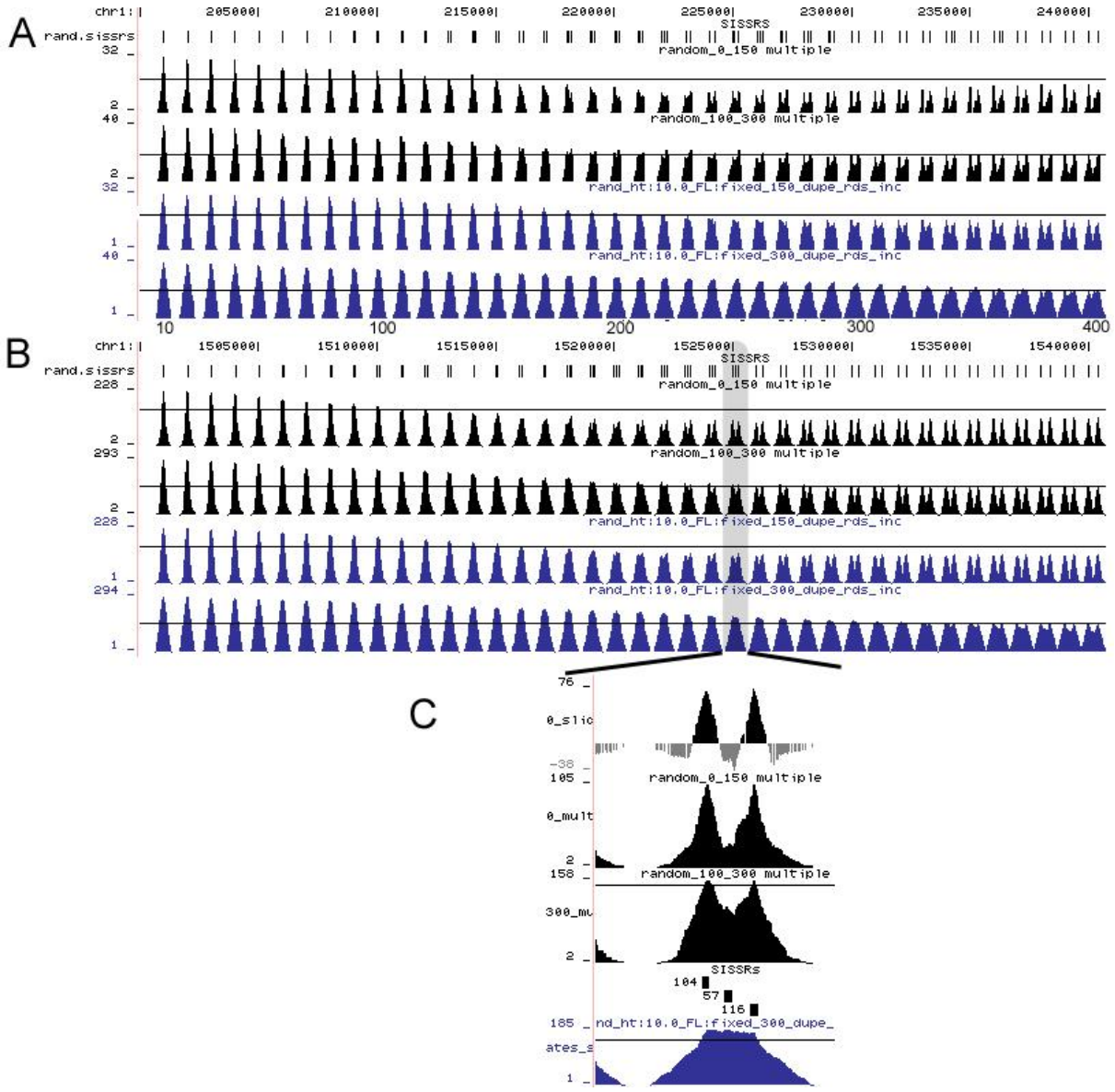

Figure S2. Effect of read matching on simulated data. To visualize differences in peak calling algorithms we generated peaks from random positions using a similar distribution as in the sequencing (simulating 100-300 base fragments) and placed two copies of the peak at an increasing distance, with the first peak centered on every 1000th base and . Distances between peaks are from 10 to 400 bases in A) and B) above. In A) we simulated 20 reads and in B) 150 reads. Peak calling results are visualized in UCSC genome browser for from top to bottom: SISSRs, our algorithm (black) by matching and extending reads within 0-150 bases and 100-300 bases, and in blue corresponding results using fixed extension of 150 and 300 bases (visualized with FindPeaks 3.2). SISSRs was able to identify the two separate peaks with ~100 base separation but in many cases reported an extra peak in between. Fragment matching gave a better separation of peak compared to the fixed extension and was also better for predicting the correct peak height as indicated by the horizontal lines. C) A closer view of the peaks with 150 reads separated by 250 bases where the fixed extension gave the highest point at the midpoint between the peaks. A better separation of peaks can be achieved by calculating the difference of the number of reads directed towards and away from each position, as visualized in the top track using a 200 bp window.

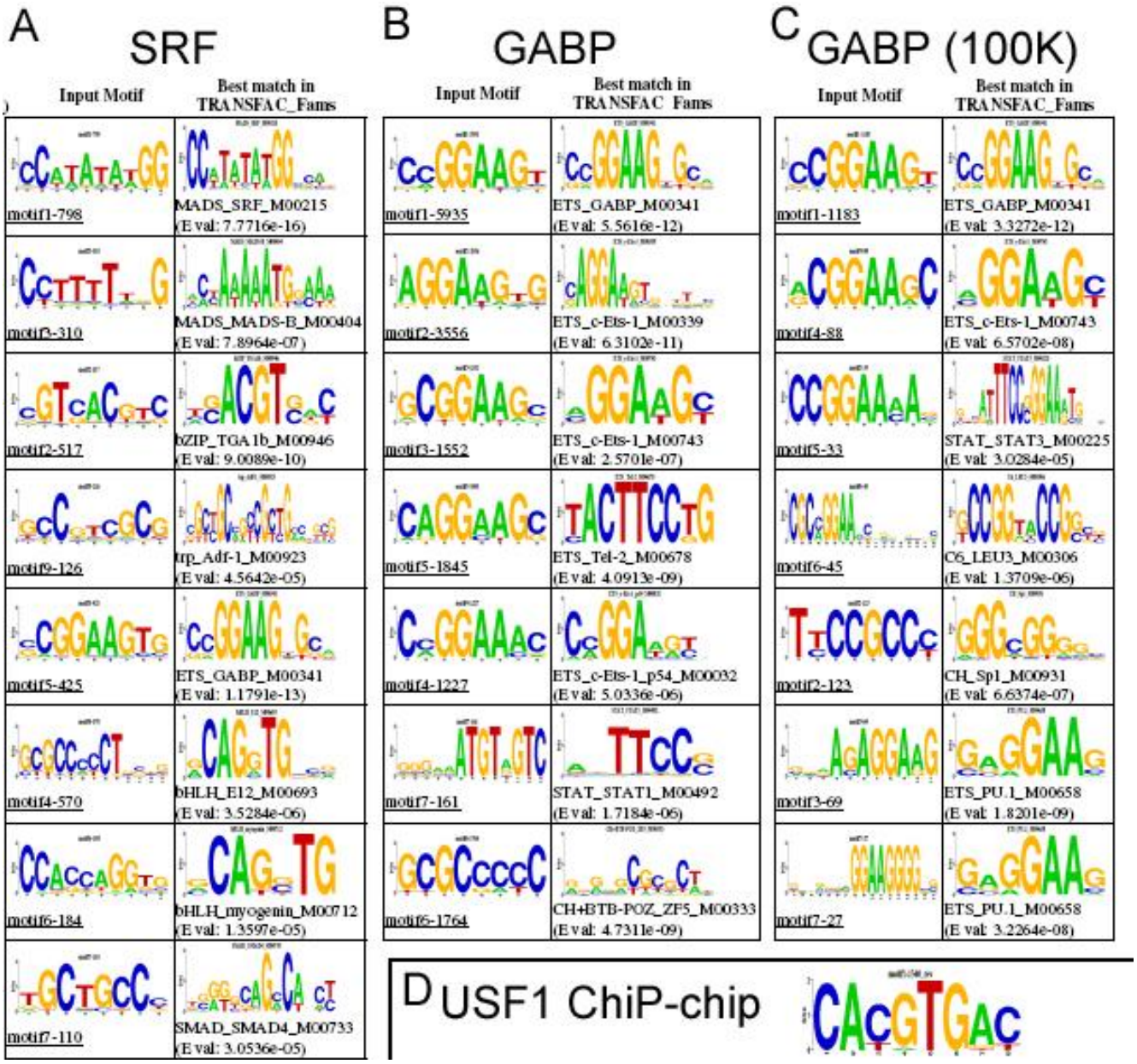

Figure S3. Results of the proposed motif finding algorithm on public datasets. A-C) Alignments were downloaded from the Valouev *et al.* paper and peaks and motifs were identified as described in the main text by repeatedly searching for and removing the peak with the most enriched motifs of a given length. For each peak the centermost motif match was used to define the matrix and for each motif the number of peaks it was centered in is reported. For SRF a word length of 9 with 2 mismatches was used and for GABP the same settings as for our dataset was used (word length 8 and 1 mismatch). In concordance with the Valouev paper we could identify the GABP / ETS motif independently of the SRF motif in that dataset. For GABP we first used a larger dataset containing > 19000 peaks, and find that the GABP motif is present in more than 14000 peaks. We did not identify the SP1 motif as was reported previously for this dataset, indicating that this motif was identified by co-localisation in the CpG islands since our method uses the sequence surrounding peak centers as a background. Due to the high enrichment of reads in GABP peaks we found that to correctly identify GABP peaks and motifs no more than 100 000 ChiP-seq reads was needed (C). D) To test the method on lower resolution data, chip-chip data from Iglesias *et al*. was used with a center distance of 250 bp and a background at positions 500 – 750 from peak centers. Using a word length of 8 and allowing 1 mismatch among sequences with at least 2-fold enrichment both the CACGTGAC sites as predicted by BCRANK in the original publication and additional e-boxes (CAC[G/A]TGNN) were identified, with a run time of less than 2 minutes for 2000 sites as compared to > 3 h with the suggested BCRANK settings.

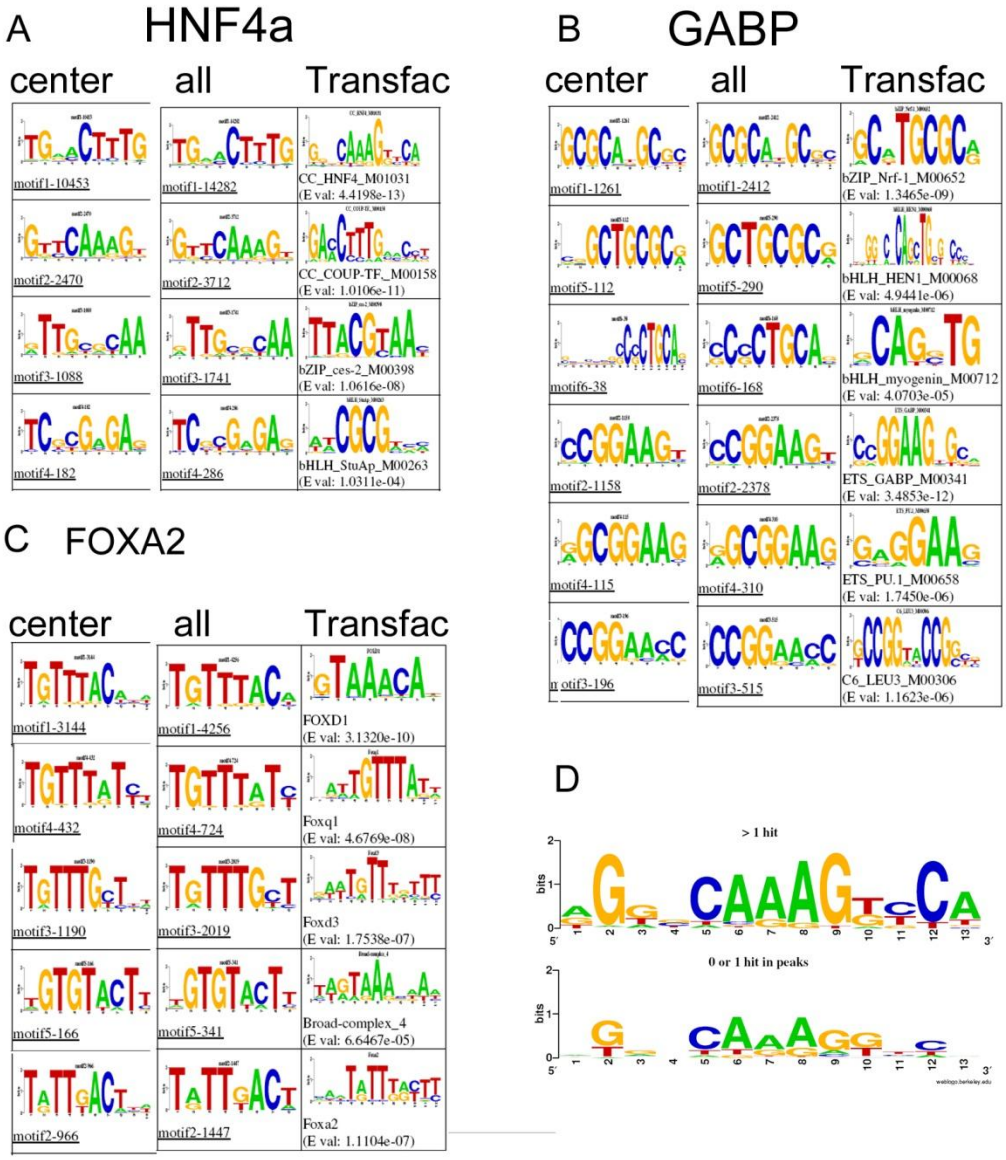

Figure S4. Results of motif discovery in peaks. A-C) For each factor motifs were generated from the most centered and from all hits within 150 bp. D) shows the difference in information content for known binding sequences with > 1 hit compared to those with 0 or one hit in peaks.

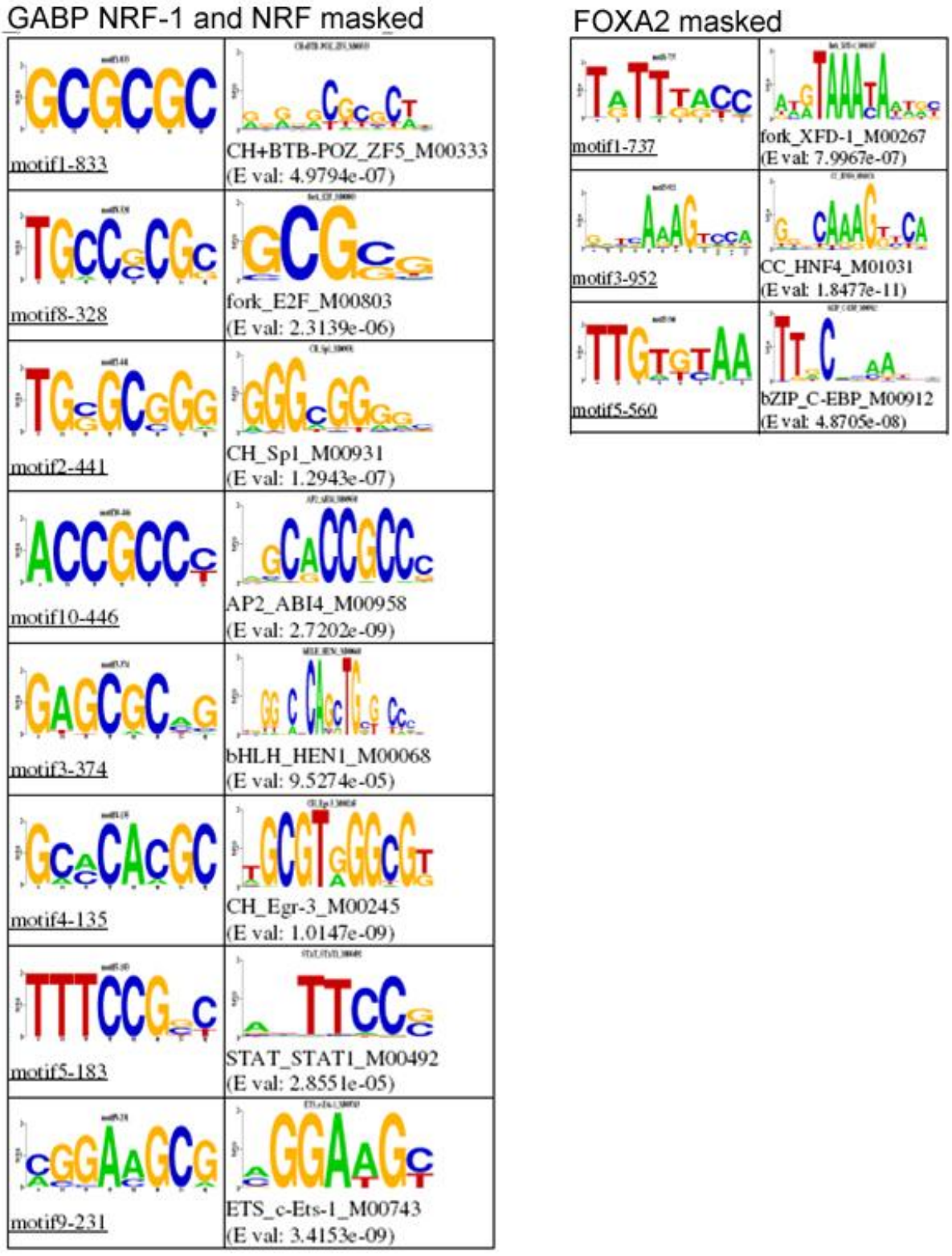

Figure S5. Motifs identified in peaks masked for GABP and FOXA2 sequences.

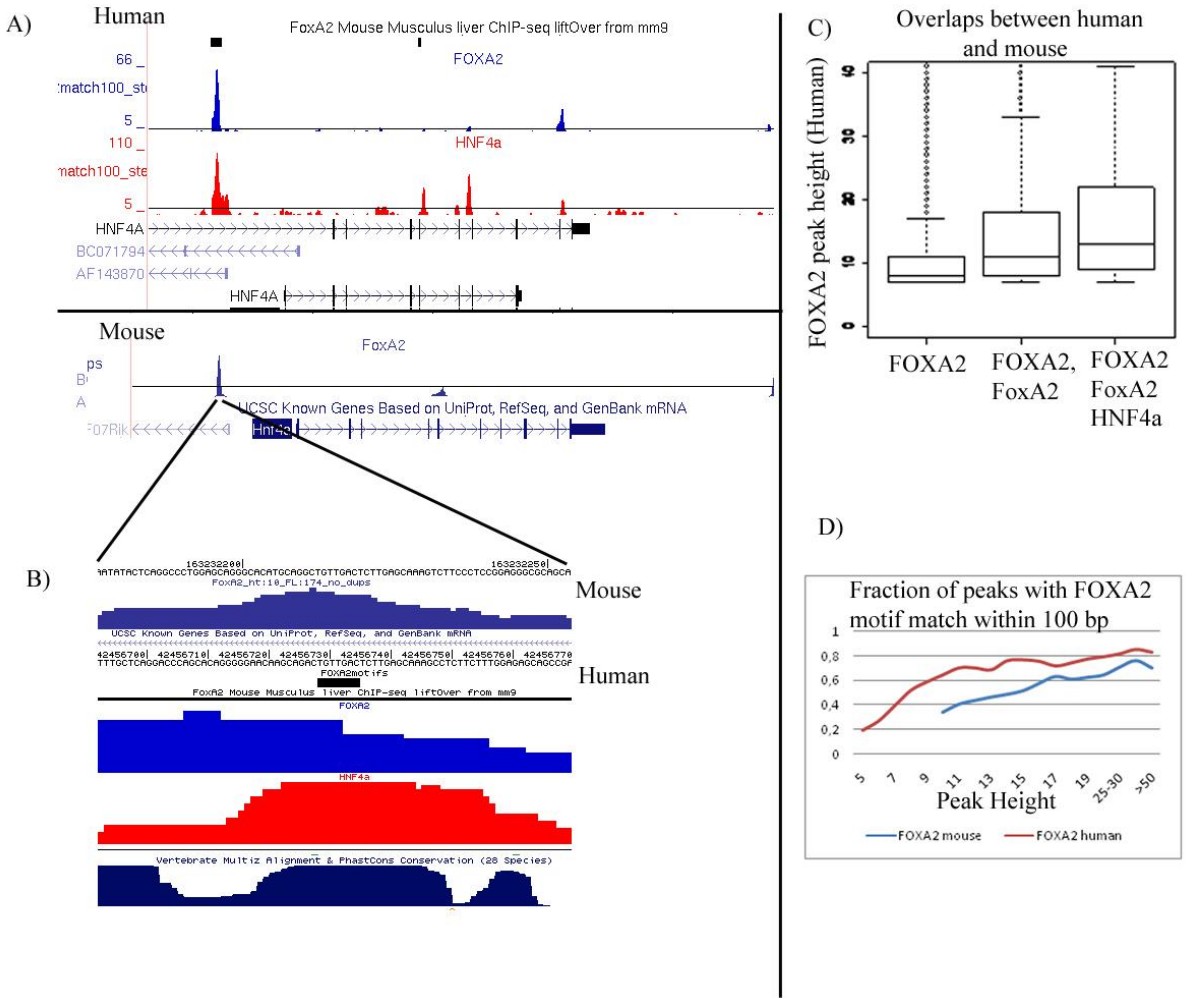

Figure S6. FOXA2 ChIP-seq in human and mouse liver. A) Overview of the bindings for FOXA2 and HNF4a in human FoxA2 in mouse for the HNF4a gene. B) Base-pair view of the identified binding sites show that FOXA2 binds to the same bases in the human and mouse genome in a region with high conservation levels and that HNF4a also binds this region . C) Peak heights for FOXA2 when either FoxA2 or FoxA2 and HNF4a signals are present. D) Presence of motif sequences in peaks for foxa2 in mouse and human for different peak heights.

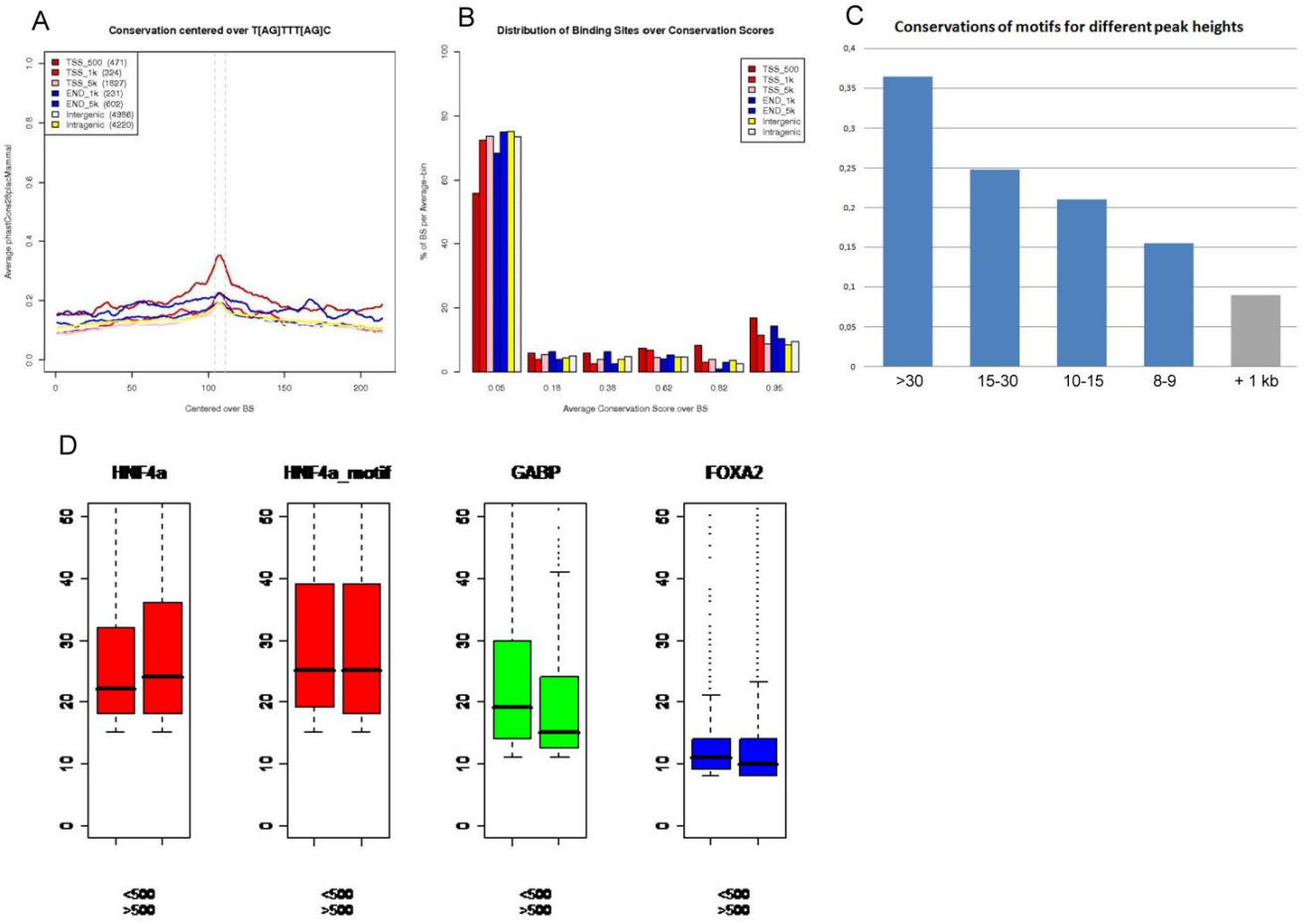

Figure S7. A) Conservation scores centered on FOXA2 sites at different genomic locations. B) Most sites had low conservation scores. C) An increase in conservation scores was seen for FOXA2 sites with increasing peak heights. D) Peak heights at TSS (<500 bp) and distal positions for the three TFs and for HNF4a peaks with and without motifs. HNF4a signals were lower at TSS, however this was not the case for sites containing the motif.

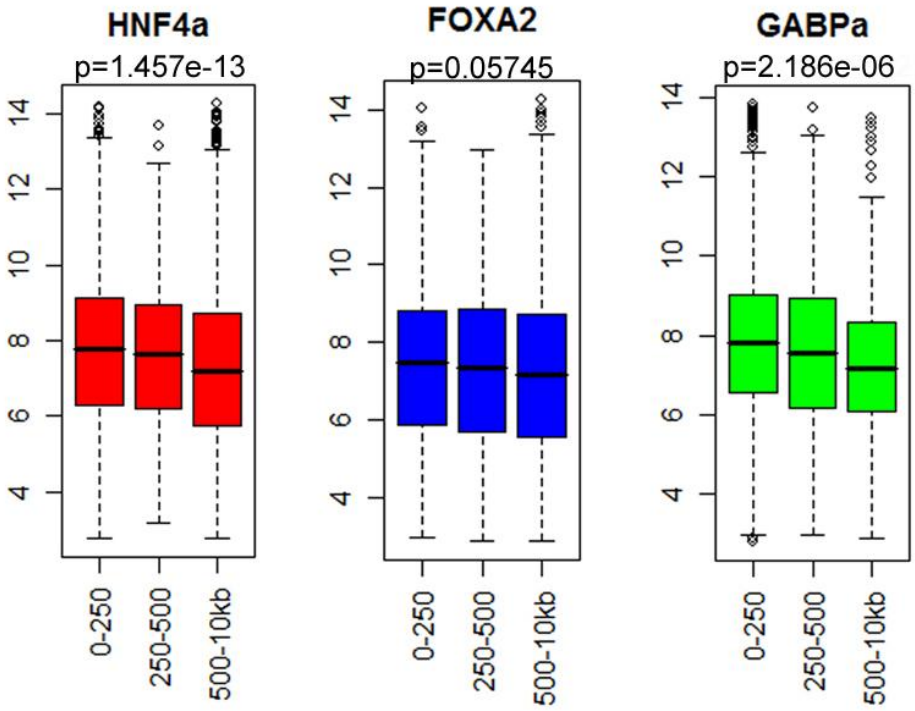

Figure S8. Expression levels for genes with bindings of the different TFs in the promoter region. Genes were grouped by distance to TF (0-250 bp, 250-500 bp and 0.5 – 10 kb) and for all factors a trend was seen with the highest expression for genes bound close to TSS, with p-values shown for differences in mean between location 0-250 bp and 0.5-10kb.

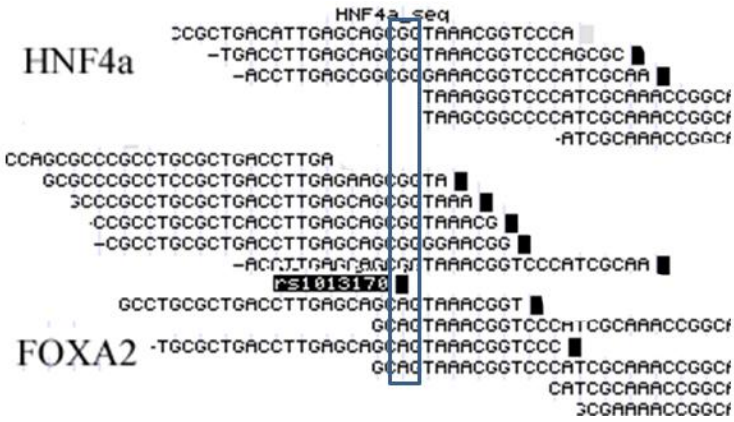

Figure S9. Potential differential binding of HNF4a and FOXA2 is found at rs1013170 with aligned reads visualized using UCSC browser. All 8 HNF4a reads are from the G allele whereas all 4 FOXA2 reads are from the A allele.

Table S1. Additional SNPs from GWAS studies within HNF4a and FOXA2 peaks.

SNP	HNF4a	FOXA2	GENE	Trait	Author
rs6060373	124		GDF5	Height	Weedon
rs5029939	58		TNFAIP3	Systemic lupus erythematosus	Graham
rs6983267	45		Intergenic	Colorectal cancer, Prostate cancer	Tomlinson, Yeager
rs10411210	36		RHPN2	Colorectal cancer	COGENT Study
rs9929218	36		CDH1	Colorectal cancer	COGENT Study
rs2054989	20		ERC2	Hip geometry	Kiel
rs17466684		18	CLU	Panic disorder	Otowa

Table S2. SNPs with preference for binding to one allele. The sequence at dbSNP positions was extracted from aligned reads and compared to results from HepG2 genotyping (Illumina 1M array). SNPs with at least 4 reads and at least 80 % from the major allele are listed.

Freq.	SNP sequence	SNP	HepG2 genotype	
HNF4a				
1	GGGGGGGG	rs1013170	A	G
1	CCCC	rs11206830	A	G
1	AAAA	rs12703055	A	G
1	TTTT	rs1615752	A	G
1	AAAAA	rs17236536	A	G
1	TTTT	rs1756867	A	G
1	GGGGG	rs17764251	A	G
1	TTTT	rs1800759	A	C
1	GGGGG	rs1871428	A	G
1	CCCC	rs2265190	A	G
1	AAAA	rs2297208	A	G
1	AAAA	rs2305062	A	G
1	TTTT	rs2372391	A	G
1	TTTT	rs2376585	A	G
1	AAAAAAA	rs2584624	A	G
1	CCCC	rs3810223	A	G
1	AAAA	rs3905288	A	G
1	AAAA	rs4383238	A	C
1	TTTT	rs4559324	A	G
1	CCCC	rs463260	A	G
1	AAAA	rs464865	A	G
1	GGGGG	rs474513	A	G
1	GGGG	rs4855884	A	G
1	GGGGG	rs4931536	A	C
1	CCCC	rs4943130	A	G
1	GGGGG	rs550015	A	G
1	TTTT	rs6470263	A	G
1	AAAA	rs6510490	A	G
1	AAAAA	rs6648	A	G
1	CCCC	rs687339	A	G

1	TTTT	rs6925895	A	C
1	AAAA	rs6950216	A	G
1	TTTT	rs7927381	A	G
1	CCCC	rs7954059	A	G
1	AAAAA	rs8083331	A	C
1	GGGG	rs818708	A	G
1	GGGG	rs9321934	A	G
0.8	AGAAA	rs10252244	A	G
0.8	GAAAA	rs10409728	A	G
0.8	TTTTG	rs10888585	A	C
0.8	AGGGG	rs11167821	A	G
0.8	TCCCCCTCCC	rs12037379	A	G
0.8	AAGAA	rs12549987	A	G
0.8	GGGAG	rs13174863	A	G
0.8	CCTCC	rs1334738	A	G
0.8	GGAGG	rs17251434	A	G
0.8	TTTTC	rs1967017	A	G
0.8	CCCAC	rs2038229	A	C
0.8	GAGGG	rs2120192	A	G
0.8	TCTTT	rs2322602	A	G
0.8	AAAAAGGAAA	rs281032	A	G
0.8	AAAAG	rs3782735	A	G
0.8	AGGGG	rs4618520	A	G
0.8	GAAAA	rs647651	A	G
0.8	GGAGG	rs6737440	A	G
0.8	AGGGG	rs6746569	A	G
0.8	CCCTC	rs737148	A	G
0.8	AAAGA	rs7572990	A	G
0.8	CTCCC	rs7583930	A	G
0.8	AAAAG	rs868943	A	G
0.8	CTCCC	rs9964346	A	G
0.82	AGGGGGGGGA	rs11236359	A	G
0.83	GGGGGA	rs2275900	A	G
0.83	CCTCCC	rs7021876	A	G
0.83	TTGTTT	rs748282	A	C
0.83	CTTTTT	rs9692401	A	G

0.86	TTTTTCT	rs1893441	A	G
0.86	TCCCCCC	rs2473375	A	G
0.875	CCCCTCCC	rs13031757	A	G
0.875	CTTTTTTT	rs2037147	A	G
0.875	AAAGAAAA	rs4920380	A	G
0.875	AGGGGGGG	rs539298	A	G
0.9	CCCCCCACCC	rs198461	A	C
FOXA2				
1	AAAA	rs1013170	A	G
1	CCCC	rs1983891	A	G
0.8	TCTTT	rs4559324	A	G
GABP				
1	GGGG	rs1880948	A	G
0.83	CCCCCT	rs12772979	A	G
0.84	TTTTTG	rs2278497	A	C