Supplementary Text

The 522 significant k-mers are sufficient to classify the enhancers from random sequences

We speculated that if the top k-mers represent the binding sites of liver-specific TFs, they should be capable to differentiate HepG2 enhancers from a random set of sequences. To validate this assumption, we trained a Support Vector Machine (SVM) classifier with a Gaussian kernel on the clusters of the top k-mers, with each feature representing the count of a k-mer cluster in HepG2 enhancer sequence. A fully independent 5-fold cross validation was applied to test the accuracy of the classifier. An independent set of k-mers was extracted every time from the training set, and sequences that has not been used for training were used for testing. The overall accuracy (measured as the area under receiver operating characteristic curve, AUC) of this approach is 85% (Figure S25). The union of the top k-mers obtained from each training set of the cross-validations contain 542 k-mers in total, which are largely overlapped with the 522 top k-mers enriched in all HepG2 enhancers (Figure S26).

To further evaluate the capability of the top 522 k-mers to differentiate enhancers from random sequences, the P300 peaks located outside the HepG2 enhancers were used for 5-fold cross validation test, with features representing the count of the k-mer clusters of the 522 top k-mers in enhancer sequence. P300 is a co-activator and its binding can accurately identify enhancers (Visel et al. 2009). The classifier has had a high overall accuracy of 0.79 (AUC) (Figure S4), indicating that the top k-mers could be used to distinguish HepG2 enhancers from random sequences.

Correlation between the p-value threshold and abundance of fKMPs/fRSPs

For the p-value threshold of 1e-3 with multiple-test correction we used in the manuscript, 97.6% of enhancers have top k-mers, and 96.7% of enhancers have fKMPs, 87.2% of enhancers have fRSPs. To show the dependency of the abundance of fKMPs/fRSPs on the threshold of p-value used in defining the significant k-mers, different cutoffs of the p-values were tested. For the p-value threshold 0.01 without multiple-test correction, 99.8% enhancers have top k-mers, 98.3% of enhancers have fKMPs, and 98.2% of enhancers have fRSPs. If we increase the stringency 100 fold (1e-5 with multiple-test correction), 94% enhancers have top k-mers, 92.8% of enhancers have fKMPs, and 68.7% of enhancers have fRSPs.

Deleterious effect of 555 engineered genetic variants on enhancer activity

In the section of evaluating the deleterious effect of KMPs and RSPs using the data of a massively parallel reporter assay, there are 580 engineered genetic variants in total. Besides the 413 genetic variants overlapping with our mutation system (153 KMs at fKMPs, 71 KMs at sKMPs, 35 RSs at sRSPs, 3 RSs at fRSPs, and 151 controls), 167 variants on the engineered enhancers did not belong to the above three categories of our predictions, including 70 Drop Within Significant (DWS) mutations (mutations that decrease the binding significance of the original k-mer but keep the mutated k-mer within the significant k-mer set), and 72 Increase Within Significant (IWS) set mutations (mutations that increase the significance of the top k-mers). The remaining 25 variants located in the repetitive genomic region were filtered out by our method and thus were left out of analysis. The impact of DWS mutations on enhancer activity was weakly stronger than the control set (Mann-Whitney test P = 0.01006) compared to that of KMs. Similar to the RS mutations, the IWS mutations also do not show significant increase on enhancer activity (Mann-Whitney test P = 0.06667) (Figure S27A).

As for examining the tendency to diminish the enhancer activity of mutations in all six categories of engineered enhancer variants, the DWS mutations have larger proportion (58.6%) of deactivating mutations than control mutations, but smaller than that of KMs (65.2%). In contrast, similar to RS, IWS mutations have the second smallest portion (33.3%) to decrease the enhancer activity (Figure S27B). In all, 65.6% of KMs at fKMPs cause decreasing expression, and 64.4% of KMs at sKMPs lead to decreasing expression (Figure S27B), indicating KMs at fKMPs and sKMPs have similar tendency on decreasing enhancer activity.

Also, to evaluate the accuracy of KM predictions, we need the confident “true” experimental data for KMs. There are three kinds of non-random motif manipulations: max 1-bp decrease, max 1-bp increase, and least 1-bp change for binding sites of the regulators. These motif manipulations were conducted in the assay because they reduce, improve, and make the smallest change to the PWM match score. The regulators (HNF4 and FOXA for HepG2; GATA and NFE2L2 for K562) included in our study are all activators, so max 1-bp decrease manipulations are supposed to decrease the expression level by disrupting the activator binding; while the max 1-bp increase manipulations are supposed to increase the expression level by strengthen the binding affinity of the activator. As for the least 1-bp change, the change of expression level should be random since the direction of PWM score change caused by these mutations is not clear. In fact, the max 1-bp decrease manipulations do have the largest portion (61%) of decreasing expression level compared to the other two categories (Figure S28A), although a litter smaller than that of KMs at either fKMPs (65.6%) and sKMPs (64.4%). Therefore, the decreasing expression caused by the max 1-bp increase and least 1-bp change can not provide an accurate linkage between motif variants and the expression level change, considering the noise generated during the experiment and noise generated by the unclear designation of motif manipulations in least 1-bp change, as well as the complex mechanism for the correlation between sequence information and downstream expression. For this reason, only the 68 max-1-bp-decrease motif manipulations leading to a decrease on expression levels and the 71 max-1-bp-increase motif manipulations leading to an increase on expression levels were used to test the accuracy of the designation of KMs (at both fKMPs and sKMPs) and RSs. To this end, 139 motif disruptions were included for the accuracy test (Figure S28B). KMs at fKMPs have high precision as 98%, the precision of KMs at sKMP is even better (100%); when evaluating KMs as a whole, its precision is 98.4%, and the overall sensitivity is 88% (among all these KMs, 76% are KMs at fKMP, 12% are KMs at sKMP). RSs have very high precision (100%) but low sensitivity (23%) to predict increasing expression caused by max 1-bp increase manipulation.

KMPs and RSPs

The correlation between the functional constraints (conservation) of fKMPs and their deleterious effect is largely caused by different proportions of binding sites with high binding significance, instead of false positive rate of predicted binding sites. In other words, the greater deleterious effect in the left bins must be caused by larger enrichment of binding sites with high binding significance to a large extent (Figure S16). To validate this conclusion, considering the top k-mers chiefly represent the binding sites of major liver TFs: HNF4, FOXA, and AP1 (c-Fos and c-Jun) whose binding sites (k-mers) tend to have high binding significance, we first investigated the correlation between the proportion of fKMPs located within a major liver TF peaks and the levels of their deleterious effects. The result shows that fKMPs with greater deleterious effect (left bins) tend to have larger portion of binding sites of major liver TFs (Figure S29A); by contrast, the fKMPs with smaller deleterious effect (right bins) have larger portion of binding sites of non-major liver TFs (the HepG2 ChIP-seq peaks of the remaining 53 TFs) (Figure S29B).

More importantly, we found that the fKMPs located within ChIP-seq peaks have stronger positive correlation between the functional constraint and deleterious effect (Figure S30 A-C), especially for the fKMPs located inside the major liver TF binding sites (Figure S30A). A weak positive correlation between functional constraint and deleterious effect is also observed for the fKMPs located outside ChIP-seq peaks, which might be caused by the false positive rate of binding sites in each bin to some extent. However, there could also be some false negative binding sites in each bin, which might be missed by those ChIP-seq peaks. To summarize, the positive correlation between conservation level and deleterious effect of fKMPs is largely caused by the varying levels of binding significance of binding sites in each bin.

Interestingly, ~43% of RSPs locate within the ChIP-seq peaks of major liver TFs, indicating that these RSPs could be potential binding sites without canonical motifs. In fact, similar proportion (~49%) of KMPs locate within ChIP-seq peaks of major liver TFs. Again, we can not determine the false positive of KMPs/RSPs simply based on their relative positions to ChIP-seq peaks. According to the result of MPRA, for the RSPs both inside and outside ChIP-seq peaks, the restoring mutations could increase the enhancer activity; by contrast, the KMs both inside and outside ChIP-seq peaks could decrease the enhancer activity: 54% of KMs that have deactivating effects are located within sites, wherease 65% of RSs that have increasing effects are located within sites.

Enrichment of fKMPs in the regulatory domain of tissue-specific genes

To further study the underlying mechanism of the impact of fKMPs on tissue–specific (TS) gene expressions, we examined if fKMPs are located near the TS genes. We first selected the 200 most highly expressed genes in each of the 79 tissues of the Human U133A/GNF1H Gene Atlas (Su et al. 2004) as the potential TS genes in each tissue, after filtering the 3,804 housekeeping genes out (Eisenberg and Levanon 2013). We next investigated whether the fKMPs and fRSPs are likely to reside in the regulatory domains (RDs) (the upstream and downstream intergenic regions and the intronic regions of the gene) (McLean et al. 2010) of each of the 200 TS genes. The RDs were assigned to genes based on the single-nearest-gene rule (McLean et al. 2010). The HepG2 fKMPs are most greatly enriched in the RDs of the top 200 TS genes in liver compared to in other tissues, so are the HepG2 fRSPs and random enhancer positions, though to a less extent (Figure S31). In addition, the HepG2 fKMPs, fRSPs and HepG2 random enhancer positions are also enriched in the RDs of top 200 TS genes in several other tissues with similar or related functions of liver, such as colorectal adenocarcinoma, fetal liver and lung (fold enrichment >2.6, binomial P-value < 2.23e - 308) (Figure S31). Likewise, the LCL fKMPs are enriched in the RDs of the top 200 TS genes in the tissues BDCA4+ dendritic cells, CD19+ B cells, Lymphoma Burkitts, CD56+ NK cells and 721 B lymphoblastoid cells (fold enrichment > 3.2, binomial P-value < 2.23e - 308), so are the LCL fRSPs and LCL random enhancer positions (fold enrichment > 2.9, binomial P-value < 2.23e - 308). The reason that the LCL fKMPs has a little higher extent of enrichment in the RDs of the TS genes in BDCA4+ dendritic cells, CD19+ B cells, Lymphoma Burkitts, CD56+ NK cells than in 721 B lymphoblastoid cells might be largely caused by the functional similarity across these cells which are all essential/tumor tissues associated with the immune system and therefore utilize similar cohorts of genes (Figure S32). In addition, in both the HepG2 cell line and LCL, the fKMPs are also enriched in the RDs of the 3,804 housekeeping genes, although to a less extent compared to the top TS genes in the corresponding tissue (fold enrichment > 1.7, binomial P-value < 2.23e - 308). However, the enrichment of fKMPs in RDs of housekeeping genes is greater than those of the top TS genes in the majority of other tissues (Figure S32).

To identify what gene categories the fKMPs are likely to regulate, we applied the enrichment analysis of Gene Ontology (GO) information (Ashburner et al. 2000) to the 2,000 most highly expressed non-housekeeping genes in both HepG2 and LCLs, and examined the enrichment of fKMPs in the RDs of gene categories associated with the clustered enriched GO terms (Benjamini–Hochberg corrected P-value < 0.01) (Supplementary Table S8). The fKMPs, fRSPs as well as the random enhancer positions are all likely to be located in the vicinity of genes with tissue-specific functions: the HepG2 fKMPs are mainly enriched in the RDs of genes involved in the lipid metabolic process, blood coagulation, and immune response; in contrast, the LCL fKMPs are principally enriched in the RDs of the genes engaged in immune system development (Figure S33). Although fKMPs, fRSPs also locate near the genes with basic cellular functions such as mitosis, and translation, the enrichment is not as strong as that of random enhancer regions, further indicating the tissue-specific regulation role of fKMPs (Figure S33).

By combining the information of the TFBSs overlapping the HepG2 fKMPs and fKMP-associated genes of certain functional groups, we rebuilt the map linking the major TFs in liver (Figure S34) to their involved biological pathways. We investigated whether the TFBSs overlapping fKMP clusters are likely to reside in the RDs of genes within a certain GO term. The binomial approach implemented in GREAT (McLean et al. 2010) was applied here to analyze the enrichment of RDs of genes associated with a certain GO terms in TFBSs (those overlapping fKMPs clusters) relative to random enhancer regions. For each enhancer, given the fKMPs-associated binding sites of a TF, we randomly selected the same amount of segments with the same lengths as the TFBSs in that enhancer, and repeated the process for 100 times. We then took the average of enrichment scores across 100 iterations to estimate the gene enrichment in random enhancer regions. Finally, the enrichment of RD of the genes associated with a GO term in the TFBSs overlapping fKMPs clusters relative to random enhancer regions was estimated by formula (6) - (9):

$$Enrichment_{[TFBS\cap RD(GO)]/[Enhancer\cap RD(GO)}$$

$$= Enrichment_{TFBS\cap RD(GO)} / AverageEnrichment_{Enhancer\cap RD(GO)} \qquad (6)$$

$$Enrichment_{TFBS\cap RD(GO)} = P_{TFBS} / P_{GO} \qquad (7)$$

$$P_{TFBS} = \frac{|TFBS \cap RD(GO)|}{|TFBS|} \qquad (8)$$

$$P_{GO} = \frac{sum(length(RD(GO)))}{sum(length(RD(allGenes))} \qquad (9)$$

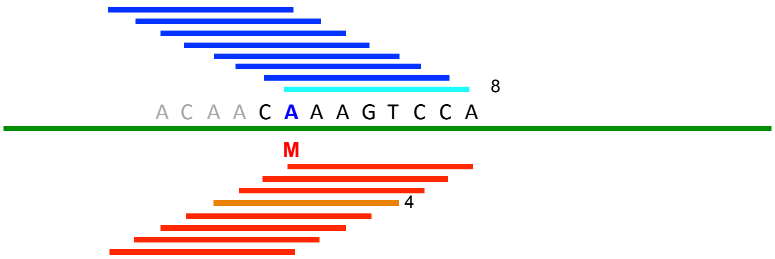
Where $|TFBS|$ represents the total number of binding sites of a TF, and $|TFBS \cap RD(GO)|$ represents the number of binding sites of the TF residing in the RD of genes with the particular GO term.

We speculated that clusters of TFs would perturb key biological pathways once KMs disrupt their binding sites: HNF4α, FOXA1, PPARγ, TFAP2, RXRA are mainly in charge of the lipid metabolic system, such as triglyceride-rich lipoprotein particle remodeling, HDL and LDL assembly/remodeling/clearance, cholesterol esterification; LEF1, TBP, ZFX, SP1, IRF1, SPI1, PPARγ, RAD21, ESR1, FOXO1 are mainly in charge of the immune response and B-cell/lymphocyte/leukocyte mediated immunity (Figure S34). This map might minimize the scope of suspects for diagnosis of certain disturbed regulatory pathways: mutations that happened at the fKMPs of these binding sites are likely to be the potential candidates associated with the disruption of the corresponding biological pathways.

Supplementary References

Ashburner M, Ball CA, Blake JA, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
Eisenberg E and Levanon EY. 2013. Human housekeeping genes, revisited. Trends Genet 29: 569-574.
Mahony S and Benos PV. 2007. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res 35: W253-258.
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM and Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28: 495-501.
Su AI, Wiltshire T, Batalov S, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062-6067.
Visel A, Blow MJ, Li Z, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457: 854-858.

Supplementary Figure 1

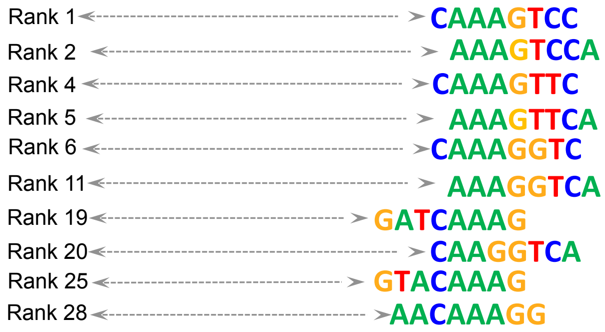


M (Mutation) : C, T or G

$$\Delta Sig = -\log_{10} P_1 - (-\log_{10} P_2)$$

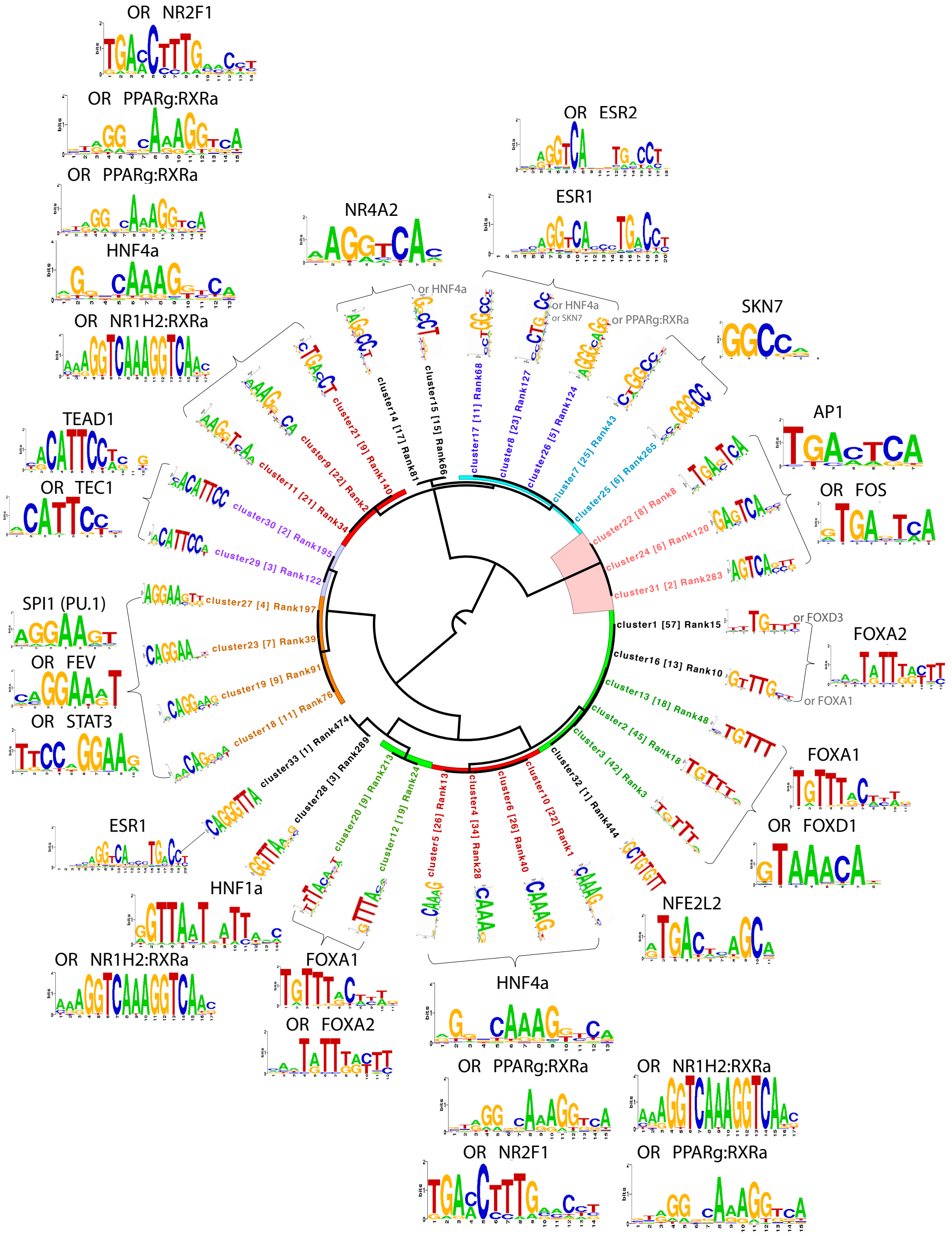
Modified IGR model. The blue bars are the k-mers associated with the wild-type nucleotide A (in blue color), the 8th k-mer in light blue is the k-mer with the highest binding significance (-$\log_{10}P_1$). The red bars are the k-mers associated with the mutated nucleotide M (C, T, or G), the 4th k-mer in orange is the k-mer with the highest binding significance (-$\log_{10}P_2$) associated with M. The change of the binding significance ΔSig is used to quantify the deleterious effect on the binding fitness caused by mutation M. If the mutation M changes the significant k-mer (light blue #8) to an insignificant one (orange #4), this position would be identified as KMP; M would be identified as a candidate killer mutation (KM). If all three mutations are KMs, this position is defined as fragile KMP (fKMP), otherwise as stable KMP (sKMP).

Supplementary Figure 2



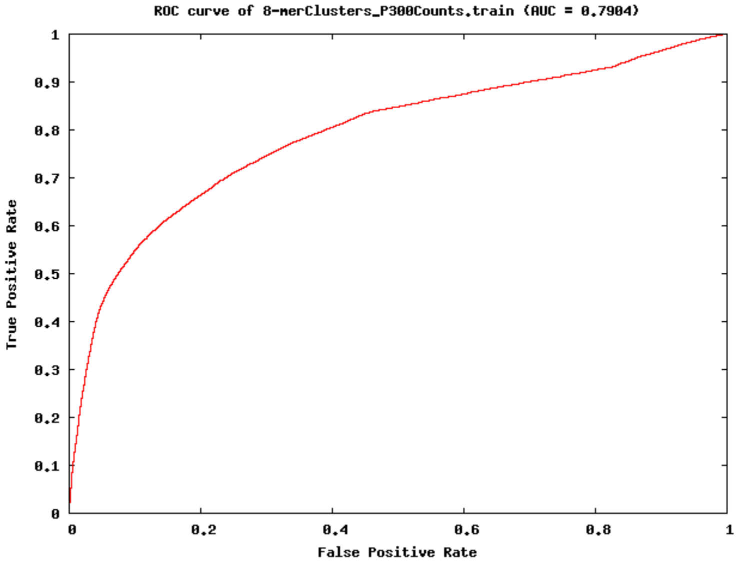
Top K-mers are redundant and tend to overlap with each other.

Supplementary Figure 3



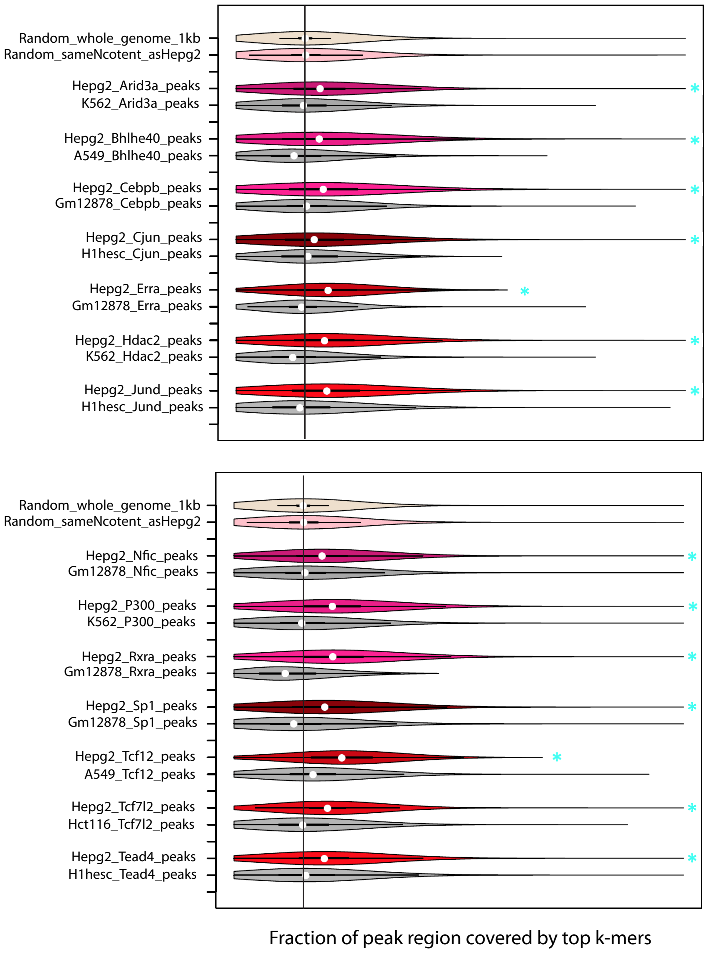
Clusters of 522 top k-mers mapping to known TFBS. Thirty-three k-mer clusters were aligned and merged into 14 motif clusters. STAMP (platform for Similarity, Tree-building, and Alignment of DNA Motifs and profiles) (Mahony and Benos 2007) identified 22 known TFBS in these clusters, from which 14 are liver-specific TFs. The inner-circle logos are the motifs for each k-mer cluster, the outer-circle logos correspond to a known motifs from TRANSFAC and JASPAR matching motif clusters. Beside the common known TFBSs for all k-mer clusters within a motif cluster, alternative known TFBSs for a sub-k-mer cluster were labeled on its the side in grey. The number within the parentheses indicates the number of k-mers in each k-mer cluster.

Supplementary Figure 4



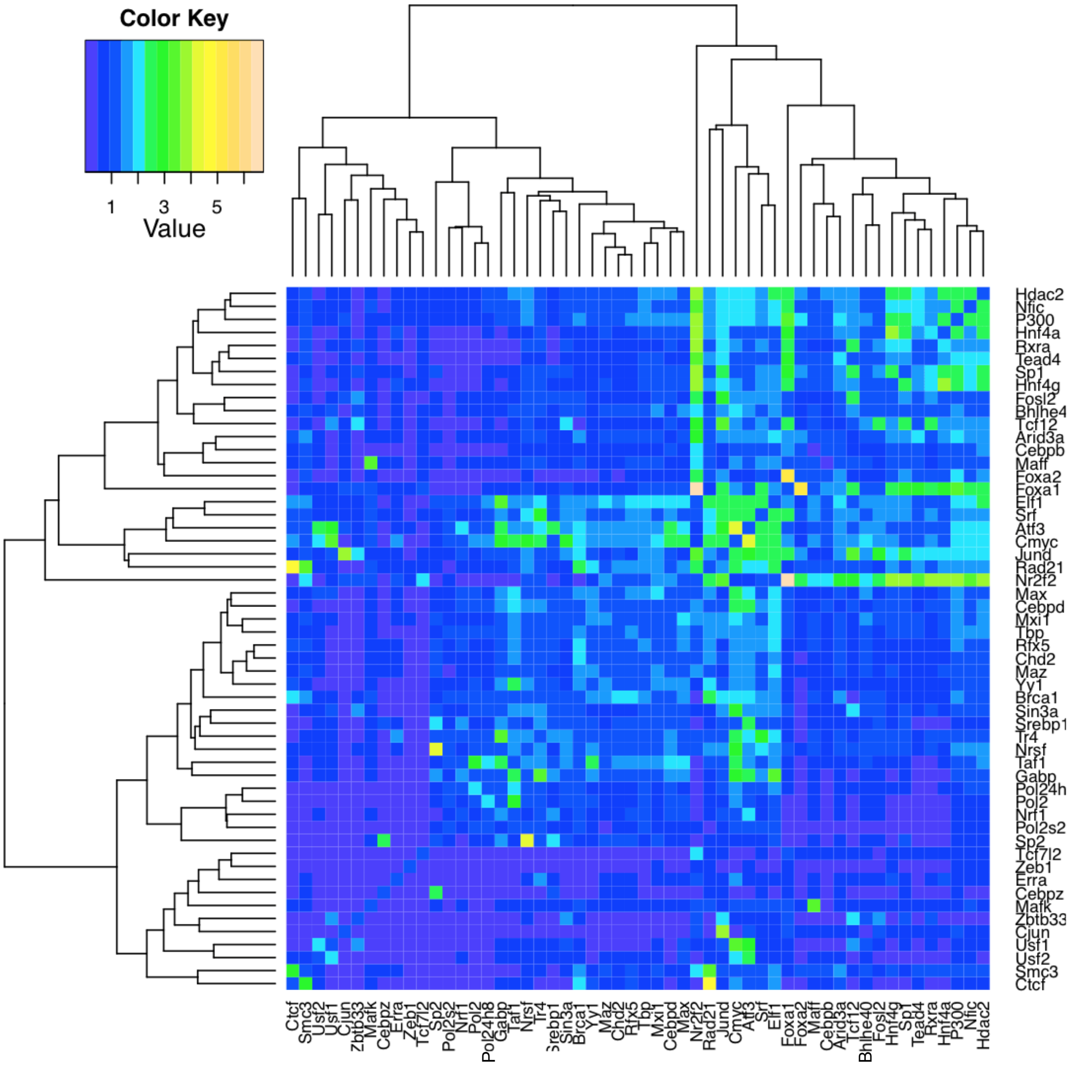
Fourteen k-mer clusters of the top 522 k-mers are sufficient to distinguish P300 enhancers (outside ChromHMM HepG2 enhancers) from random sequences.

Supplementary Figure 5



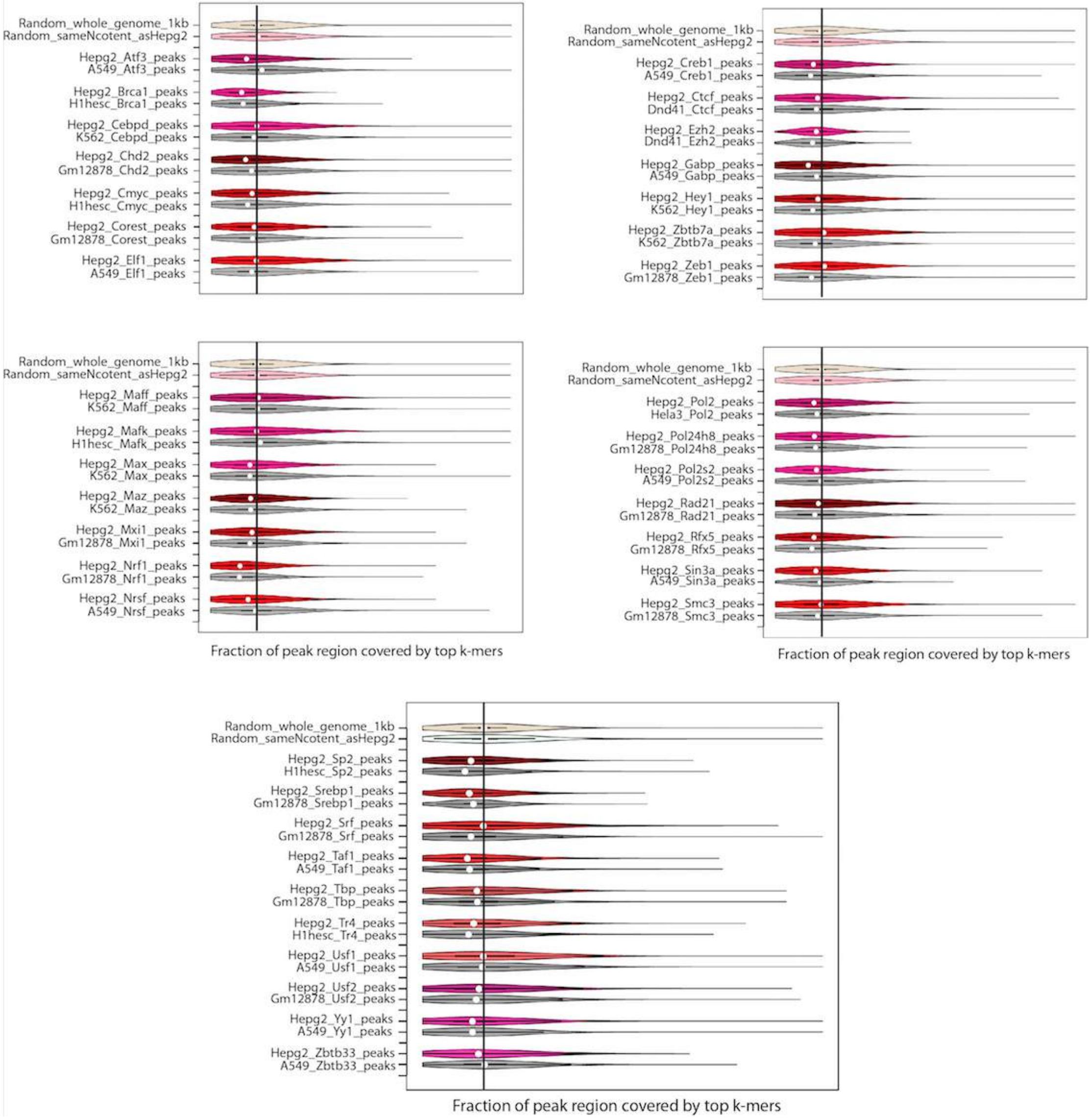
Fourteen TFs with top k-mers enriched in their ChIP-seq peaks only in HepG2 cell line. The significant enrichment of top k-mers in the peak regions compared to the random background is highlighted by cyan asterisks (Mann-Whitney test p-value < 0.001).

Supplementary Figure 6



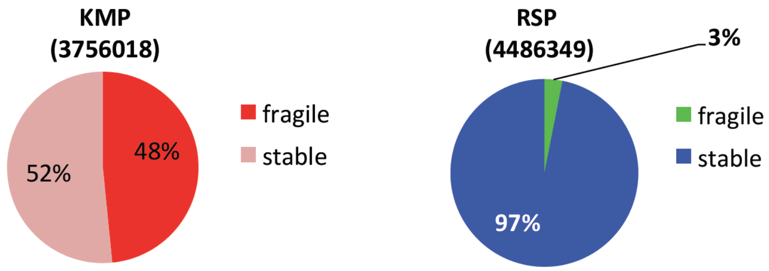
A heatmap showing fold enrichment of overlap between TF ChIP-seq peaks relative to expectation. Neighboring peaks with less than 50 bp between the two peak centers were considered as overlapping peaks. The null expectations were constructed via randomly generated 1000 independent pairs of TF ChIP-seq peaks, each with similar size as the tested TFs.

Supplementary Figure 7



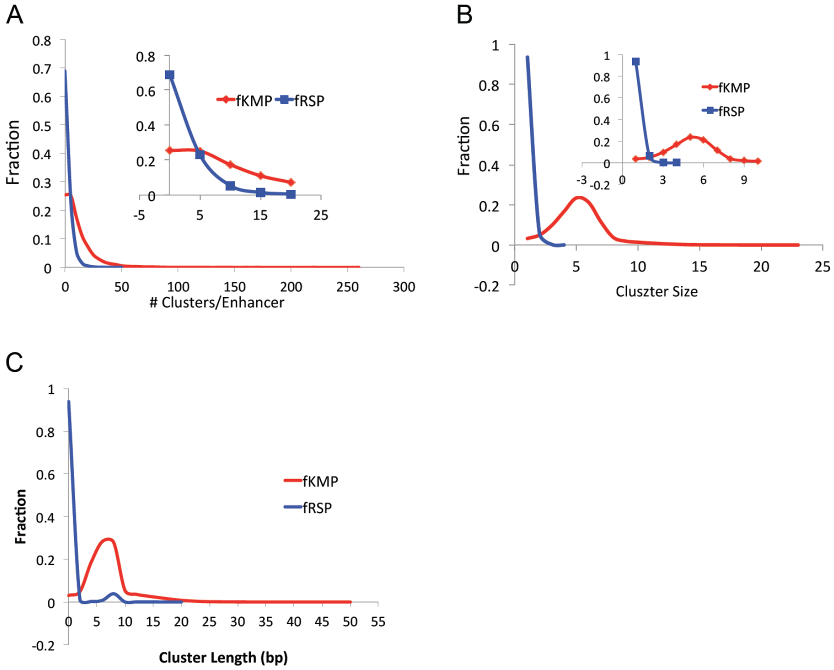
The top k-mers were not enriched in the ChIP-seq peak regions of 38 TFs in HepG2 cell line.

Supplementary Figure 10



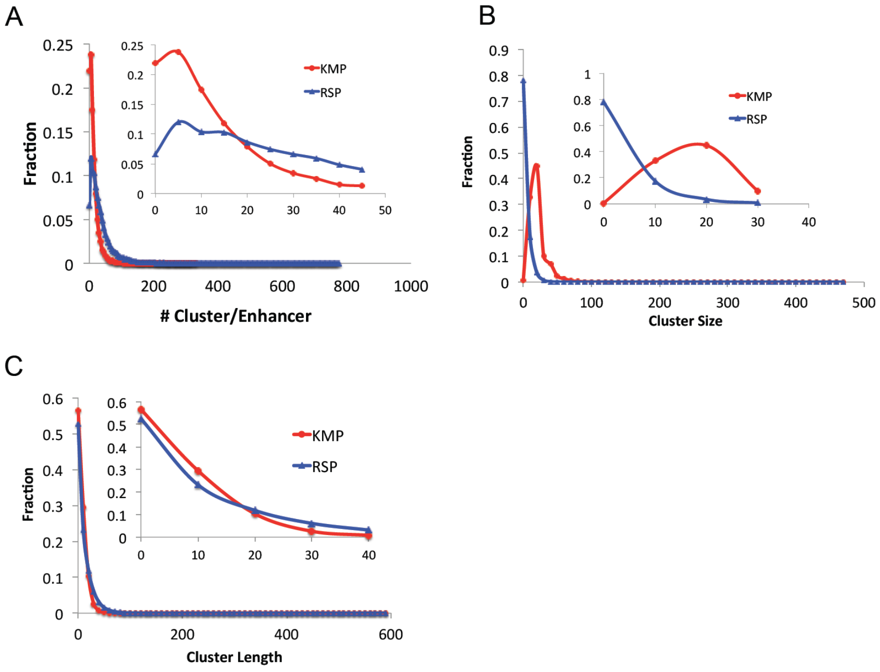
Proportions of fKMP and fRSPs.

Supplementary Figure 11



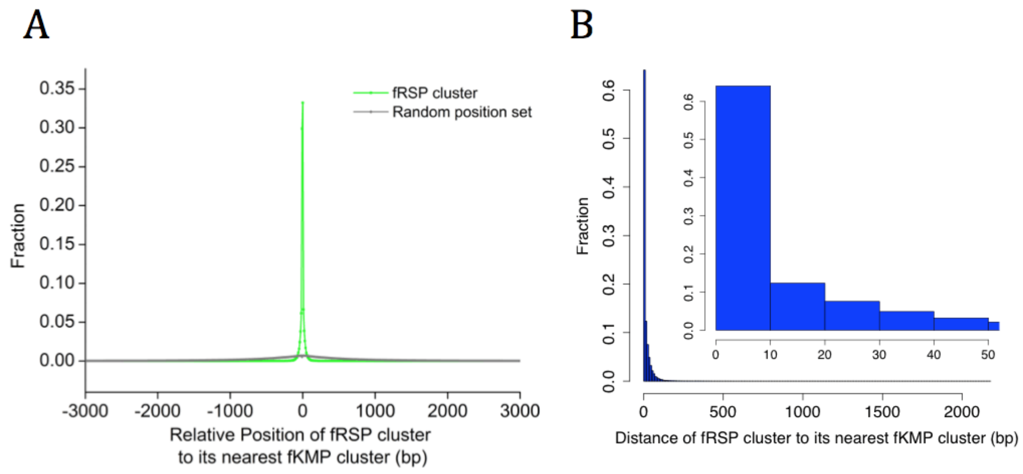
Properties of fKMP/fRSP clusters. A) Distribution of # fKMP/fRSP clusters per enhancer. B) Distribution of # fKMP/fRSP positions in each cluster. C) Distribution of fKMP/fRSP cluster lengths.

Supplementary Figure 12



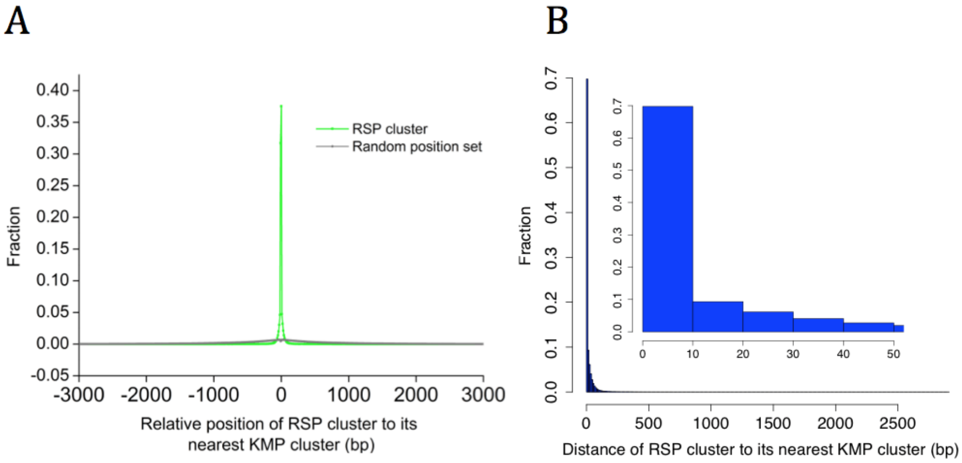
Properties of KMP/RSP clusters. A) Distribution of # KMP/RSP clusters per enhancer. B) Distribution of # KMP/RSP positions in each cluster. C) Distribution of KMP/RSP cluster lengths.

Supplementary Figure 13



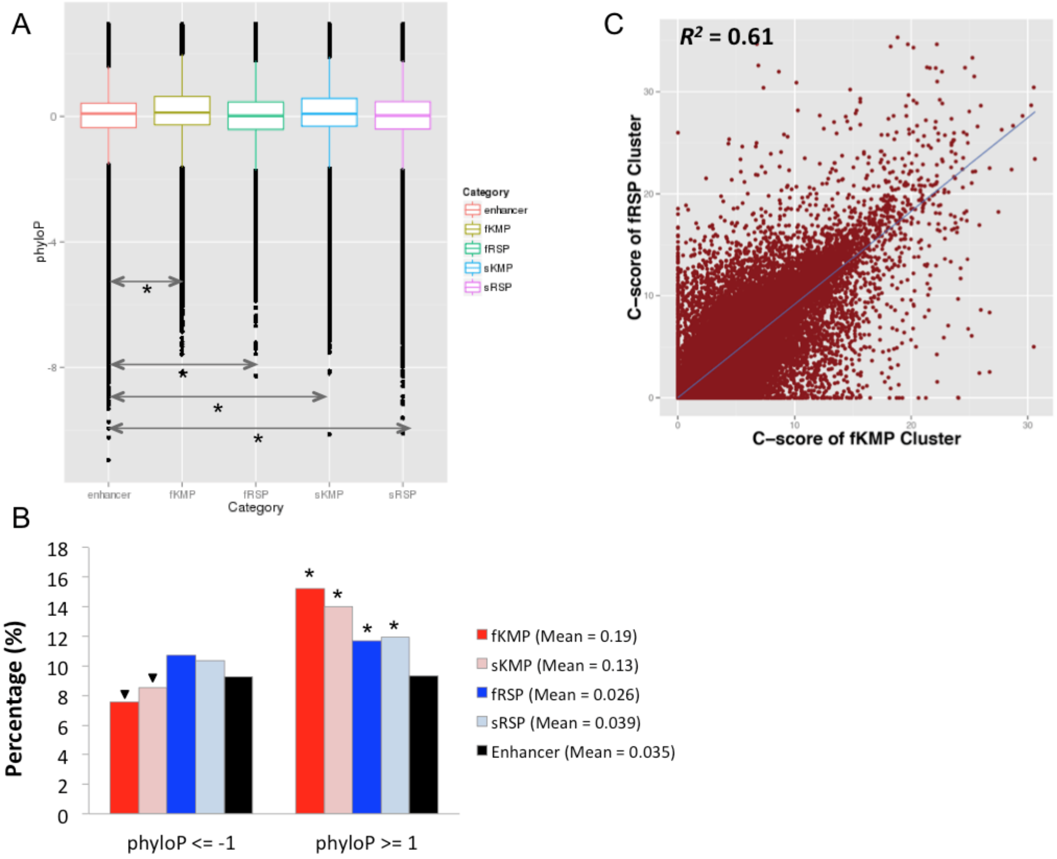
fRSP clusters locate bias to the neighborhood of the fKMP clusters. A) Relative positions of fRSP cluster to its nearest fKMP cluster compared to those of the empirical random background. B) Distribution of distance of fRSP cluster to its nearest fKMP cluster against those of the empirical background. Around 60% of fRSP clusters locate within 10-bp neighborhood of their nearest fKMP clusters.

Supplementary Figure 14



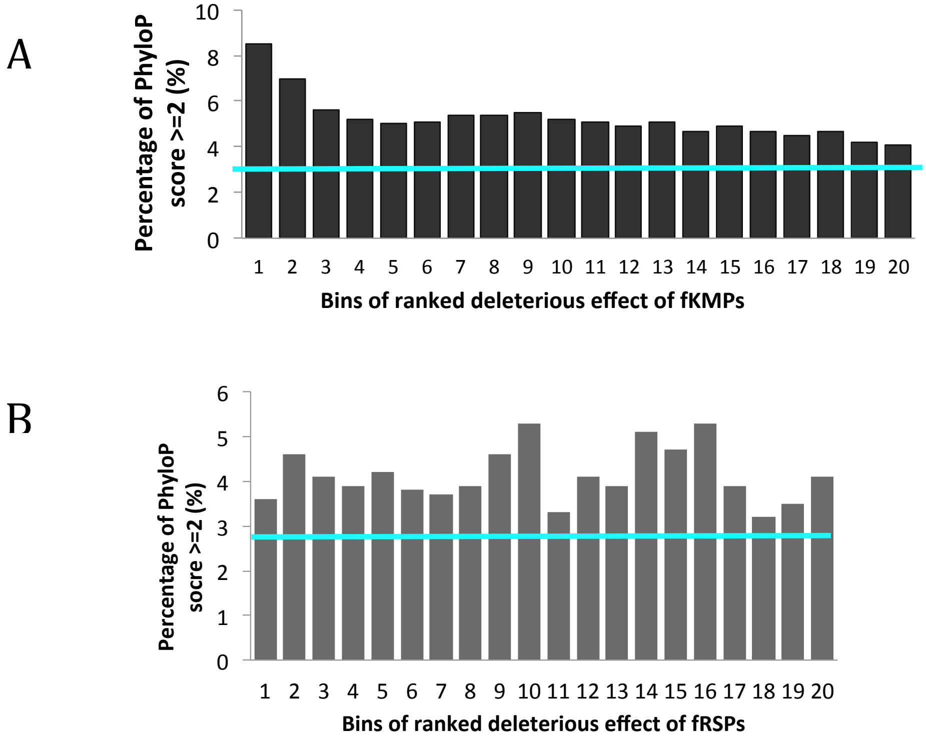
RSP clusters locate bias to the neighborhood of the KMP clusters. A) Relative positions of RSP cluster to its nearest KMP cluster compared to those of the empirical random background. B) Distribution of distance of RSP cluster to its nearest KMP cluster against those of the empirical background. Around 70% of RSP clusters locate within 10-bp neighborhood of their nearest KMP clusters.

Supplementary Figure 15



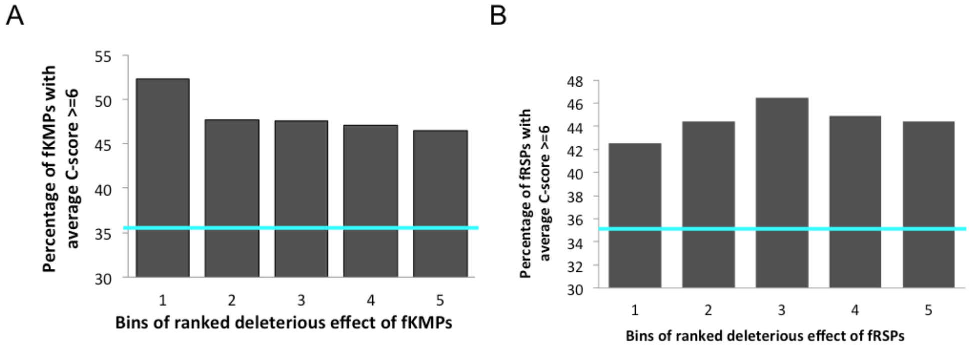
KMPs have higher PhyloP score than RSPs and enhancer position. A) Box plot of phyloP score of KMPs, RSPs, and regular enhancer positions, respectively. KMPs are the most conserved category, followed by RSPs. The P-values are calculated using Mann-Whitney test. **B)** KMPs and RSPs are both enriched in the conserved category (PhyloP ≥ 1), but only KMPs are depleted in the non-conserved category (PhyloP ≤ -1). The mean values in the legend are the average PhyloP score for the corresponding category. Both the P-values for enrichment and depletion were calculated using fisher's exact test. Asterisk represents significant enrichment compared to control (Enhancer), P < 2.2e-16. Triangle represents significant depletion compared to control (Enhancer), P < 2.2e-16. **C)** The average C-score of fKMP clusters is positively correlated with the average C-score of the nearest fRSP clusters. R2 = 0.61 (PCC = 0.78).

Supplementary Figure 16



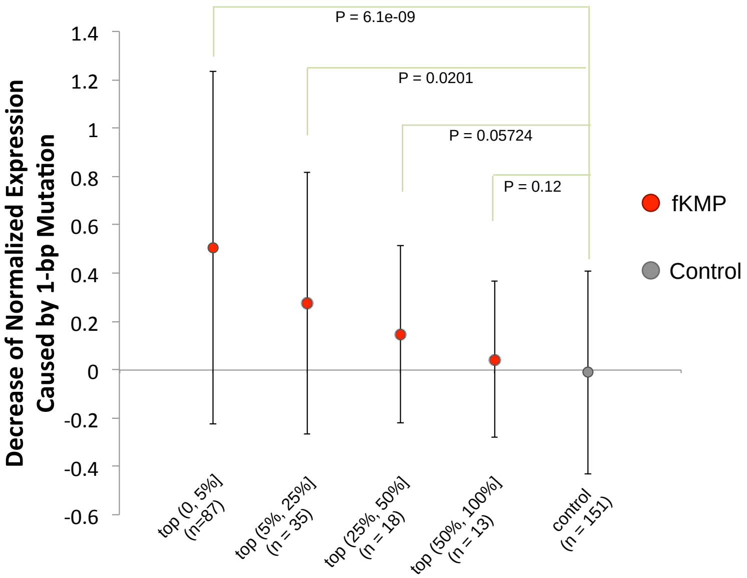
Positive correlation is observed between minimum binding significance drop of fKMPs and their conservation levels but no apparent correlation is observed between deleterious effect increase of fRSPs and their conservation levels. A) fKMPs are sorted in a decreasing manner by the minimum significance drop caused by KMs, and separated into 20 bins. **B) f**RSPs are sorted by the absolute value of the minimum significance increase caused by RSs in a decreasing manner, and separated into 20 bins. The cyan line high lights the percentage (2.84%) of positions with PhyloP score ≥ 2 in regular enhancer region.

Supplementary Figure 17



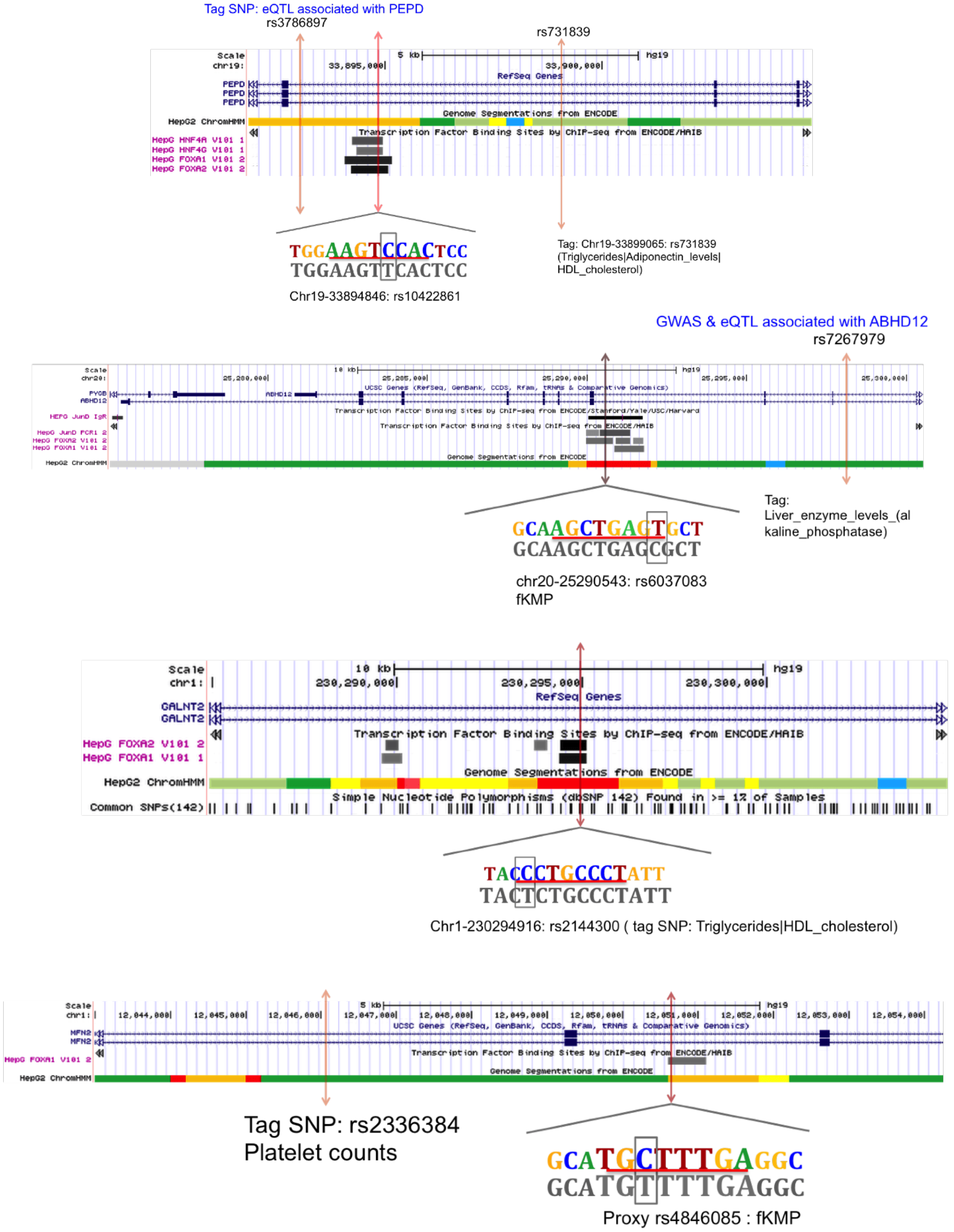
Correlations between deleterious effect of fKMPs/fRSPs and their average C-scores. Both **f**KMP (A) and fRSP (B) are sorted by the deleterious effect of the positions in a decreasing manner, and separated into five bins. The cyan line highlights the percentage (35.12%) of positions with C-score ≥ 6 in regular enhancer region.

Supplementary Figure 18



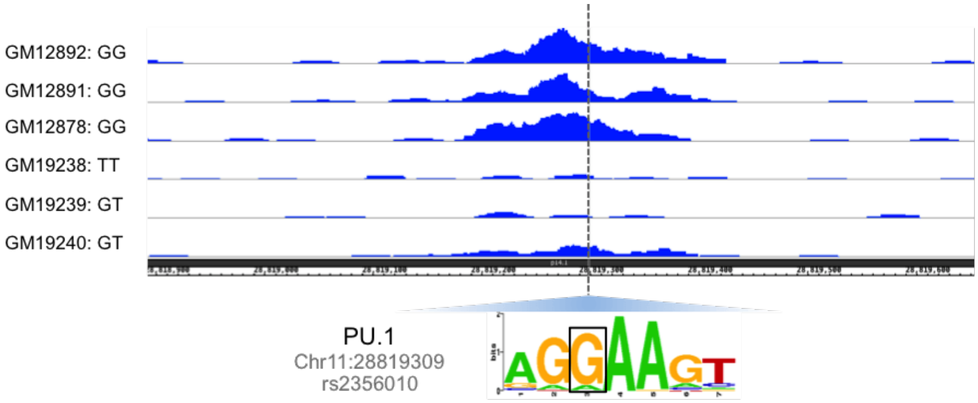
Stratification of expression decrease caused by mutations at fKMPs sorted decreasingly by their deleterious effects. The sorted fKMPs were separated to four bins. The numbers in the parentheses under each percentile interval indicate the size of mutations in the corresponding bin. P-values were obtained by applying Mann-Whitney test**.** Each error bar indicates the mean and the standard deviation of the normalized expression level drop in each bin.

Supplementary Figure 19


Tag SNP: eQTL associated with PEPD
rs3786897
rs731839
Scale
chr19:
5 kb
hg19
33,895,000
33,900,000
RefSeq Genes
PEPD
PEPD
PEPD
Genome Segmentations from ENCODE
HepG2 ChromHMM
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
HepG HNF4A V101 1
HepG HNF4G V101 1
HepG FOXA1 V101 2
HepG FOXA2 V101 2
TGGAAGTCCACTCC
TGGAAGTTCACTCC
Chr19-33894846: rs10422861
Tag: Chr19-33899065: rs731839
(Triglycerides|Adiponectin_levels|
HDL_cholesterol)
GWAS & eQTL associated with ABHD12
rs7267979
Scale
chr20:
10 kb
hg19
25,280,000
25,285,000
25,290,000
25,295,000
25,300,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PYGB
ABHD12
ABHD12
Transcription Factor Binding Sites by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard
HEPG JunD IgR
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
HepG JunD PCR1 2
HepG FOXA2 V101 2
HepG FOXA1 V101 2
Genome Segmentations from ENCODE
HepG2 ChromHMM
GCAAGCTGAGTGCT
GCAAGCTGAGCGCT
chr20-25290543: rs6037083
fKMP
Tag:
Liver_enzyme_levels_(al
kaline_phosphatase)
Scale
chr1:
10 kb
hg19
230,290,000
230,295,000
230,300,000
RefSeq Genes
GALNT2
GALNT2
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
HepG FOXA2 V101 2
HepG FOXA1 V101 1
Genome Segmentations from ENCODE
HepG2 ChromHMM
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
TACCCTGCCCTATT
TACTCTGCCCTATT
Chr1-230294916: rs2144300 (tag SNP: Triglycerides|HDL_cholesterol)
Scale
chr1:
5 kb
hg19
12,044,000
12,045,000
12,046,000
12,047,000
12,048,000
12,049,000
12,050,000
12,051,000
12,052,000
12,053,000
12,054,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
MFN2
MFN2
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
HepG FOXA1 V101 2
Genome Segmentations from ENCODE
HepG2 ChromHMM
Tag SNP: rs2336384
Platelet counts
GCATGCTTTGAGGC
GCATGTTTTGAGGC
Proxy rs4846085 : fKMP

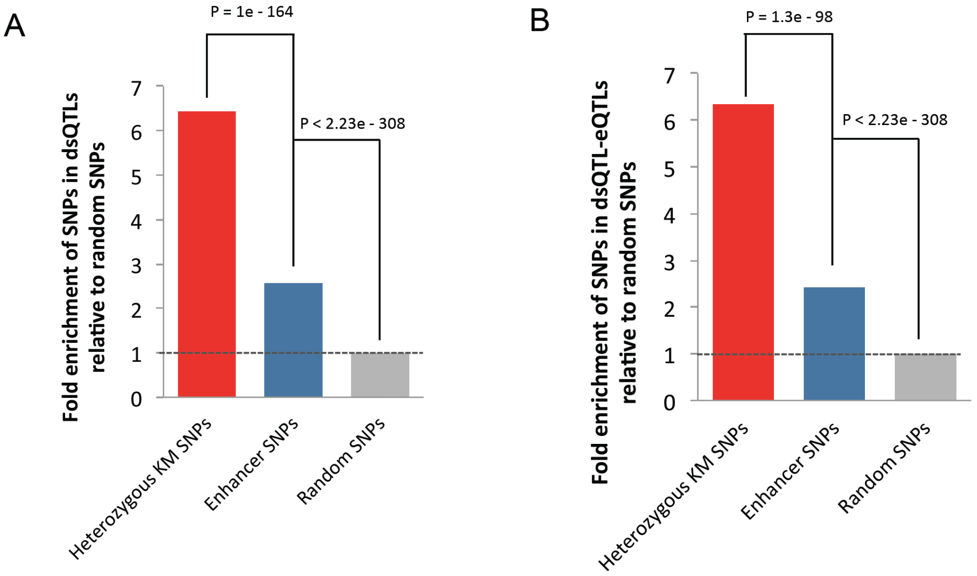
Examples of KM SNPs that are likely to be the causal SNPs of the liver/liver-related traits. The significant k-mers are highlighted by a red underscore line.

Supplementary Figure 20



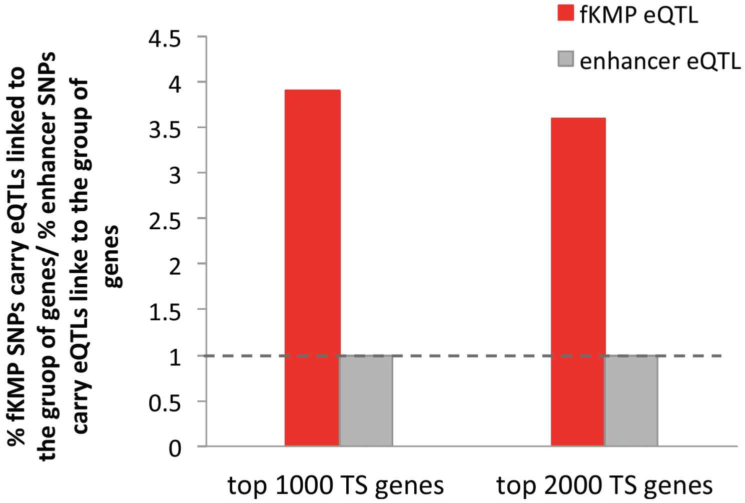
One example showing KM/RS is the causal variant that cause different PU.1 binding. The regions shown here is 1.2kb long.

Supplementary Figure 21



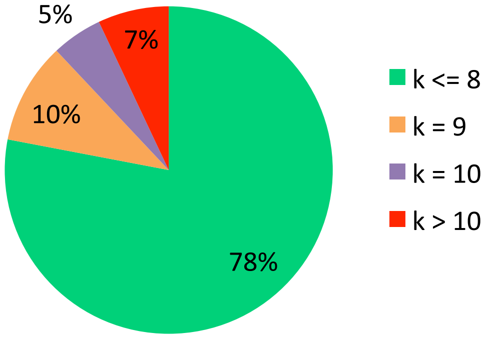
Heterozygous KM SNPs are strongly enriched in dsQTL and dsQTL-eQTL. A) The heterozygous KM SNPs were enriched in dsQTLs relative to LCL enhancer SNPs. The LCL enhancer SNPs were also enriched in dsQTLs relative to the 500 sets of matched random SNPs with the same distance to TSS.**B**) The heterozygous KM SNPs were enriched in dsQTLs-eQTLs relative to LCL enhancer SNPs. The LCL enhancer SNPs were also enriched in dsQTLs-eQTLs relative to the 500 sets of matched random SNPs with the same distance to TSS.

Supplementary Figure 22



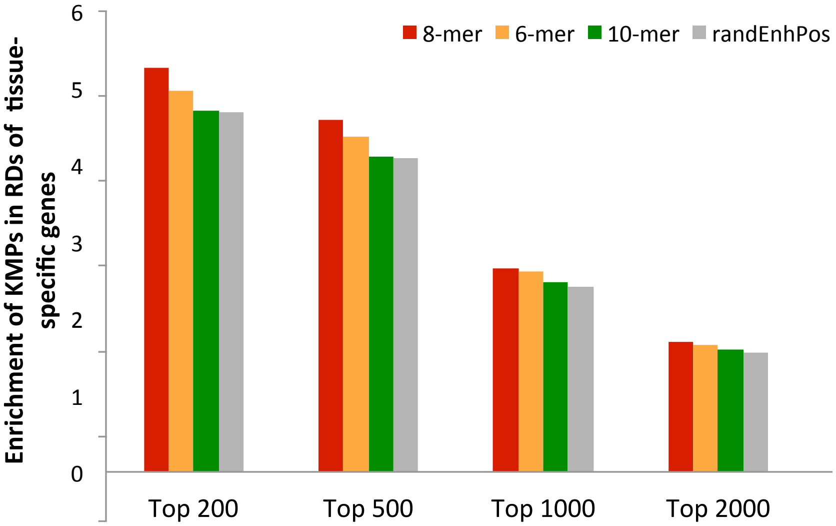
The SNPs at fKMPs are much strongly enriched in eQTLs linked to genes with higher expression level specifically in LCL cell lines. The dashed line indicates the base line where fold enrichment = 1.

Supplementary Figure S23



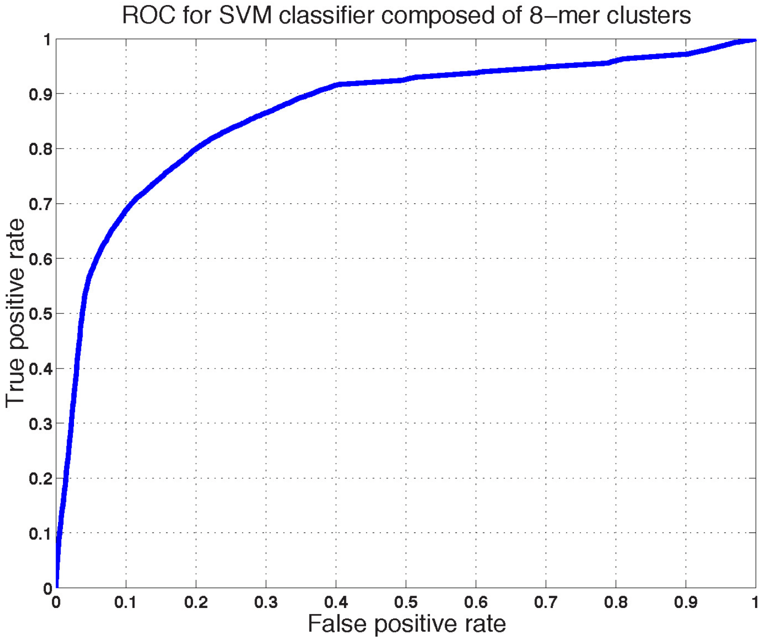
Distribution of length of informative regions of known TF binding motifs in TRANSFAC and JASPAR. To obtain the informative motif region, for a motif, we filtered the marginal region and the internal gap region (consecutive 3+bp) with low information content (<0.8).

Supplementary Figure S24



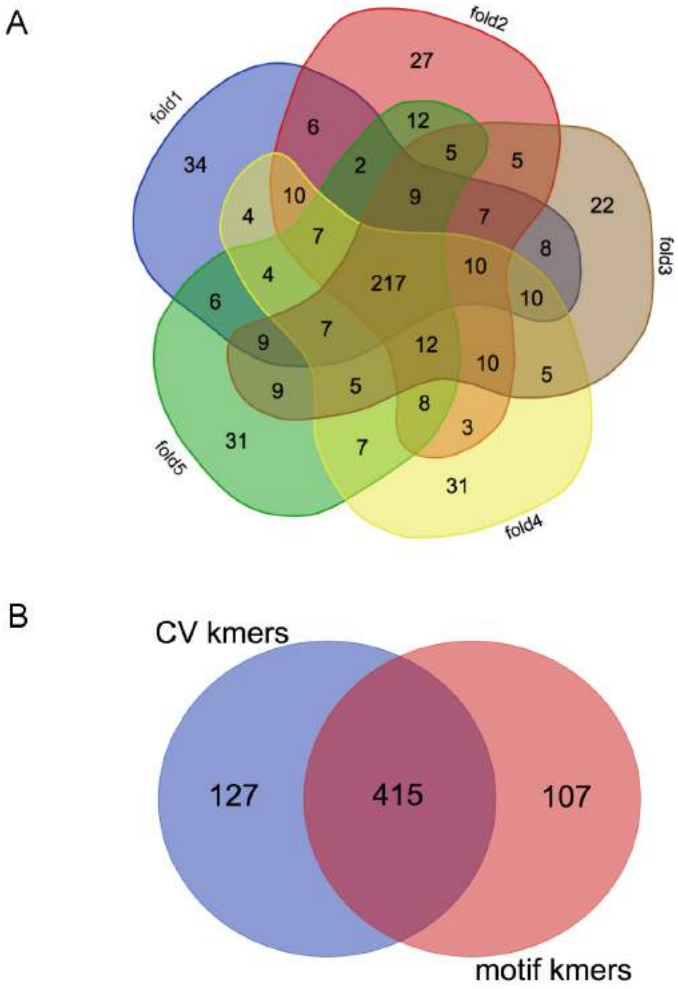
Enrichment of fKMPs in the regulatory domains of the most highly expressed liver-specific genes. The KMPs were identified using 8-mers, 6-mers, and 10-mers, respectively. We selected the 200, 500, 1000, and 2000 most highly expressed genes in liver compared to other tissues of the Human U133A/GNF1H Gene Atlas (Su et al. 2004) as the potential liver-specific genes.

Supplementary Figure 25



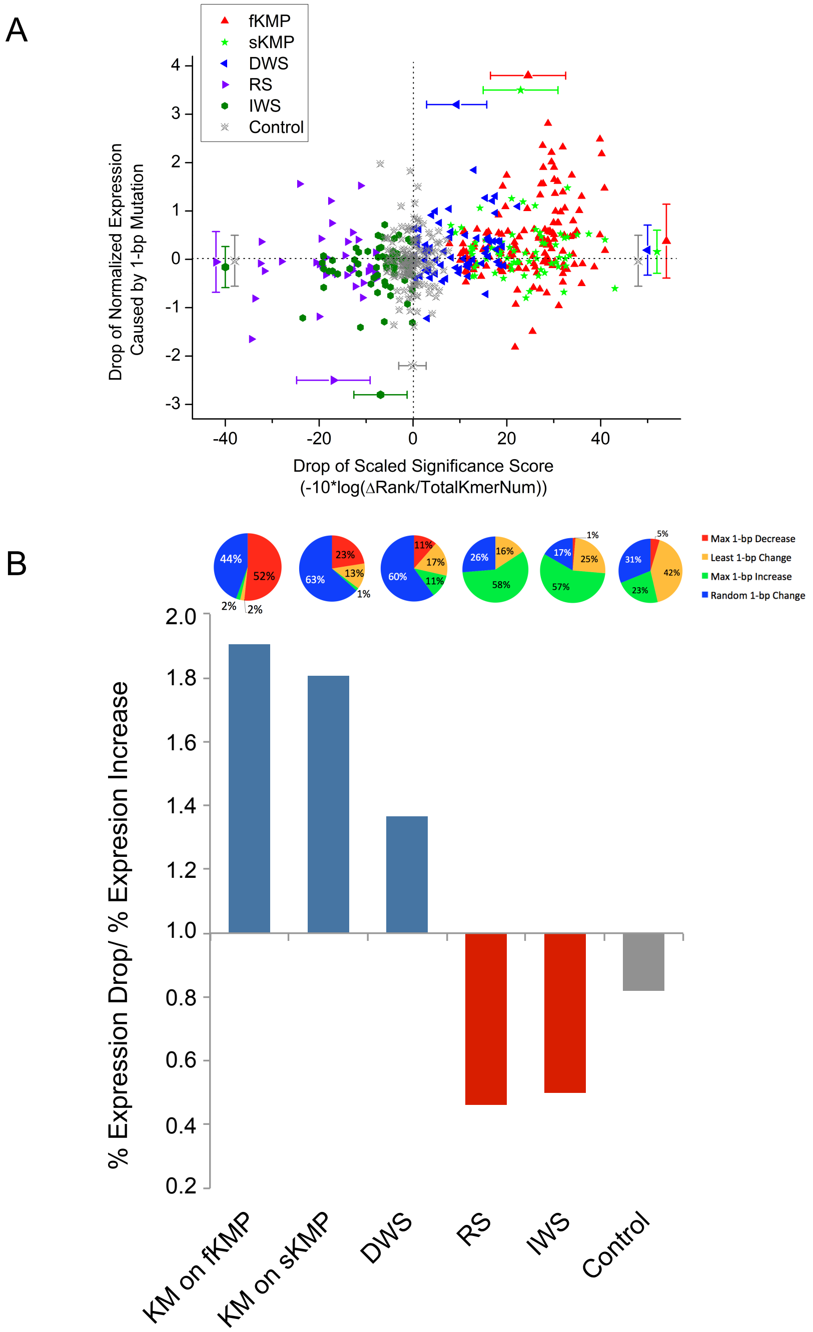
ROC curve for a 5-fold cross validation of independent sets of k-mers. The AUC is 85%.

Supplementary Figure 26



Correlation between the top k-mers obtained from enrichment of 5-fold cross validation and all enhancers. A) Venn diagram showing the overlapping across the top k-mers in each training set of the 5-fold cross validation. There are 542 significant k-mers overall. B) Venn diagram showing the overlapping between the union of 542 top k-mers in 5-fold Cross Validation (CV k-mers) and 522 top k-mers enriched in all HepG2 enhancers (motif k-mers).

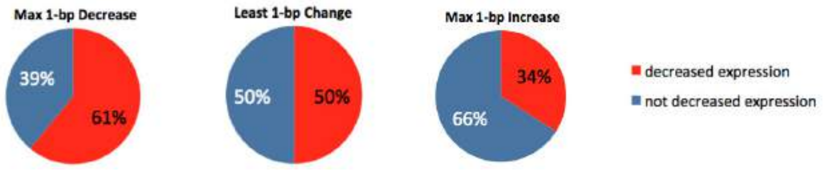
Supplementary Figure 27



Deleterious effect of all genetic variants on enhancer activity. A) Relationship between drop of the scaled significance score and the drop of normalized expression level of the 145-bp enhancer. The Y-axis is the normalized expression level drop caused by the 1-bp mutation, and X-axis is the drop of the scaled significance of the mutated k-mer relative to the original k-mer. Error bars indicate the mean and the standard deviation of the corresponding axis in each category. B) Tendency of decreased expression of each predicted mutation category. The y-axis is the ratio of proportion of mutations that decrease expression to that of mutations increase expression. KM: Killer mutations at both fKMPs and sKMPs. RS: Restoration mutations on both fRSPs and sRSPs. DWS: Drop within significant set mutation. IWS: Increase within significant set mutation. The pie charts above indicate the composition of PWM manipulations in each of our predicted category.

Supplementary Figure 28

A

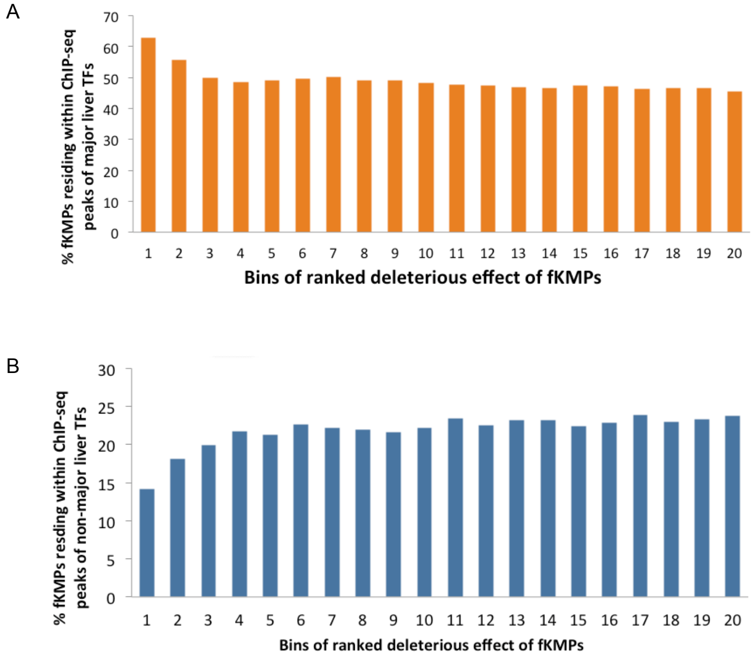


B

	KMs at fKMP	KMs at sKMP	DWS	RS	IWS	control
expression decrease caused by max 1-bp drop (68)	52	8	5	0	0	3
expression Increase caused by max 1-bp increase (71)	1	0	5	16	28	21

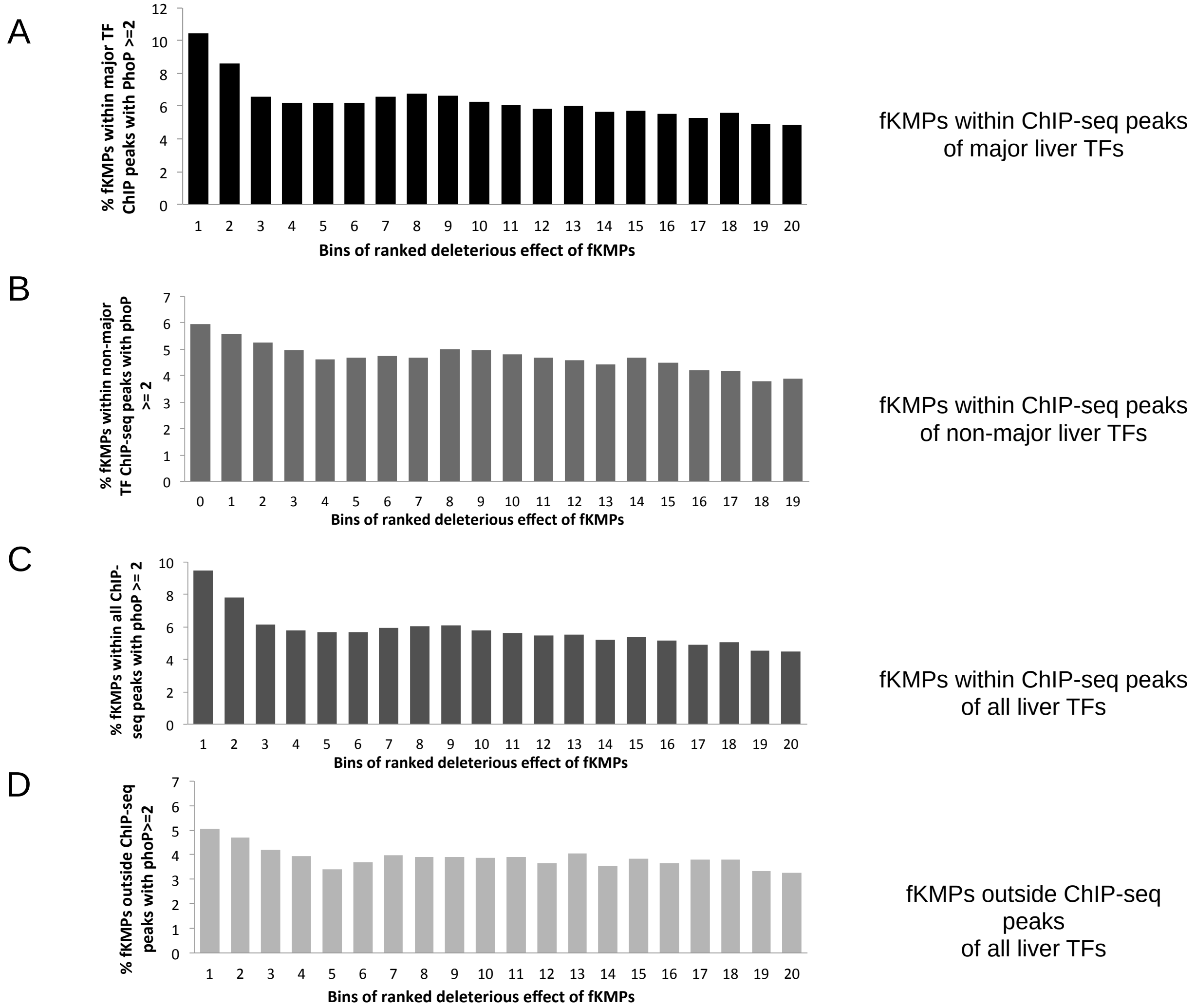
Performance of MPRA and our predictions. Proportions of decreasing enhancer activity in each PWM-score manipulation category, A). Distribution of the true/false expression change in our predictions B).

Supplementary Figure 29



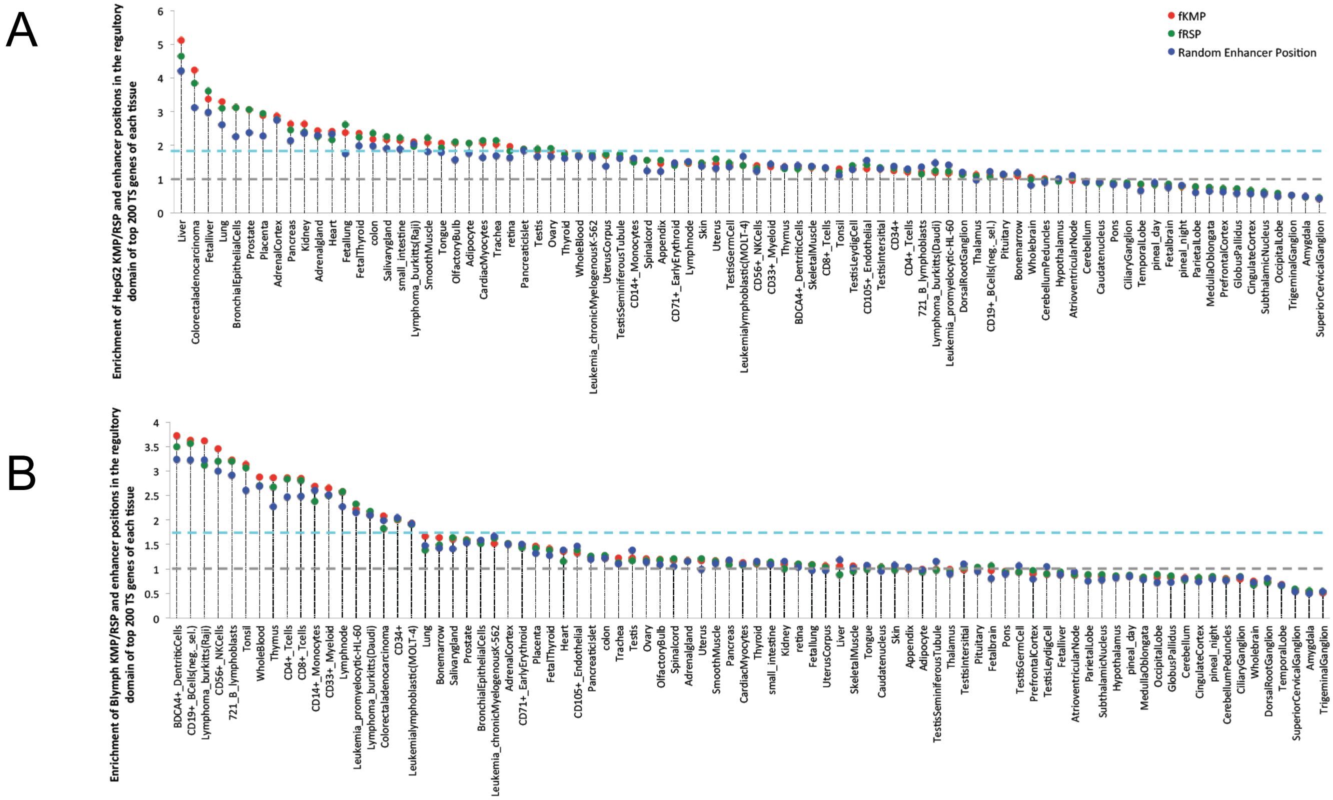
Proportion of fKMPs in each bins located within a ChIP-seq peak of a major liver TF (A), and non-major liver TF (B).

Supplementary Figure 30



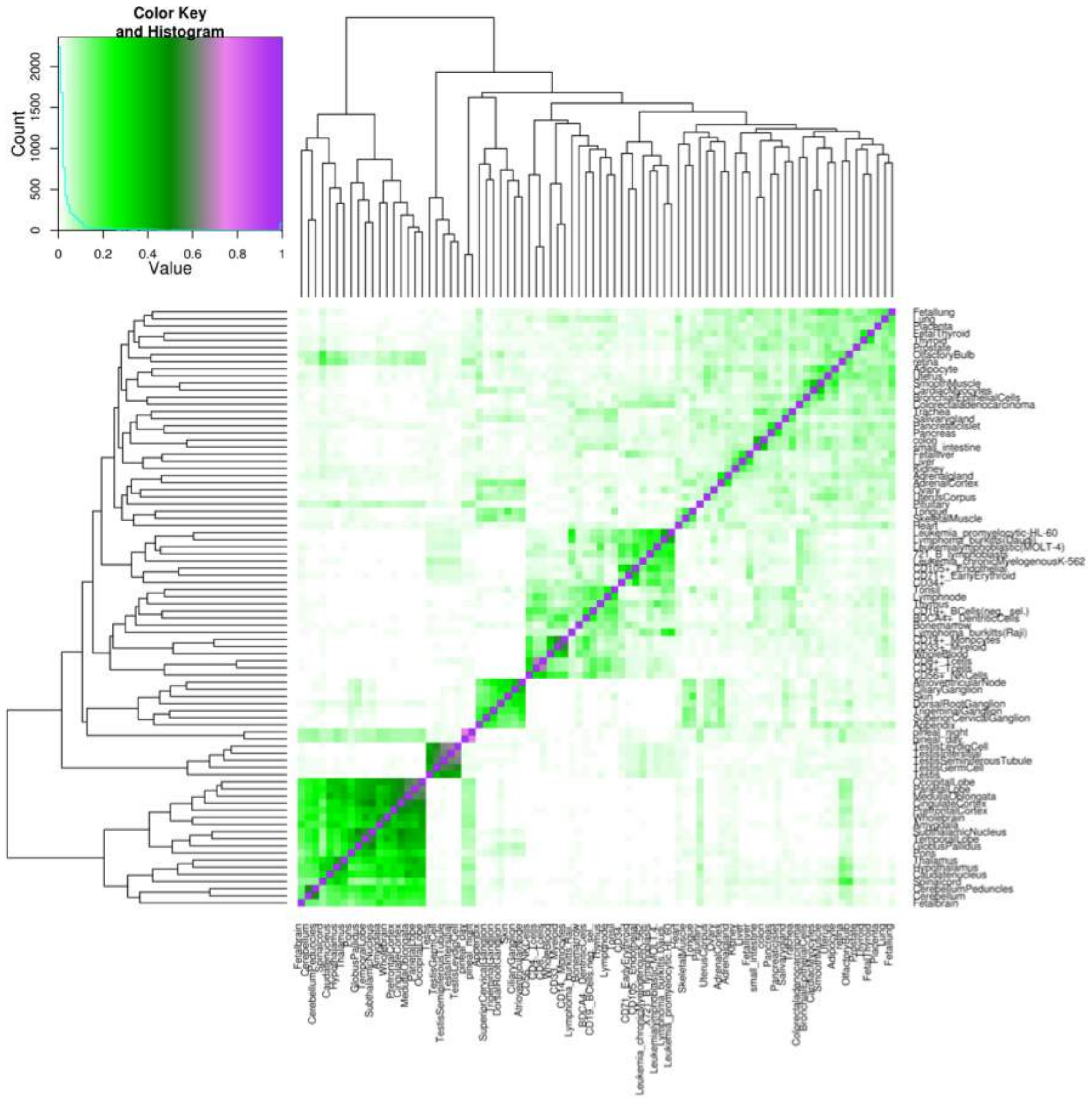
Positive correlation is observed between minimum binding significance drop of fKMPs and their conservation levels for fKMPs located within ChIP-seq peaks of major liver TFs (HNF4, FOXA, AP1), A); fKMPs located within ChIP-seq peaks all other liver TFs, B); fKMPs located within ChIP-seq peaks of all liver TFs, C); fKMPs located outside ChIP-seq peaks of any liver TFs, D). fKMPs are sorted in a decreasing manner by the minimum significance drop caused by KMs, and separated into 20 bins.

Supplementary Figure 31



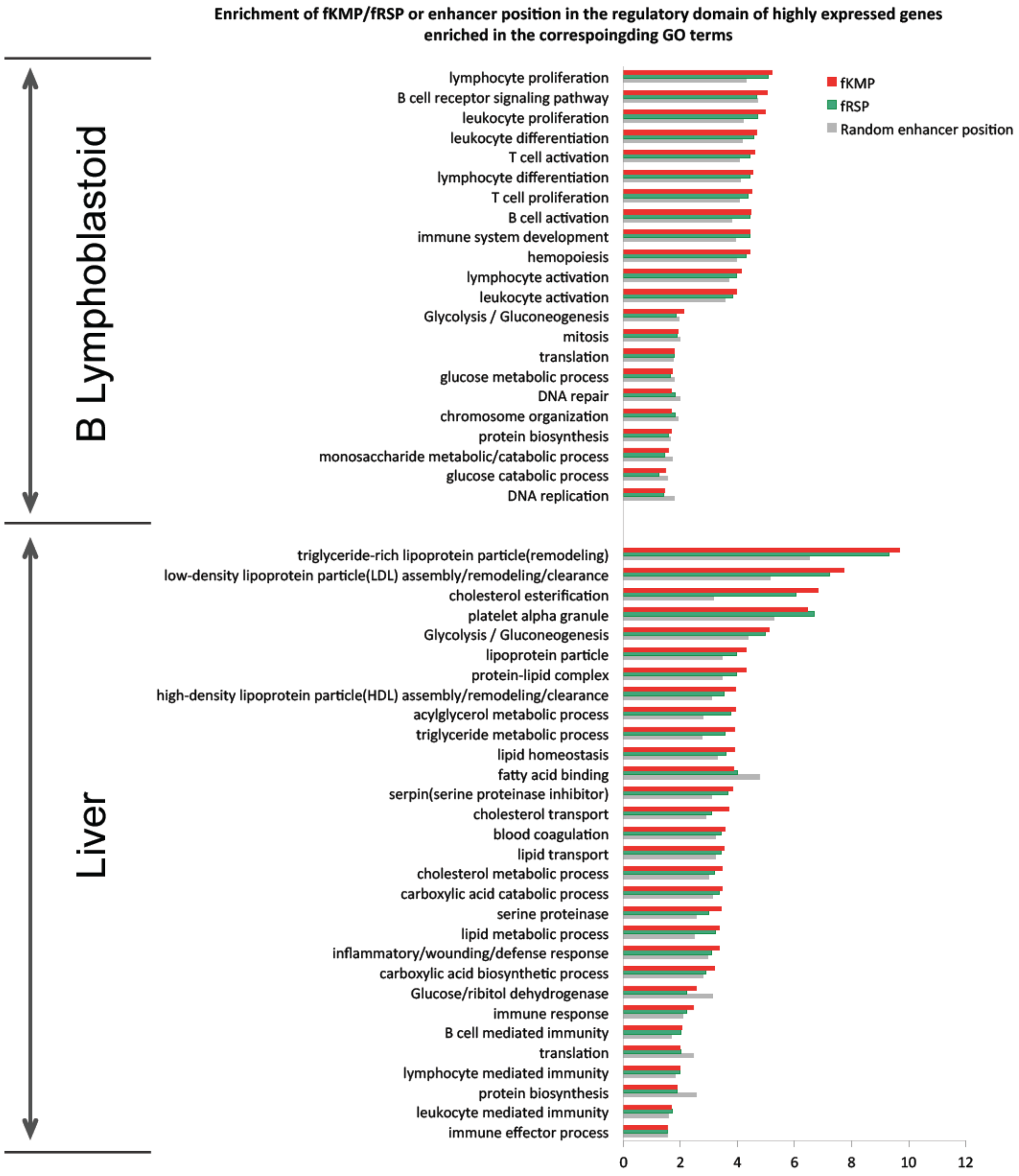
fKMPs and fRSPs are significantly enriched in the regulatory domains of tissue-specific genes. A) HepG2 fKMPs and fRSPs are mostly enriched in the regulatory domains of the top 200 tissue-specific genes in liver B) B lymphoblastoid fKMRs and fRSPs are mostly enriched in the regulatory domains of the top 200 tissue-specific genes in BDCA4+ dentritic cells, CD19+ B cells, CD56+ NK cells, Lymphoma burkitts cells and 721 B lymphoblastoid cells. The cyan dashed line labels the enrichment of KMPs and RSPs in the regulatory domains of 3804 identified housekeeping genes: 1.89 in liver and 1.72 in B lymphoblastoid. The grey dashed line labels the fold enrichment for the background.

Supplementary Figure 32



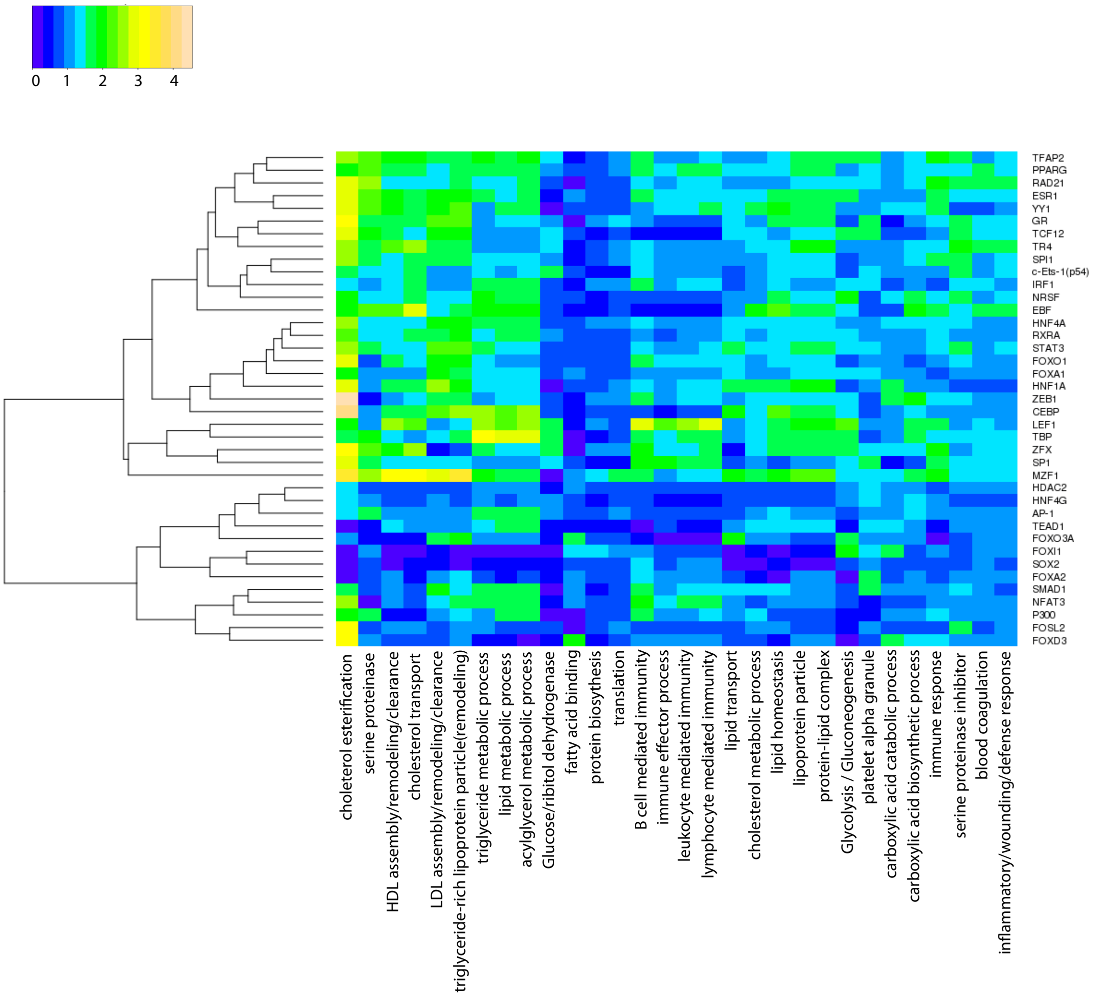
Fraction of overlap between tissue-specific sets of highly expressed genes. Fractions of genes overlapping between the sets of top 500 highly expressed genes for any two tissues.

Supplementary Figure 33



fKMPs and fRSPs are significantly enriched in the regulatory domains of tissue-specific genes. The fKMPs/fRSPs and random enhancer positions are significantly enriched in the regulatory domains of all the listed GO categories with binomial P value < 1.45e-205.

Supplementary Figure 34



Reconstruction of map linking fKMPs to their associated regulatory pathways. Binding sites of major TFs in liver are significantly enriched in the regulatory domains (RDs) of tissue-specific gene categories relative to the random enhancer positions. The value in the heat map represent the $Enrichment_{|TFBS\cap RD(GO)|/|Enhancer\cap RD(GO)}$ which evaluates the enrichment of the TFBS in the regulatory domain of certain GO terms compared to enhancer regions (Supplementary Text). Green and yellow indicate high enrichment of TFBS in the regulatory domain of the corresponding genes.