Published online 17 December 2014 *Nucleic Acids Research, 2015, Vol. 43, No. 1* **225–236**
doi: 10.1093/nar/gku1318

Identifying causal regulatory SNPs in ChIP-seq enhancers

Di Huang and Ivan Ovcharenko*

Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892, USA

Received September 19, 2014; Revised December 04, 2014; Accepted December 05, 2014

ABSTRACT

Thousands of non-coding SNPs have been linked to human diseases in the past. The identification of causal alleles within this pool of disease-associated non-coding SNPs is largely impossible due to the inability to accurately quantify the impact of non-coding variation. To overcome this challenge, we developed a computational model that uses ChIP-seq intensity variation in response to non-coding allelic change as a proxy to the quantification of the biological role of non-coding SNPs. We applied this model to HepG2 enhancers and detected 4796 enhancer SNPs capable of disrupting enhancer activity upon allelic change. These SNPs are significantly over-represented in the binding sites of HNF4 and FOXA families of liver transcription factors and liver eQTLs. In addition, these SNPs are strongly associated with liver GWAS traits, including type I diabetes, and are linked to the abnormal levels of HDL and LDL cholesterol. Our model is directly applicable to any enhancer set for mapping causal regulatory SNPs.

INTRODUCTION

Common phenotypically associated single nucleotide polymorphisms (SNPs) map predominantly to the non-coding DNA regions of the human genome (1–5). More than 90% of SNPs collected in the National Human Genome Research Institute (NHGRI) Genome-wide Association Study (GWAS) catalog (6) are located within non-coding regions (7), the majority of which lacks haplotype protein-coding variants (8), suggesting that the vast majority of SNPs disrupt gene regulation rather than alter the protein-coding sequence or protein structure. Many risk-associated non-coding SNPs (ncSNPs) have been found to affect the activity of regulatory elements. For example, it has been reported that rs2670660—a SNP residing in an intergenic DNA region (∼30 kb from NLRP1 gene)—is transcribed into a non-coding RNA and exerts regulatory effect on monocyte/macrophage transdifferentiation (9,10). The SNPs rs10811656 and rs10757278, located in distal enhancers, were observed to disrupt chromatin conformation and STAT1 binding, inhibit expression of neighboring genes and promote the risk of coronary artery disease (11). In another example, enhancer SNP rs6983267 has been strongly associated with colorectal cancer (12,13). Mutations at this SNP position impair binding of TCF7L2 and alter the transcription of MYC proto-oncogene in colorectal cancer cells (14,15). In addition, the common SNP rs4590952, located in a p53 binding site, has been reported to alter p53 binding activity and significantly influence human cancer risk (16). Although the evidence of individual risk-associated ncSNPs is rapidly emerging, a large (or genome-wide) scale identification of such ncSNPs and the understanding of the mechanisms of regulatory disruption have remained challenging because of the lack of functional annotation of non-coding DNA regions. So far, efforts to prioritize ncSNPs have extensively relied on evolutionary conservation (17,18). With the advance of sequencing techniques, multiple functional genomics lines of evidence became available for more accurate ncSNP classification (19). In RegulomeDB (20) and HaploReg (21), for example, ChIP-seq profiling of histone modifications and transcription factors (TFs), together with the presence of characterized binding motifs, is used to predict functional ncSNPs. Trynka *et al.* developed a computational model exploring H3K4me3 ChIP-seq across cells/tissues to identify potential casual variants (22). ChIP-seq profiling of FOXA1 and ESR1 in breast cancer cells successfully identified risk-associated SNPs and revealed that these SNPs drive allele-specific gene expression through changing the binding affinity of FOXA1 (23). More recently, Kircher *et al.* integrated ChIP-seq data of TFs and histone modification with other genomic features (such as conservation, genomic position, the distribution of CpG sites) into a C-score measuring the deleteriousness of all possible sequence mutations (24).

Here we propose a computational approach for prioritization of SNPs residing in enhancers (dubbed enhSNPs) and prediction of enhSNPs with deleterious properties (Supplementary Figure S1; see the Materials and Methods section). After assembling a set of sequence motifs charac-

*To whom correspondence should be addressed. Tel: +1 301 435 8944; Fax: +1 301 480 2288; Email: ovcharen@nih.gov

teristic to a group of enhancers, we identified enhSNP variants that transform an underlying characteristic motif into a motif uncharacteristic of that enhancer group (dubbed deleterious enhancer SNPs or deSNPs for brevity). We speculated that deSNPs are more likely to increase disease risk or cause a phenotypic change than other enhSNPs, and this speculation is supported by our analysis of genes flanking deSNPs, expression quantitative trait loci (eQTLs) and GWAS experimental reports. We also observed that deSNPs have a substantial impact on the binding affinity of TFs but only a modest impact on the distribution of histone modifications, suggesting a mechanism by which deSNPs cause phenotypic changes.

MATERIALS AND METHODS

Identification of deSNPs

We downloaded 45 107 DNA sequences marked as strong enhancers by ChromHMM in HepG2 cells (25). These sequences constituted our input set of HepG2 enhancers. To analyze sequence signatures of these enhancers, we first generated a control set of sequences through matching the length, GC content and repeat content of enhancer sequences by randomly sampling the sequence of the human genome.

To capture sequence features of given enhancers, we used k-mer sequences, i.e. all DNA fragments of k-bps long. We counted all k-mers in enhancers and controls and ran a series of Fisher's exact tests to identify k-mers significantly enriched in HepG2 enhancers ($P < 1.0 \times 10^{-5}$ after Bonferroni multiple testing correction), and then dubbed these as signal k-mers. The remaining k-mers were named neutral. Next, to account for degeneration and displacement of TFs recognizing their binding sites, we adopted the method of intragenomic replicates (23). Given an SNP and one of its allele, we looked into all DNA k-mers carrying the tested SNPs and sorted out the one(s) that has the highest enhancer enrichment. We marked deSNPs as the SNPs in which all one allele correspond to at least one signal k-mer while other alleles corresponds to neutral k-mers.

To determine the optimal length of k-mers, we investigated the informative regions of known TF binding motifs from TRANSFAC and JASPAR (Supplementary materials), and observed that 80% of these motifs have informative regions of 8 bps or shorter (Supplementary Figure S2). Previously, 8-mers have been successfully used to exploit ChIP-seq data to characterize the binding sequence of FOXA1 in breast cancer (23). Six-mers were used to build support vector machine (SVM) models to predict enhancers in melanocytes (26). To provide a good trade-off between computational complexity and the sequence specificity, we evaluated the selection of 6-mers, 8-mers and 10-mers and observed that HepG2 deSNP identified using 8-mers are the most enriched in proximity of liver-specific genes (i.e. genes highly expressed in the liver as compared with other tissues; Supplementary Figure S3). As such, we used 8-mers for modeling TF binding sites, and identified a total of 549 8-mers significantly enriched in HepG2 enhancers ($P < 1.0 \times 10^{-5}$ after Bonferroni multiple testing correction).

Discovering *de novo* motifs by clustering signal k-mers

To discover *de novo* binding motifs, we clustered the 549 signal 8-mers. Based on the presence/absence of all possible 4-mers along these 8-mers, we clustered all 8-mers using agglomerative hierarchical algorithm, which starts by assigning each 8-mer to an individual cluster and continues by merging two closest clusters until a convergence criterion is reached (Supplementary Figure S4). The distance between two clusters was computed as the average of all inter-cluster 8-mer distances (an inter-cluster distance is the distance between an 8-mer from one cluster and an 8-mer from the other cluster). The clustering process was stopped when the minimal distance between two clusters was greater than 0.6. As a result, we clustered the 549 8-mers into 13 clusters, consisting of 7 to 124 8-mers each.

Next, we aligned the 8-mers within a cluster using the dominant 4-mers, i.e. the 4-mers that occur most frequently in all 8-mer instances of the tested cluster. For each cluster of 8-mers, we started by identifying the 4-mer with the largest occurrence among the 8-mers as a dominant 4-mer. After that, we only examined the 8-mers not harboring the dominant 4-mer(s) and identified the 4-mer occurring most frequently among these 8-mers. This step was repeated until all given 8-mers harbored at least one dominant 4-mer. Using the dominant 4-mers as an anchor, we aligned all given 8-mers together and then derived a position weight matrix (PWM) for the studied 8-mer cluster (Supplementary Figure S5). For each derived PWM, we ran STAMP (27) with default setting to identify the best-matching known TF binding motifs in JASPAR and TRANSFAC (Supplementary Figure S5).

Identification of heterozygous HepG2 enhSNPs

We used ChIP-seq data of histone marks, TFs, P300 and HDAC2 to identify heterozygous SNPs in HepG2 cells. We downloaded ENCODE HepG2 ChIP-seq data from three directories of ENCODE project (Supplementary Table S1),

(i) http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/;
(ii) http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/;
(iii) http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/.

Consistently high polymerase chain reaction bottleneck coefficient values across different ChIP-seq datasets (0.90 ± 0.12) have been observed, indicating high quality of reads and sufficient sequencing library complexity in all utilized ChIP-seq data sets. We only used the uniquely mapped reads and focused on the enhSNPs having the coverage of more than 20 reads, and analyzed whether the presence of two alleles at a SNP position is significantly different from a random expectation using the binomial distribution:

$$P_{\text{ref}} = \text{binomial.test}\left(n_{\text{ref}}, n_{\text{ref}} + n_{\text{alt}}, r_{\text{ref}}\right), \quad (1)$$

where n_{ref} and n_{alt} are the counts of reference and alternative alleles in ChIP-seq reads, respectively. r_{ref} is the expected ratio of the reference allele in ChIP-seq reads. At a

heterozygous site, r_{ref} is expected to be 0.5. However, mapping ChIP-seq reads to the reference genome leads to a bias toward the reads carrying reference allele. To correct for this bias, we estimated r_{ref} using ENCODE ChIP-seq reads at the HepG2 heterozygous sites reported in a previous study (28) where 1 000 000 SNPs were genotyped in HepG2, of which 222 828 sites were detected as heterozygous in HepG2. Among the studied enhSNPs, 9855 sites coincide with these genotyped SNPs, of which 2097 sites were previously reported as heterozygous and others were marked as homozygous (28). We used these genotyped sites as a golden standard reference (GSR). We binned enhSNPs into different categories according to their nucleotide combinations of reference and alternative alleles. In each category, we averaged the reference allele ratios at the GSR heterozygous sites within this category, and used it as the expected reference ratio r_{ref}. When less than 500 GSR heterozygous sites were available, we used the global estimate, i.e. the averaged reference allele ratios of all GSR heterozygous sites, as r_{ref}. Also, our prediction strategy featured high accuracy on GRS (the area under the receiver operating characteristic curve AUC = 0.96, Supplementary Figure S6), further supporting the reliability of our predictions. We chose a cutoff of $P_{ref} > 0.0001$ to detect heterozygous enhSNPs since this setting leads to an optimal balance between the false positive and true positive rates of prediction, i.e. 10% of false positive rate and 95% of true positive rate (Supplementary Figure S6). To the end, we identified 9828 heterozygous enhSNPs, among which 942 were deSNPs.

Allele preference of TFs and histone marks

Using ChIP-seq data targeting a regulatory factor, we focused on the identified heterozygous sites and used a binomial test to evaluate the reference allele preference (i.e. P_{ref}) of the tested regulatory factor at these sites (as illustrated in Equation (1)). The cutoff $P_{ref} < 0.001$ was used to detect significant allele preference (SAP) sites. We applied this procedure on ENCODE ChIP-seq data of TFs and histone markers obtained in HepG2. For regulatory factors with multiple ChIP-seq data sets, we combined all reads together to obtain the possible largest read set. In summary, we obtained ChIP-seq read data for 51 data sets, including 12 chromatin marks and 37 TFs/P300/HDAC2 (Supplementary Table S1).

Nucleotide divergence

For each SNP, we recorded its ancestor allele from the chimpanzee genome and its major human allele. For each set of SNPs, the nucleotide divergence was measured as the fraction of SNPs where the ancestor allele disagrees with the major allele among all human–chimpanzee alignable SNPs. Also, to account for the variable selective constraint across the sequence of human genome, we compared deSNPs with their neighboring enhSNPs. For each deSNP, we collected the most proximal 10 enhSNPs, five downstream and five upstream from the tested deSNPs within the distance up to 1 Mbp.

Neutral reference

To establish a neutral reference, we used pseudogenes, dysfunctional gene homologs (29). Pseudogenes were downloaded from the Peseudogene.org database (30).

Mapping TF binding motifs in DNA sequences

We collected 754 vertebrate TF binding motifs from TRANSFAC (version 2010.3) and JASPAR (31,32) databases and used FIMO (33) with the default settings to map these TF binding motifs along the DNA sequences harboring SNPs.

ChIP-seq TF binding site analysis

The binding sites we used were predicted using ChIP-seq experiments in HepG2 cells. We downloaded narrow peaks generated by the ENCODE Analysis Working Group (AWG) using a uniform processing pipeline (34). In total, we downloaded 77 ChIP-seq data sets for 59 TFs/P300/HDAC2 (Supplementary Table S2). To provide the reliable enrichment estimations, we filtered out the TFs of with ChIP-seq peaks harboring less than 5% of enhSNPs. In total, we retained 22 TFs, including HNF4A, P300 and FOXA1, for the further analysis.

Functional analysis of SNPs based on associated genes

SNPs were associated with genes according to the rule of proximity. That is, an intronic SNP was linked to the host gene, and an intergenic SNP was linked to the most proximal gene. We used a binomial test to address the variable number of the SNPs linked to genes. Given a set of SNPs, we counted these SNPs associated with a given gene category and evaluated the significance of the association between these SNPs and the tested gene category by using

$$P_{alt} = \text{binomial.test}\left(|S_{\cap G}|, |S|, f\right), \quad (2)$$

where S and G are a SNP set and gene category. $|S|$ represents the number of SNPs in S, while $|S_{\cap G}|$ is the number of SNPs in S that are linked to at least one gene in G. f is the fraction of all SNPs that are associated with a gene in G.

We also associated SNPs with genes according to liver eQTLs consisting of 3834 SNP-gene links (35). We used linkage disequilibrium (LD) to expand the liver eQTL SNP set. For a liver eQTL SNP, we identified all SNPs in its tight LD ($r^2 > 0.8$ based on 1000 Genomes Project), and linked these SNPs to the genes associated with the lead SNP (i.e. the tested liver eQTL SNP). After linking genes to eQTL SNPs, we used a binomial test (Equation (2)) to assess the eQTL-based correlation between a SNP set and a gene category. In this case, $|S|$ represents the number of eQTL SNPs in S and $|S_{\cap G}|$ is the number of eQTL SNPs in S that are linked to G, and f is the fraction of all eQTL SNPs that are linked to G.

Functional analysis of SNPs based on GWAS SNP

We downloaded the NHGRI GWAS Catalog in February 2014 (6). From this catalog, we identified 541 traits associated with at least five SNPs each. Next, we extended

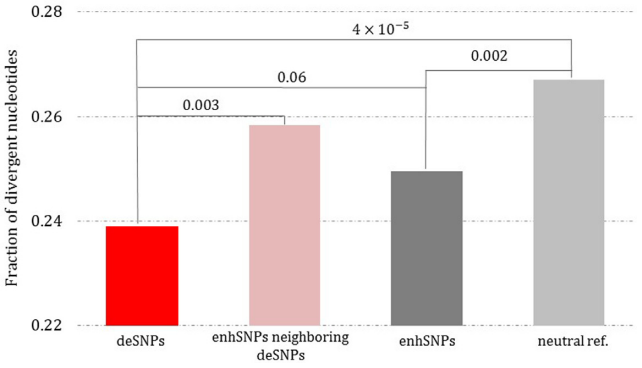

Figure 1. Nucleotide divergence of enhSNPs. Nucleotide divergence is the fraction of enhSNPs where the ancestor alleles disagree with the major alleles among all alignable enhSNPs. The finding that enhSNPs neighboring deSNPs are more conserved than other enhSNPs may be explained by the observation that enhSNPs neighboring deSNPs tend to reside at cytosine nucleotide in CpG sites (Supplementary Figure S8) and that the substitution rate of C to T in CpG has been reported to be significantly higher than other nucleotide substitutions along human genome. Neutral reference are the common SNPs that are located pseudogenes. The values between bars are the significance *P* values of binomial tests.

GWAS SNP sets by identifying all SNPs in a tight LD with a GWAS SNP ($r^2 > 0.8$ based on at one population of 1000 Genomes Project CEU, YRI and CHBJPN) and associated these SNPs with the corresponding traits.

Given a set of SNPs and a GWAS trait, we counted the number of SNPs coinciding with a tested trait. To evaluate the significance of this association, we established a null distribution by randomly generating 1000 independent SNP sets, each being in the same size as the tested SNP set. For each randomly generated SNP set, we counted the number of SNPs associated with the studied trait and used the Poisson distribution to compute significance *P*-value of the observed association.

Enhancers in different cells

We applied our model to other cell lines, including H1-hESC, GM12878, K562, HUVEC, HSMM, NHEK, HeLa-S3 and NHLF (Supplementary Table S3). In all these cell lines except HeLa-S3, the active enhancers have been predicted using ChromHMM (25). In HeLa-S3, we marked H3K27ac ChIP-seq peaks as enhancers. To evaluate the prediction of deSNPs, we also downloaded RNA-seq data published by ENCODE project, and compared cell specificity of deSNPs and enhSNPs using Equation (2).

RESULTS

DeSNPs in HepG2 cells

We started our analysis with the human HepG2 cell line derived from the hepatocellular carcinoma liver tissue. First, we collected all 45 107 DNA sequences marked as strong enhancers by ChromHMM (25)—a computational approach that segments human genome based on the distributions of ChIP-seq histone modification marks—in HepG2 (named HepG2 enhancers for brevity). We analyzed the abundance of all possible 8-mers (8-bp DNA sequences) in HepG2 enhancers and identified 549 8-mers significantly over-represented in these sequences compared to randomly generated controls with matching GC-content, repeat density and length (Fisher's exact test $P < 1.0 \times 10^{-5}$ after accounting for multiple testing) and 31 862 8-mers with insignificant enrichment ($P > 0.01$), which were named signal and neutral k-mers, respectively.

Our framework used these two categories of 8-mers to identify SNPs that disrupt one of the signal k-mers and transform it into a neutral 8-mer (deSNPs; see the Materials and Methods section). Using common SNPs from the 1000 Genomes Project we extracted 78 183 HepG2 enhSNPs. Out of these enhSNPs, 4797 were identified as deSNPs (Supplementary Table S4). Among the identified deSNPs, 1408 deSNPs (30%) feature at least one other deSNP located in the same enhancer, which is significantly higher than expected (student *t*-test $P = 5.0 \times 10^{-9}$; Supplementary Figure S7). This finding is in line with the 'multiple enhancer variant' hypothesis, which postulates that multiple variants located within the same locus cooperatively affect gene expression associated with a common trait/disease (36). By measuring human–chimpanzee nucleotide divergence at enhSNP sites, we observed that enhSNP and deSNP sites show lower nucleotide divergence than the neutral reference (composed of SNPs residing within pseudogenes, binomial test $P < 0.002$, see the Materials and Methods section; Figure 1). Also, considering the fact that the human genome evolves under non-uniform selective pressure (37), we designed a localization strategy to compare deSNPs with their neighboring enhSNPs (see the Materials and Methods section), which demonstrated that deSNP sites exhibit significantly lower nucleotide divergence than their local enhSNP counterparts ($P = 0.003$; Figure 1 and Supplementary Figure S8).

De novo motifs of top 8-mers represent binding sites of liver TFs

Although evolutionary constraint has been widely used to prioritize candidate variants, its statistical power was reported to be marginal at the single-nucleotide level (17,37). We therefore decided to investigate directly if the strong evolutionary conservation of deSNPs is indicative of their co-localization with active TF binding sites.

First, we noticed that the top five ranking 8-mers correspond to the core region (i.e. the region with the highest information content) of the binding motifs of well-known liver TFs HNF4A and FOXO1 [Figure 2a; (38–40)].

Next, to comprehensively investigate all identified 549 signal 8-mers, we clustered these 8-mers based on their sequence similarity. For this purpose, each 8-mer was represented as a collection of all possible 4-mers, and a hierarchical agglomerative algorithm has been applied to bin these 8-mers into 13 clusters, consisting of 7 to 124 8-mer instances each (Supplementary Figures S4 and S5; see the Materials and Methods section). Position weight matrices (PWMs) were then derived to model sequence specificity of each cluster. These PWMs were then aligned with TB binding motif PWMs from the JASPAR and TRANSFAC databases (41) to identify the best matching TF PWM using STAMP (27) (see the Materials and Methods section). We first observed

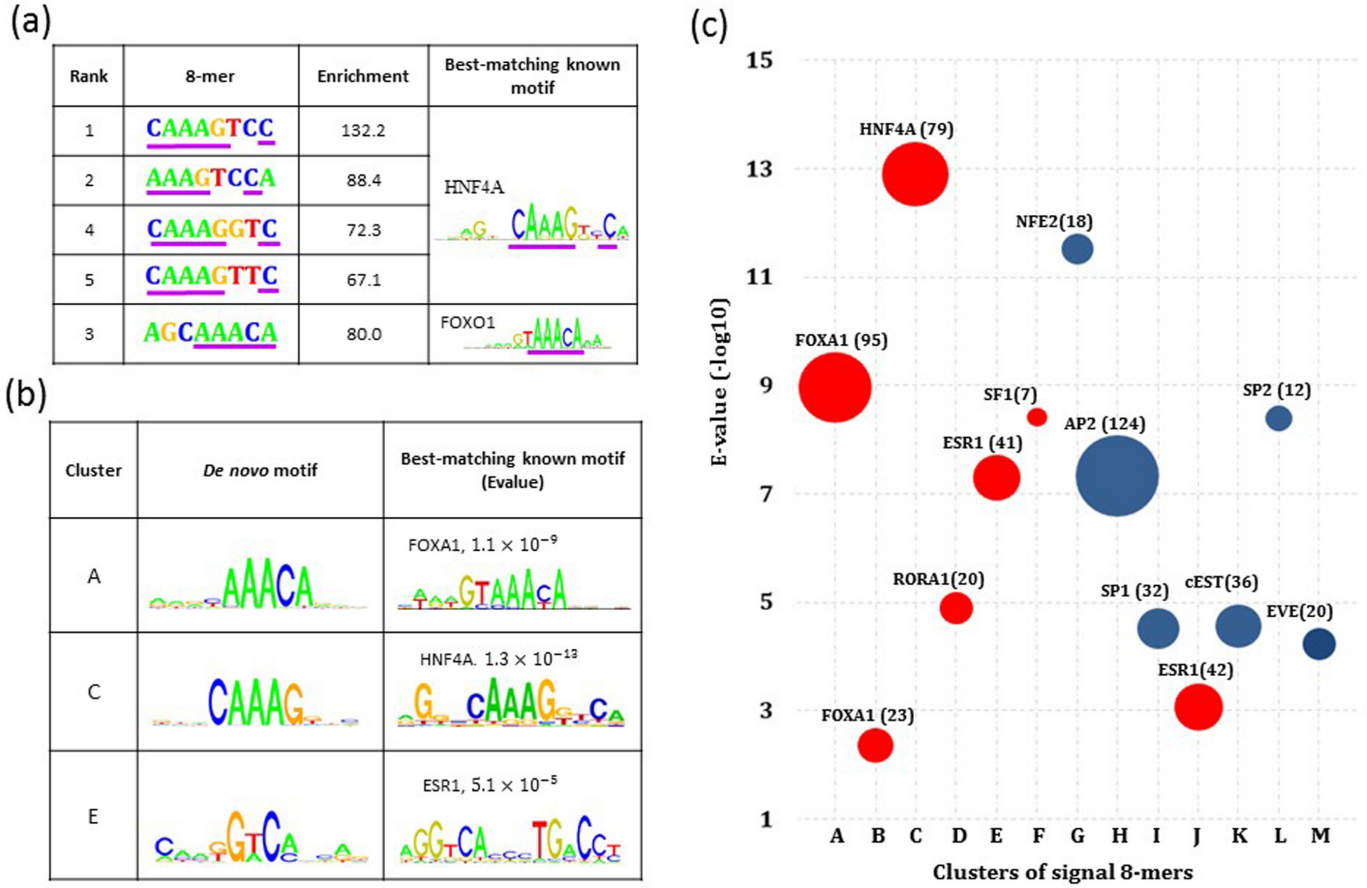

Figure 2. The identified signal 8-mers. (**a**) Top five 8-mers with the strongest enrichment in enhancer sequences and their best matching TF binding motifs reported in TRANSFAC and JASPAR data sets. The aligned positions are underlined in purple. Enrichment is the over-representation of an 8-mer in enhancer sequences as compared to background, measured as $-\log_{10}$(Fisher's test P value). (**b**) Selected three signal 8-mer clusters. For each *de novo* motif, the best-matching known motif was detected using STAMP with the default setting. The values next to motif names are STAMP E-values. A small STAMP E-value indicates a significant similarity between tested motifs. (**c**) Distribution of E-values between *de novo* motifs and their best-matching known motifs. The label by a node presents the TF corresponding to the known motif, and the number of 8-mers in the cluster (which is also indicated by the size of the node). The red nodes mark known liver-specific TFs, while blue nodes indicate generic TFs active in the liver.

that the binding motifs of FOXA1 and HNF4A were almost identical to two PWMs retrieved by our clustering approach (STAMP E-value $<1.0 \times 10^{-9}$; Figure 2b and Supplementary Figure S4). A motif of another cluster matched one arm of the 19-bps palindromic binding sequence of ESR1 (Figure 2b and Supplementary Figure S4), suggesting that the identified 8-mers are capable of characterizing long palindromic TF binding motifs. In addition, the binding motifs of AP2, NFE2 and SP2 matched three other clusters sufficiently well (STAMP E-value $<5 \times 10^{-5}$; Supplementary Figure S5). Overall, 41% (222/549) of signal 8-mers were in the clusters highly correlated with known liver TF motifs (STAMP E-value $< 1 \times 10^{-7}$; Figure 2c and Supplementary Figure S9—see the Materials and Methods section), indicating that the identified signal 8-mers largely correspond to liver TF binding sites in HepG2 enhancers.

We further observed that deSNPs are significantly more over-represented in liver TF binding motifs than enhSNPs (Table 1 and Supplementary Table S5). For example, the binding motifs of nuclear receptors HNF4, PRARA and NUR77 (42–44) are enriched in deSNPs (binomial test $P < 1 \times 10^{-16}$).

While our analysis of signal 8-mer sequence specificities suggested that mutations at deSNPs will likely result in disruption of liver TF binding, we were interested in validating the association of deSNPs with liver TF binding sites directly. For that purpose, we used 59 available HepG2 ChIP-seq data sets (see the Materials and Methods section). Compared to enhSNPs, deSNPs were significantly over-represented in ChIP-seq peaks of 10 TFs ($P < 1 \times 10^{-5}$; Table 2 and Supplementary Table S6). Among these TFs are well-known liver TFs HNF4A, HNF4G, FOXA1 and RXRA, as well as the ubiquitous transcriptional co-activator P300. Also, as compared to enhSNPs, deSNPs were more enriched within ChIP-seq peaks for the epigentic regulator HDAC2 that has been reported to play a critical role in the development of liver cancer (45). The significant over-representation of deSNPs in ChIP-seq peaks of liver TFs, coupled with the motif analysis described above, indicates that deSNP mutations are likely located in the binding sites of liver TFs, providing significant chance that deSNP mutations change the binding affinity of liver TFs, and thus disrupted liver TF binding.

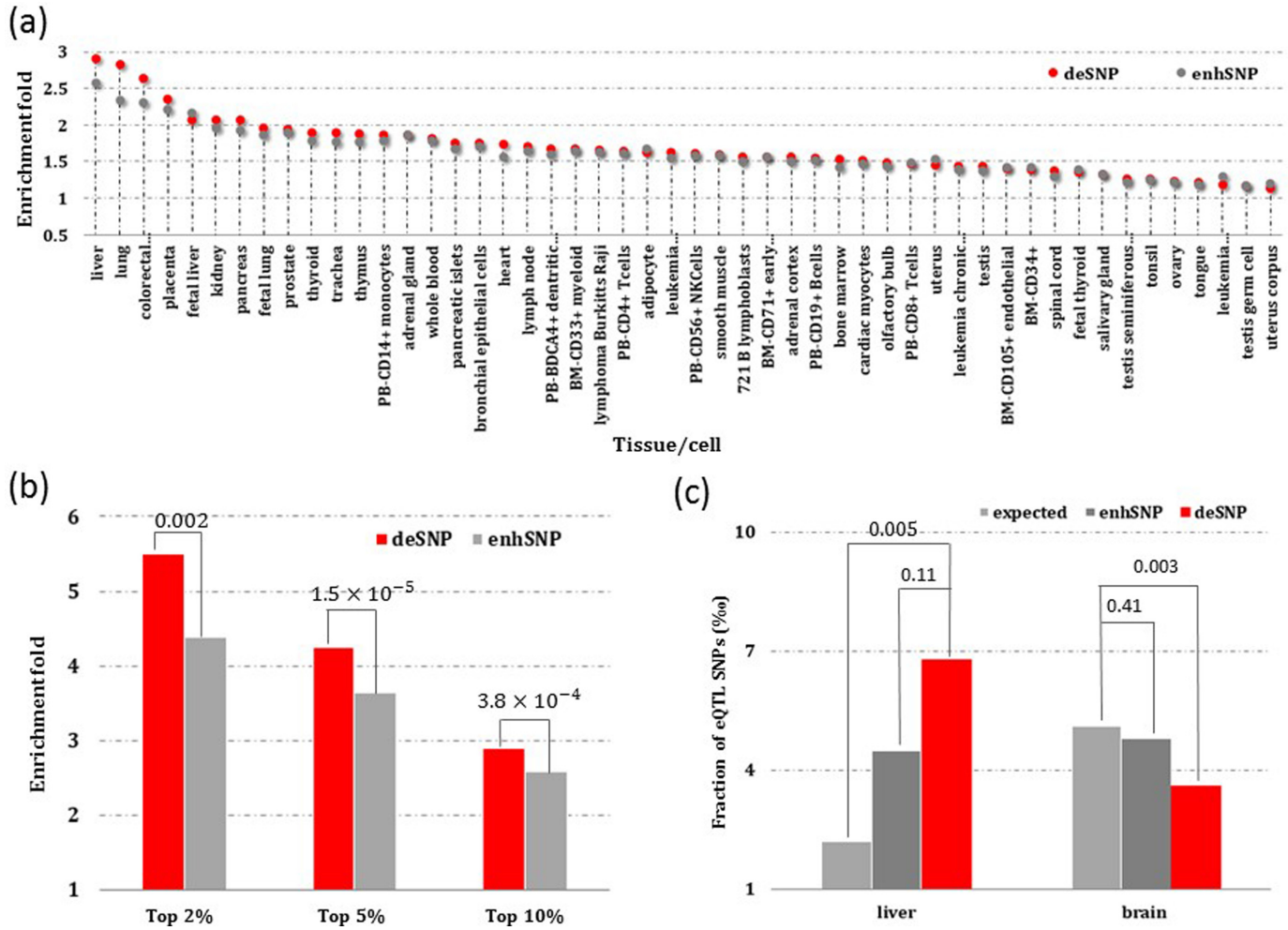

Figure 3. Function analysis of enhSNP based on their associated genes. (**a**) Enrichment of enhSNPs in proximity of tissue-specific genes across 79 tissues/cells. See Supplementary Table S7 for details. (**b**) Enrichment of enhSNPs in proximity of liver-specific genes identified using different criteria. Genes are ranked in a descending order of their expression levels in the liver with respect to other tissues. Top 2%, 5% and 10% of genes in this list were used as liver-specific genes in separate evaluations. The values between bars are binomial test significance *P* values. (**c**) Enrichment of SNPs for eQTLs in the liver and the brain. See Supplementary Table S8 for detailed results.

DeSNPs are located close to liver-specific genes and enriched in liver eQTLs

To evaluate the biological function of deSNP mutations, we turned to the genes associated with these SNPs. We adopted the rule of genomic proximity and linked intronic SNPs to their host genes and intergenic SNPs to the most proximal gene (46). We started with the top 10% of genes highly expressed in each of the 79 human tissues/cell lines (47) and noticed that both enhSNPs and deSNPs demonstrate the highest enrichment in the neighborhood of liver-specific genes (enrichment-fold > 2.5, binomial test $P < 1 \times 10^{-116}$; Figure 3a and Supplementary Table S7). In addition, deSNPs are more enriched in the proximity of liver-specific genes than enhSNPs (enrichment fold = 1.12, hypergeometric test $P = 3.8 \times 10^{-4}$; Figure 3a and Supplementary Table S7). This trend is further escalated with the increasing stringency of selecting liver-specific genes (Figure 3b). As compared to enhSNPs, deSNPs are 1.2 times more frequent in the proximity of the top 5% genes highly expressed in the liver ($P = 1.5 \times 10^{-5}$), and 1.3 times more frequent in the proximity of the top 2% liver genes ($P = 2.3 \times 10^{-3}$). While liver enhSNPs are expected to be located close to liver genes, our results suggest that HepG2 enhancers harboring deSNPs are more likely to be involved in liver regulation than the set of HepG2 enhancers overall.

Since the genomic proximity is not always a reliable indicator of enhancer–gene relationships (48–50), we also studied eQTLs, in which the changes in gene expression are directly associated with genetic variants. Genome-wide eQTL data sets are available for several tissues, including liver and brain (35,51), from the Genotype-Tissue Expression database (52). Relative to all SNPs, both deSNPs and enhSNPs are significantly enriched in liver eQTLs (enrichment fold > 2.1 and hypergeometric test $P < 5 \times 10^{-3}$; Figure 3c and Supplementary Table S8) while deSNPs show stronger liver eQTL association than enhSNPs (enrichment = 1.5, $P = 0.1$). By contrast, the identified deSNPs are significantly depleted of brain eQTLs ($P = 0.003$), while enhSNPs show slight depletion of brain eQTL, as compared

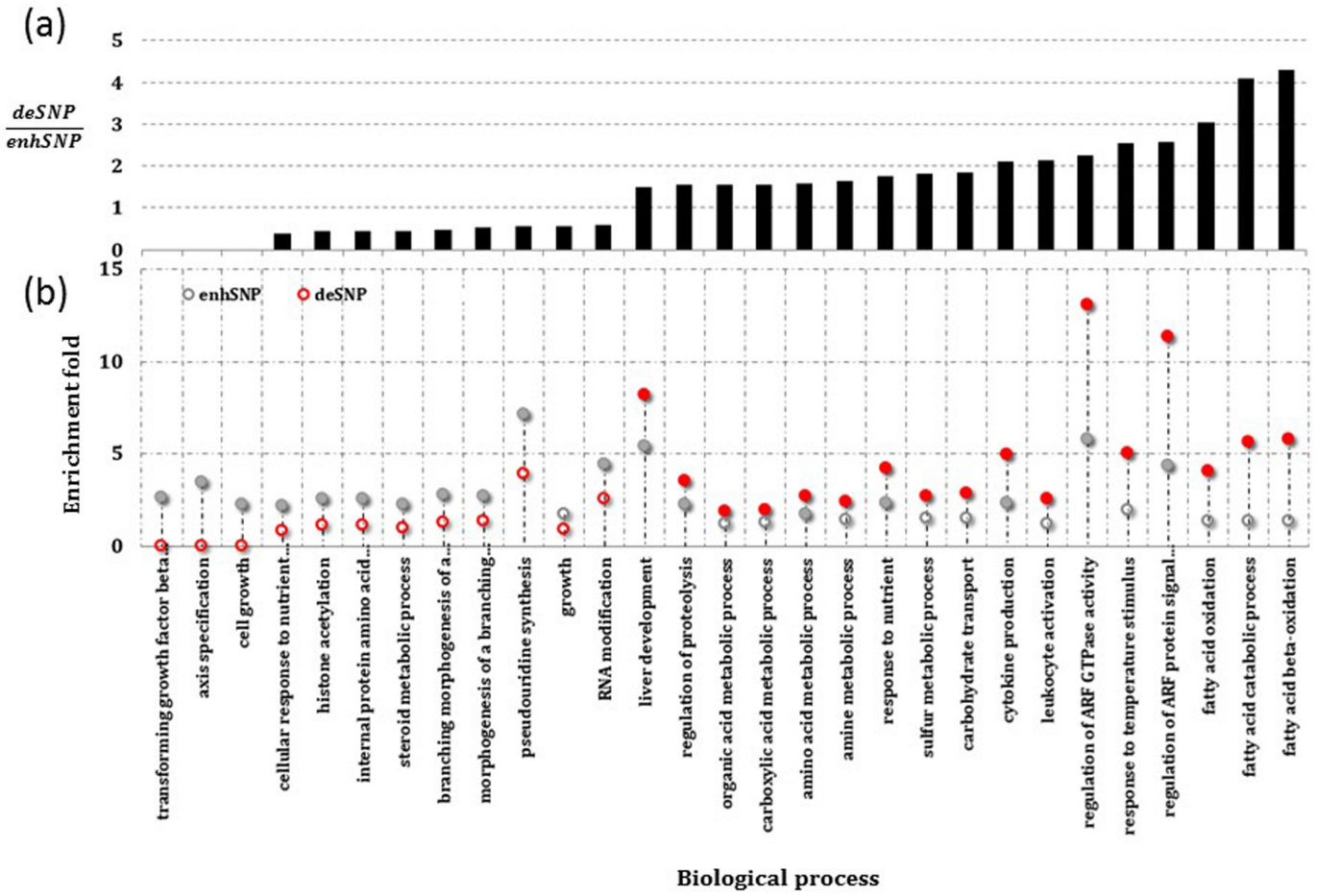

Figure 4. eQTL-based functional analysis of enhSNPs. (**a**) Ratio between deSNP enrichment and enhSNP enrichment across GO biological processes. (**b**) Enrichment fold of deSNPs and enhSNP as compared to expected values. Solid dots suggest a significant enrichment ($P < 0.05$). See Supplementary Table S9 for detailed results.

Table 1. Enrichment of the genome sequences harboring enhSNPs for the known binding motifs collected in TRANSFAC and JASPAR

		Enrichment		Fraction of SNPs located in motif (%)	
Rank	TF motif	deSNP	*P*	deNP	enhSNP
1	FOXP3_01	3.23	<1E-16	8.28	2.56
2	FOXO1_Q5	2.49	<1E-16	7.4	2.97
3	HNF4_Q6_03	2.49	<1E-16	3.67	1.47
4	HNF4_Q6	2.38	<1E-16	4.88	2.05
5	AP2ALPHA_01	2.29	<1E-16	5.82	2.54
6	Zfx-MA0146.2	2.24	<1E-16	6.57	2.93
7	PPARA_Q6	2.22	<1E-16	5.98	2.7
8	NUR77_Q5	2.19	<1E-16	6.71	3.06
9	AP2_Q6_01	2.16	<1E-16	9.94	4.6
10	AREB6_04	2.16	3.33E−16	2.84	1.32
11	HNF3A_01	2.08	<1E-16	5.32	2.55
12	LEF1_Q2_01	2.07	<1E-16	4.09	1.98
13	HNF4_Q6_01	2.04	<1E-16	9.13	4.49
14	PPARG_Q6	2	<1E-16	4.19	2.1
15	SP1_Q4_01	1.97	<1E-16	6.96	3.53
16	E2F1_Q6_01	1.97	1.97E−13	2.86	1.45
17	SOX_01	1.94	<1E-16	4.23	2.18
18	HNF3_Q6	1.9	<1E-16	6.46	3.4
19	Rrxa-MA0512.1	1.87	2.46E−12	2.98	1.59
20	TFAP2C-MA0524.1	1.85	<1E-16	5.17	2.8

Table 2. Enrichment of SNPs in ChIP-seq peaks of TFs/P300/HDAC2

		Enrichment		Fraction of SNPs located in ChIP-seq peak (%)	
Rank	TF	deSNP	*P*	deSNP	enhSNP
1	HNF4A	1.3	6.72E−11	12.31	9.49
2	HDAC2	1.29	9.48E−11	12.62	9.79
3	RXRA	1.25	1.63E−07	10.45	8.35
4	HNF4G	1.25	1.19E−07	10.74	8.59
5	MXI1	1.24	3.86E−06	8.97	7.25
6	RAD21	1.23	9.36E−05	6.57	5.33
7	FOSL2	1.23	9.45E−06	8.59	6.98
8	SP1	1.23	4.52E−09	15.21	12.40
9	SIN3	1.2	4.53E−04	6.61	5.50
10	TBP	1.2	9.48E−04	5.84	4.86
11	CEBPB	1.2	1.52E−04	7.82	6.51
12	POL2	1.2	9.85E−09	17.42	14.50
13	BHLHE40	1.16	7.24E−03	5.19	4.46
14	FOXA2	1.15	1.35E−04	12.37	10.72
15	FOXA1	1.15	9.85E−06	16.5	14.31
16	P300	1.15	9.56E−06	17.58	15.33
17	NFIC	1.14	1.80E−04	14.43	12.70
18	TEAD4	1.13	2.47E−03	10.85	9.64
19	MBD4	1.12	1.65E−02	6.45	5.73
20	JUND	1.12	3.07E−03	10.81	9.63
21	ARID3	1.09	2.88E−02	9.32	8.56
22	MYBL2	1.09	6.33E−03	15.81	14.54
23	DnaseI	1.26	3.22E−15	20.5	16.23

to expected. All of these indicate a potential role deSNPs play in changing of liver-specific gene expression pattern.

Next, we used liver eQTL–gene associations to estimate the function of studied enhSNPs (see the Materials and Methods section) and observed that several liver-related Gene Ontology (GO) categories (53) display significantly higher eQTL-based association with deSNPs than with either all SNP or enhSNPs (Figure 4 and Supplementary Table S9). These GO categories include fatty acid catabolism, ADP-ribosylation factor (ARF), GTPase activity (54) and amino acid biosynthetic process (enrichment fold > 3 and binomial test $P < 3.2 \times 10^{-4}$ as compared to all SNPs; enrichment fold > 2.4 and hypergeometric test $P < 2 \times 10^{-2}$ as compared to enhSNPs; Figure 4 and Supplementary Table S9). Also, stronger association with liver development was observed when comparing deSNPs to all SNPs (enrichment fold $= 11.0$ and $P = 3.7 \times 10^{-5}$) and comparing deSNPs to enhSNPs (enrichment fold $= 1.7$ and $P = 7 \times 10^{-2}$).

Taken together, our results show that deSNP mutations likely cause the change in liver-specific gene expression and are subsequently altering liver-related biological processes.

DeSNPs are a contributing factor to liver-related GWAS traits

We used the NHGRI GWAS Catalog (6) to assess if HepG2 deSNPs are associated with liver disorders and traits. Given a set of SNPs, we calculated the fraction of these SNPs annotated to a GWAS trait and compared this fraction to a random expectation (see the Materials and Methods section). As compared to all SNPs, deSNPs are significantly over-represented in GWAS traits related to liver metabolism, HDL cholesterol level, fasting glucose level, type 1 diabetes and several other liver traits (Figure 5a and Supplementary Table S10). Moreover, as compared to enhSNPs, deSNPs are significantly overrepresented in a total of 7 GWAS traits (hypergeometric test $P \leq 0.05$), including four liver traits (fasting glucose level, serum albumin level, HDL and liver enzyme level) and myocardial infarction, which is known to be strongly associated with liver metabolism and cholesterol production (55). Also, liver or liver-related GWAS traits exhibited significantly higher enrichment ratios of deSNPs versus enhSNPs than other traits ($P = 5 \times 10^{-6}$; Supplementary Figure S10).

As an example, one of deSNPs, rs17744121, overlaps ChIP-seq peaks of HNF4A and FOXA1 (Figure 5b) and resides upstream of NT5E, a gene over-expressed in HepG2 cells, which is under-expressed in liver disease (56). NT5E harbors no intronic enhSNPs, indicative of the intergenic nature of its regulatory variants. As such, we suspected that the genetic variation at rs17744121 might be causal of the expressional change of NT5E. This speculation is further supported by rs17744121 being in a strong LD with rs4954562, an SNP associated with a metabolic GWAS trait. Also, the deSNP rs6726639, located within the binding site of HNFa in HepG2, is in close LD with rs4374383 that has been associated with hepatitis C-induced liver fibrosis by a GWAS study (57). Our results indicate that rs6726639, rather than rs4374383, is likely to be the causal SNP in this locus. In seeking additional evidence for the role of rs6726639 in gene regulation in the liver, we found a previous study that directly linked this SNP to fasting insulin-related disorder, further supporting our deSNP prediction (58).

In addition, rs3756066, one of the identified deSNPs, is located within an LD block with liver eQTL SNP rs10020432 ($r^2 = 0.84$) that has been significantly associated with the transcription regulation of AFP in a liver eQTL study (35) (Supplementary Figure S11). AFP is a

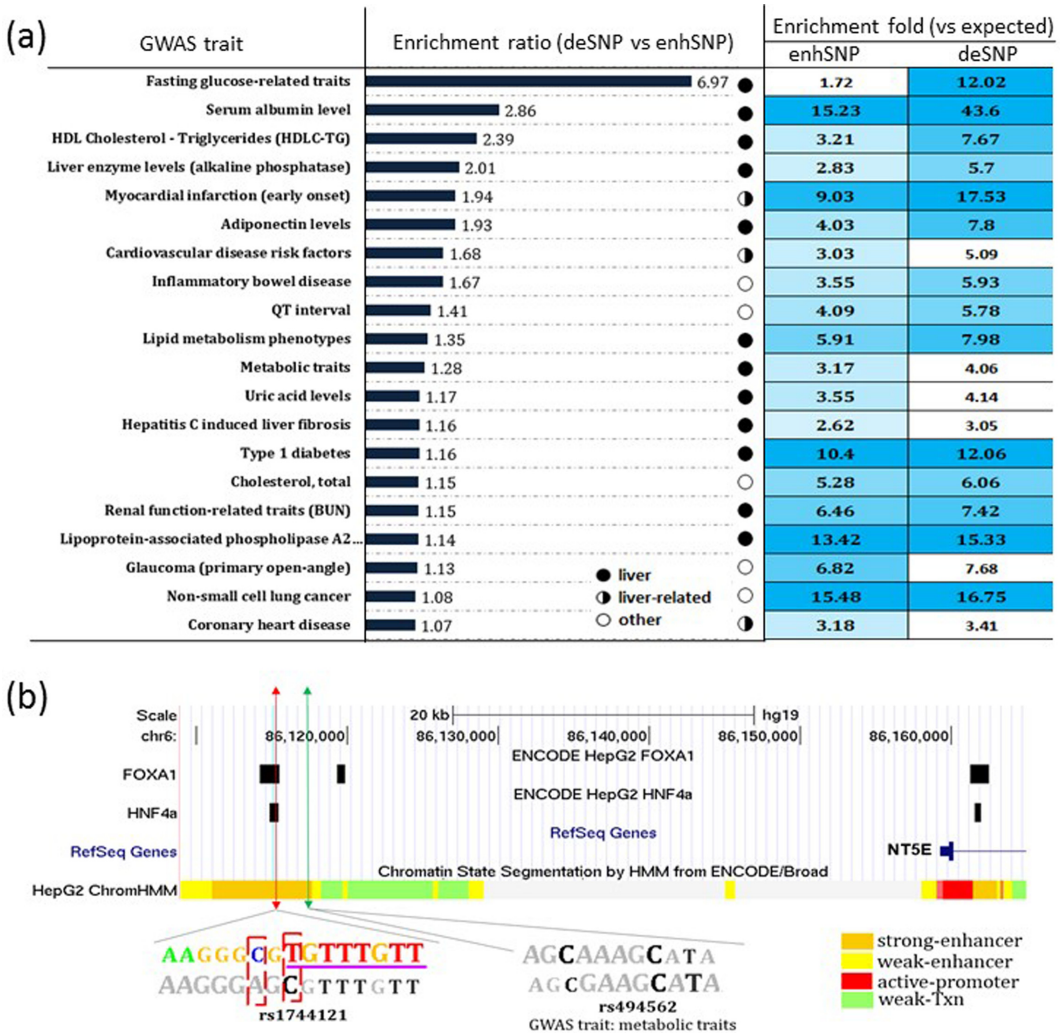

Figure 5. Function analysis of enhSNP based on NHGRI GWAS Catalog. (**a**) Enrichment of deSNPs and enhSNPs for GWAS SNPs as compared with random expectation. A white cell indicates an insignificant enrichment ($P > 10^{-5}$). Full results are presented in Supplementary Table S10. Enrichment ratio is the ratio of enrichment fold of deSNPs versus enhSNPs. (**b**) rs17744121 is demonstrated as an example of deSNPs. The DNA sequences carrying this SNP are presented with its alleles. The color sequences harbor a signal 8-mer as underlined in purple, while the gray sequences harbor neutral 8-mers. Additional examples are presented in Supplementary Figure S11.

gene highly expressed in adult and fetal liver (59), and is a well-known marker of liver cancer (60). Additional deSNP example (rs1109166) is provided in Supplementary Figure S11. These GWAS-based analyses and examples further suggest a strong association between deSNPs and liver-related diseases and phenotypes.

Allele-specific TF binding associated with deSNPs

We used allele-specific distribution of ChIP-seq reads to identify heterozygous enhSNP nucleotides in HepG2 cells. We collected all TF and histone modification ChIP-seq data sets available and predicted that 9828 of 78 183 ChromHMM enhSNPs are heterogeneous in HepG2 (see the Materials and Methods section).

Next, we used these heterozygous sites to explore whether TF binding and histone modification significantly prefer one allele over all other alleles. A high allele preference indicates a regulatory factor selectively recognizing alleles for action. We evaluated the allele-specific preference of TF binding and histone modification at heterogeneous enhSNP positions using HepG2 ChIP-seq data (see the Materials and Methods section). Based on the allele-specific preference profiles, we calculated the fraction of heterozygous enhSNPs showing significant allele preference (SAP, see the Materials and Methods section). We observed that deSNPs are more likely to be SAPs than enhSNPs for liver TFs (FOXA2: 26% deSNPs versus 21% enhSNPs, $P = 0.05$; HNF4A: 21% versus 14%, $P = 0.009$; Figure 6 and Supplementary Table S11). This indicates that a substantial fraction of deSNP mutations result in differential TF binding. We also observed that deSNP variations alter activity of H3K27ac, the histone mark associated with active enhancers (Figure 6 and Supplementary Table S11). However histone modifications show lower allele specificity than liver TFs, which is consistent with the report that histone modifications are less sequence-dependent than TF binding (61–63).

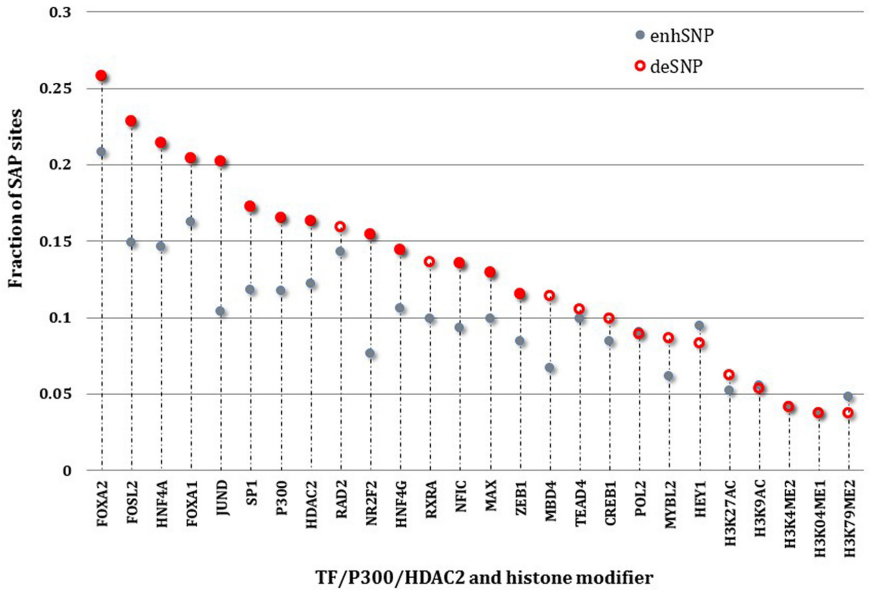

Figure 6. Fraction of SAP sites in deSNPs and enhSNPs across different regulatory factors. A high fraction indicates a strong impact of genetic mutations on regulatory activity. Here, a heterozygous site is predicted as SAP when its reference allele preference $P_{ref} < 0.001$. The solid dots correspond to significant enrichment of deSNPs as compared to enhSNPs ($P \leq 0.05$).

DISCUSSION

In this study, we modeled sequence composition of HepG2 enhancers and identified deSNPs—enhancer SNPs with likely deleterious impact on the biological function of the enhancers. We demonstrated that deSNPs exhibit lower nucleotide divergence than enhSNPs and commonly reside in active TF binding sites. After associating deSNPs with their target genes using either the rule of proximity or the reported liver eQTLs, we observed that the deSNPs display significant association with liver-specific genes and liver-related biological processes. In addition, the deSNPs are significantly associated with liver-related GWAS traits and commonly exhibit allele-specificity of TF binding. All these observations support the hypothesis of a strong connection between deSNPs and disrupted regulation of genes. The developed computational framework can be applied to other tissues/cell lines if the corresponding data of TFs and epigenetic marks is available.

We characterized the allele-specific impact of SNPs, which enables us to discriminate deleterious enhSNPs from the enhSNPs that are located in TF binding sites but have no disruptive impact on TF binding. This is the major difference between our framework and the related methods (20–22,24), which led us to prioritize causal regulatory SNPs in an effective way. For example, the previously published method Cscore combines various informative data (including ChIP-seq-based genome annotations) to estimate damaging extent of all possible sequence mutations (24). While Cscore has not been tailored to identify mutations deleterious to enhancer function, it could be used as an independent validation of the deleterious effects of deSNPs. The identified deSNPs exhibit a significantly higher Cscore than all enhSNPs (student t-test $P = 2.1 \times 10^{-3}$; Supplementary Figure S12), indicating a general agreement between our prediction and Cscore. However, enhSNPs in which all the alleles correspond to at least one signal 8-mer also correspond to high Cscores ($P = 1.7 \times 10^{-9}$, versus all enhSNPs), which are even slightly greater than deSNPs ($P = 8 \times 10^{-2}$). This indicates a lack of specificity in Cscore assignments as compared to our method. Therefore, our model, in which allelic specificity of enhSNPs is quantified, is better suited for the identification of disease-causal enhancer SNPs.

To check how generalizable our approach is, we identified deSNPs in other cell types, in which enhancer maps have been established previously (25), including GM12878, H1-hESC, K562, HSMM, HeLa-S3, HUVEC, NHEK and NHLF (Supplementary Table S3). The identified deSNPs are significantly enriched in the neighborhood of the genes highly expressed in the corresponding cells using 8-mers ($P < 1 \times 10^{-5}$, deSNP versus enhSNPs; Supplementary Figure S14). These results indicate that our approach can be directly applied to a different cell type. Also the enrichment of the identified deSNPs over enhSNPs varied across different cell lines (Supplementary Figure S14). It is possible that different k-mers might be an optimal choice for different cells.

With the increasing amount of genomic and epigenomic data, we anticipate to expand our framework to model allele-specific regulatory activity and then to identify disease-associated SNPs and to reveal how they impact gene regulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENT

We are grateful to Gesine Cauer for critical reading of the manuscript.

FUNDING

Intramural Research Program of the National Institutes of Health, National Library of Medicine. Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.
Conflict of interest statement. None declared.

REFERENCES

1. Fehrmann,R.S.N., Jansen,R.C., Veldink,J.H., Westra,H.-J., Arends,D., Bonder,M.J., Fu,J., Deelen,P., Groen,H.J.M., Smolonska,A. *et al.* (2011) Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.*, **7**, e1002197.
2. Nica,A.C., Montgomery,S.B., Dimas,A.S., Stranger,B.E., Beazley,C., Barroso,I. and Dermitzakis,E.T. (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.*, **6**, e1000895.
3. Schork,A.J., Thompson,W.K., Pham,P., Torkamani,A., Roddey,J.C., Sullivan,P.F., Kelsoe,J.R., O'Donovan,M.C., Furberg,H., Schork,N.J. *et al.* (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet.*, **9**, e1003449.
4. Conde,L., Bracci,P.M., Richardson,R., Montgomery,S.B. and Skibola,C.F. (2013) Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. *Am. J. Hum. Genet.*, **92**, 126–130.
5. Yao,L., Tak,Y.G., Berman,B.P. and Farnham,P.J. (2014) Functional annotation of colon cancer risk SNPs. *Nat Commun.*, **5**, 1–13.
6. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
7. Kumar,V., Westra,H.-J., Karjalainen,J., Zhernakova,D.V., Esko,T., Hrdlickova,B., Almeida,R., Zhernakova,A., Reinmaa,E., Võsa,U. *et al.* (2013) Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet.*, **9**, e1003201.
8. The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
9. Glinskii,A.B., Ma,J., Ma,S., Grant,D., Lim,C.-U., Sell,S. and Glinsky,G. (2009) Identification of intergenic trans-regulatory RNAs containing a disease-linked SNP sequence and targeting cell cycle progression/differentiation pathways in multiple common human disorders. *Cell Cycle*, **8**, 3925–3942.
10. Glinskii,A.B., Ma,S., Ma,J., Grant,D., Lim,C.-U., Guest,I., Sell,S., Buttyan,R. and Glinsky,G.V. (2011) Networks of intergenic long-range enhancers and snpRNAs drive castration-resistant phenotype of prostate cancer and contribute to pathogenesis of multiple common human disorders. *Cell Cycle*, **10**, 3571–3597.
11. Harismendy,O., Notani,D., Song,X., Rahim,N.G., Tanasa,B., Heintzman,N., Ren,B., Fu,X.-D., Topol,E.J., Rosenfeld,M.G. *et al.* (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-[ggr] signalling response. *Nature*, **470**, 264–268.
12. Tomlinson,I., Webb,E., Carvajal-Carmona,L., Broderick,P., Kemp,Z., Spain,S., Penegar,S., Chandler,I., Gorman,M., Wood,W. *et al.* (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.*, **39**, 984–988.
13. Haiman,C.A., Le Marchand,L., Yamamato,J., Stram,D.O., Sheng,X., Kolonel,L.N., Wu,A.H., Reich,D. and Henderson,B.E. (2007) A common genetic risk factor for colorectal and prostate cancer. *Nat. Genet.*, **39**, 954–956.
14. Pomerantz,M.M., Ahmadiyeh,N., Jia,L., Herman,P., Verzi,M.P., Doddapaneni,H., Beckwith,C.A., Chan,J.A., Hills,A., Davis,M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.
15. Sotelo,J., Esposito,D., Duhagon,M.A., Banfield,K., Mehalko,J., Liao,H., Stephens,R.M., Harris,T.J.R., Munroe,D.J. and Wu,X. (2010) Long-range enhancers on 8q24 regulate c-Myc. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 3001–3005.
16. Zeron-Medina,J., Wang,X., Repapi,E., Campbell,M.R., Su,D., Castro-Giner,F., Davies,B., Peterse,E.F.P., Sacilotto,N., Walker,G.J. *et al.* A polymorphic p53 response element in KIT ligand influences cancer risk and has undergone natural selection. *Cell*, **155**, 410–422.
17. Cooper,G.M., Goode,D.L., Ng,S.B., Sidow,A., Bamshad,M.J., Shendure,J. and Nickerson,D.A. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods*, **7**, 250–251.
18. Poitras,L., Yu,M., Lesage-Pelletier,C., MacDonald,R.B., Gagné,J.-P., Hatch,G., Kelly,I., Hamilton,S.P., Rubenstein,J.L.R., Poirier,G.G. *et al.* (2010) An SNP in an ultraconserved regulatory element affects Dlx5/Dlx6 regulation in the forebrain. *Development*, **137**, 3089–3097.
19. Kasowski,M., Grubert,F., Heffelfinger,C., Hariharan,M., Asabere,A., Waszak,S.M., Habegger,L., Rozowsky,J., Shi,M., Urban,A.E. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.
20. Boyle,A.P., Hong,E.L., Hariharan,M., Cheng,Y., Schaub,M.A., Kasowski,M., Karczewski,K.J., Park,J., Hitz,B.C., Weng,S. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, **22**, 1790–1797.
21. Ward,L.D. and Kellis,M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930–D934.
22. Trynka,G., Sandor,C., Han,B., Xu,H., Stranger,B.E., Liu,X.S. and Raychaudhuri,S. (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.*, **45**, 124–130.
23. Cowper-Sallari,R., Zhang,X., Wright,J.B., Bailey,S.D., Cole,M.D., Eeckhoute,J., Moore,J.H. and Lupien,M. (2012) Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.*, **44**, 1191–1198.
24. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
25. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
26. Gorkin,D.U., Lee,D., Reed,X., Fletez-Brant,C., Bessling,S.L., Loftus,S.K., Beer,M.A., Pavan,W.J. and McCallion,A.S. (2012) Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. *Genome Res.*, **22**, 2290–2301.
27. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.
28. Motallebipour,M., Ameur,A., Reddy Bysani,M.S., Patra,K., Wallerman,O., Mangion,J., Barker,M., McKernan,K., Komorowski,J. and Wadelius,C. (2009) Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biol.*, **10**, R129.
29. Vanin,E.F. (1985) Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.*, **19**, 253–272.
30. Balasubramanian,S., Zheng,D., Liu,Y.-J., Fang,G., Frankish,A., Carriero,N., Robilotto,R., Cayting,P. and Gerstein,M. (2009) Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol.*, **10**, R2.
31. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
32. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
33. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
34. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.

35. Schadt,E.E., Molony,C., Chudin,E., Hao,K., Yang,X., Lum,P.Y., Kasarskis,A., Zhang,B., Wang,S., Suver,C. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.
36. Corradin,O., Saiakhova,A., Akhtar-Zaidi,B., Myeroff,L., Willis,J., Cowper-Sal·lari,R., Lupien,M., Markowitz,S. and Scacheri,P.C. (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, **24**, 1–13.
37. MacArthur,D.G., Manolio,T.A., Dimmock,D.P., Rehm,H.L., Shendure,J., Abecasis,G.R., Adams,D.R., Altman,R.B., Antonarakis,S.E., Ashley,E.A. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**, 469–476.
38. Vatamaniuk,M.Z., Gupta,R.K., Lantz,K.A., Doliba,N.M., Matschinsky,F.M. and Kaestner,K.H. (2006) Foxa1-deficient mice exhibit impaired insulin secretion due to uncoupled oxidative phosphorylation. *Diabetes*, **55**, 2730–2736.
39. Wagner,M., Zollner,G. and Trauner,M. (2011) Nuclear receptors in liver disease. *Hepatology*, **53**, 1023–1034.
40. Ning,B.-F., Ding,J., Yin,C., Zhong,W., Wu,K., Zeng,X., Yang,W., Chen,Y.-X., Zhang,J.-P., Zhang,X. *et al.* (2010) Hepatocyte nuclear factor 4α suppresses the development of hepatocellular carcinoma. *Cancer Res.*, **70**, 7640–7651.
41. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
42. Pols,T.W.H., Ottenhoff,R., Vos,M., Levels,J.H.M., Quax,P.H.A., Meijers,J.C.M., Pannekoek,H., Groen,A.K. and de Vries,C.J.M. (2008) Nur77 modulates hepatic lipid metabolism through suppression of SREBP1c activity. *Biochem. Biophys. Res. Commun.*, **366**, 910–916.
43. Chakravarthy,M.V., Pan,Z., Zhu,Y., Tordjman,K., Schneider,J.G., Coleman,T., Turk,J. and Semenkovich,C.F. (2005) 'New' hepatic fat activates PPARα to maintain glucose, lipid, and cholesterol homeostasis. *Cell Metab.*, **1**, 309–322.
44. Sladek,F.M., Zhong,W.M., Lai,E. and Darnell,J.E. (1990) Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. *Genes Dev.*, **4**, 2353–2365.
45. Noh,J.H., Bae,H.J., Eun,J.W., Shen,Q., Park,S.J., Kim,H.S., Nam,B., Shin,W.C., Lee,E.K., Lee,K. *et al.* (2014) HDAC2 provides a critical support to malignant progression of hepatocellular carcinoma through feedback control of mTORC1 and AKT. *Cancer Res.*, **74**, 1728–1738.
46. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotech.*, **28**, 495–501.
47. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 6062–6067.
48. Huang,D. and Ovcharenko,I. (2014) Genome-wide analysis of functional and evolutional features of tele-enhancers. *G3: Genes|Genomes|Genetics*,**4**, 579–593.
49. Zhang,Y., Wong,C.-H., Birnbaum,R.Y., Li,G., Favaro,R., Ngan,C.Y., Lim,J., Tai,E., Poh,H.M., Wong,E. *et al.* (2013) Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, **504**, 306–310.
50. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
51. Gibbs,J.R., van der Brug,M.P., Hernandez,D.G., Traynor,B.J., Nalls,M.A., Lai,S.-L., Arepalli,S., Dillman,A., Rafferty,I.P., Troncoso,J. *et al.* (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.*, **6**, e1000952.
52. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
53. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
54. Suzuki,T., Kanai,Y., Hara,T., Sasaki,J., Sasaki,T., Kohara,M., Maehama,T., Taya,C., Shitara,H., Yonekawa,H. *et al.* (2006) Crucial role of the small GTPase ARF6 in hepatic cord formation during liver development. *Mol. Cell. Biol.*, **26**, 6149–6156.
55. von KÃnel,R., Abbas,C., Begré,S., Gander,M.-L., Saner,H. and Schmid,J.-P. (2010) Association between posttraumatic stress disorder following myocardial infarction and liver enzyme levels: a prospective study. *Dig. Dis. Sci.*, **55**, 2614–2623.
56. Snider,N.T., Griggs,N.W., Singla,A., Moons,D.S., Weerasinghe,S.V.W., Lok,A.S., Ruan,C., Burant,C.F., Conjeevaram,H.S. and Omary,M.B. (2013) CD73 (ecto-5′-nucleotidase) hepatocyte levels differ across mouse strains and contribute to mallory-denk body formation. *Hepatology*, **58**, 1790–1800.
57. Patin,E., Kutalik,Z., Guergnon,J., Bibert,S., Nalpas,B., Jouanguy,E., Munteanu,M., Bousquet,L., Argiro,L., Halfon,P. *et al.* (2012) Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. *Gastroenterology*, **143**, 1244–1252; e1212.
58. Dupuis,J., Langenberg,C., Prokopenko,I., Saxena,R., Soranzo,N., Jackson,A.U., Wheeler,E., Glazer,N.L., Bouatia-Naji,N., Gloyn,A.L. *et al.* (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.*, **42**, 105–116.
59. Lemire,J.M. and Fausto,N. (1991) Multiple alpha-fetoprotein RNAs in adult rat liver: cell type-specific expression and differential regulation. *Cancer Res.*, **51**, 4656–4664.
60. Debruyne,E.N. and Delanghe,J.R. (2008) Diagnosing and monitoring hepatocellular carcinoma with alpha-fetoprotein: New aspects and applications. *Clin. Chim. Acta*, **395**, 19–26.
61. Heinz,S., Romanoski,C.E., Benner,C., Allison,K.A., Kaikkonen,M.U., Orozco,L.D. and Glass,C.K. (2013) Effect of natural genetic variation on enhancer selection and function. *Nature*, **503**, 487–492.
62. Mullen,A.C., Orlando,D.A., Newman,J.J., Lovén,J., Kumar,R.M., Bilodeau,S., Reddy,J., Guenther,M.G., DeKoter,R.P. and Young,R.A. (2011) Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell*, **147**, 565–576.
63. Doege,C.A., Inoue,K., Yamashita,T., Rhee,D.B., Travis,S., Fujita,R., Guarnieri,P., Bhagat,G., Vanti,W.B., Shih,A. *et al.* (2012) Early-stage epigenetic modification during somatic cell reprogramming by Parp1 and Tet2. *Nature*, **488**, 652–655.