Article

Systematic analysis of binding of transcription factors to noncoding variants

https://doi.org/10.1038/s41586-021-03211-0

Received: 28 November 2018

Accepted: 11 December 2020

Published online: 27 January 2021

Check for updates

Jian Yan[1,2,3,4,12]✉, Yunjiang Qiu[2,5,12], André M. Ribeiro dos Santos[2,6,12], Yimeng Yin[4,7], Yang E. Li[2,8], Nick Vinckier[9], Naoki Nariai[9], Paola Benaglio[9], Anugraha Raman[2,5], Xiaoyu Li[1,3], Shicai Fan[9], Joshua Chiou[9], Fulin Chen[1], Kelly A. Frazer[9], Kyle J. Gaulton[9], Maike Sander[8,9], Jussi Taipale[4,7,10]✉ & Bing Ren[2,8,11]✉

Many sequence variants have been linked to complex human traits and diseases[1], but deciphering their biological functions remains challenging, as most of them reside in noncoding DNA. Here we have systematically assessed the binding of 270 human transcription factors to 95,886 noncoding variants in the human genome using an ultra-high-throughput multiplex protein–DNA binding assay, termed single-nucleotide polymorphism evaluation by systematic evolution of ligands by exponential enrichment (SNP-SELEX). The resulting 828 million measurements of transcription factor–DNA interactions enable estimation of the relative affinity of these transcription factors to each variant in vitro and evaluation of the current methods to predict the effects of noncoding variants on transcription factor binding. We show that the position weight matrices of most transcription factors lack sufficient predictive power, whereas the support vector machine combined with the gapped *k*-mer representation show much improved performance, when assessed on results from independent SNP-SELEX experiments involving a new set of 61,020 sequence variants. We report highly predictive models for 94 human transcription factors and demonstrate their utility in genome-wide association studies and understanding of the molecular pathways involved in diverse human traits and diseases.

Genome-wide association studies (GWAS) have implicated hundreds of thousands of single-nucleotide polymorphisms (SNPs) in human diseases and traits[1], but very few of them have been mechanistically characterized. This is in part due to incomplete knowledge of the DNA binding specificity of human transcription factors[2]. To systematically characterize the effects of noncoding variants on transcription factor binding to DNA, we adopted an ultra-high-throughput, multiplexed transcription factor–DNA binding assay, high-throughput (HT)-SELEX[3], to examine in vitro binding of human transcription factors to common sequence variants using a sampling scheme that surveys candidate *cis*-regulatory variants near the reported type-2 diabetes (T2D) risk loci. Whereas HT-SELEX used randomized DNA sequences as input, SNP-SELEX used a library of 40-bp DNA matching the reference human genomic sequence, with the centre position corresponding to tested SNPs permutated to all four bases (Fig. 1a, Extended Data Fig. 1a). At the start of this project, 110 distinct tagging SNPs had been linked to T2D susceptibility. We designed 6,724 DNA oligonucleotides to represent these tagging variants as well as the SNPs in linkage disequilibrium with them ($r^2 \geq 0.8$). We additionally designed oligonucleotides to cover a much larger pool of 89,162 common SNPs in annotated candidate *cis*-regulatory sequences located within 500 kb of these T2D-tagging SNPs (Supplementary Table 1). Thus, the input DNA library contained a total of 383,544 distinct oligonucleotides corresponding to 95,886 SNPs. The sequence features of these genomic fragments closely resembled those of the rest of the human reference genome, especially the fraction containing transcription factor binding sites and DNase I hypersensitive elements (Extended Data Fig. 1b, c).

The enrichment of each oligonucleotide could be used to estimate the relative affinity between the transcription factor and the DNA (Extended Data Fig. 2a). We conducted a total of 768 SNP-SELEX experiments including 751 recombinant transcription factor proteins and protein-free controls (Supplementary Table 2). Overall, 360 experiments passed quality control and were subjected to subsequent analyses (Supplementary File 1). Altogether, we obtained about 828 million measurements of transcription factor–DNA interactions.

We first computed the relative enrichment of DNA sequences in the pool as an odds ratio after each cycle of experiment and then defined the oligonucleotide binding score (OBS) as the cumulative area under the curve (AUC) of enrichment values across the six rounds of SNP-SELEX, which reflects the relative binding affinity of the 40-mer sequence to the transcription factor (Fig. 1b, Extended Data Fig. 2b). This computational strategy could effectively retain information of all SNP-SELEX

[1]School of Medicine, Northwest University, Xi'an, China. [2]Ludwig Institute for Cancer Research, La Jolla, CA, USA. [3]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong SAR, China. [4]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Solna, Sweden. [5]Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA, USA. [6]Universidade Federal do Pará, Institute of Biological Sciences, Belém, Brazil. [7]Department of Biochemistry, University of Cambridge, Cambridge, UK. [8]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA. [9]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. [10]Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland. [11]Center for Epigenomics, University of California San Diego, La Jolla, CA, USA. [12]These authors contributed equally: Jian Yan, Yunjiang Qiu, André M. Ribeiro dos Santos. ✉e-mail: jian.yan@cityu.edu.hk; ajt208@cam.ac.uk; biren@ucsd.edu

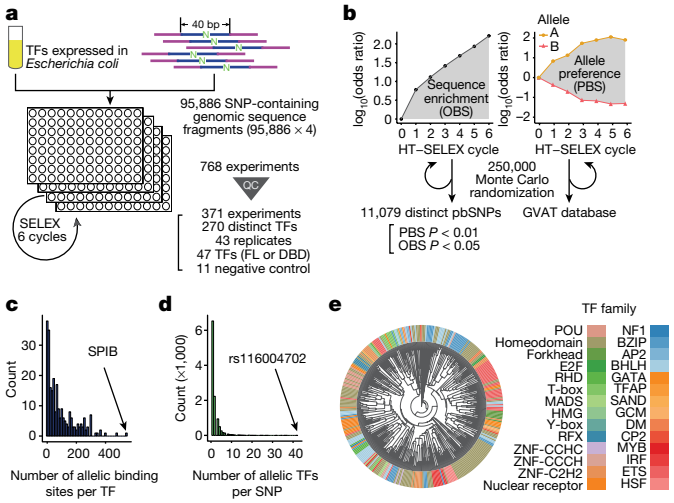


Fig. 1 | High-throughput analysis of the binding of human transcription factors to common sequence variants by SNP-SELEX. a, An overview of the SNP-SELEX experimental procedure. N indicates the position of SNPs. FL, full length; QC, quality control; TFs, transcription factors. **b**, The data obtained from each SELEX cycle were analysed to determine OBS and PBS. Two alleles of the SNP are shown, the reference allele (triangle) and the alternative allele (circle). Differential binding information for all SNPs tested is publicly available from the GVAT database. **c**, **d**, Histograms show the number of pbSNPs bound by each transcription factor (**c**), and the number of transcription factors showing allelic binding for each pbSNP (**d**). **e**, A clustering diagram of transcription factors tested in this study was generated on the basis of the pairwise Pearson correlation of their DNA binding specificity from the SNP-SELEX data. For each pair of experiments, we computed the Pearson correlation coefficient (PCC) and dissimilarity (1 − PCC) of PBS between significantly enriched oligonucleotides in both experiments and clustered them using the UPGMA algorithm.

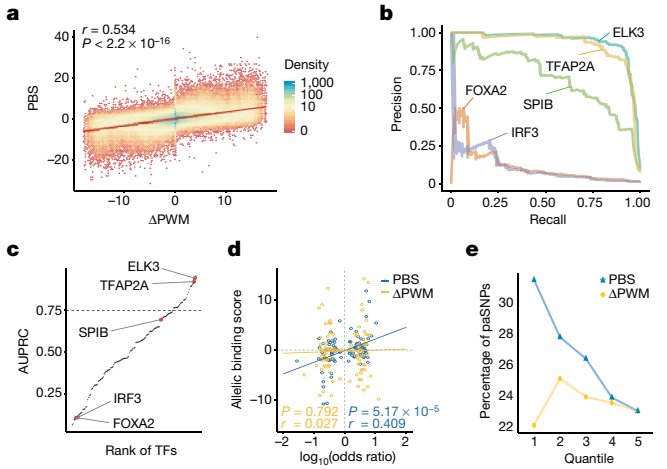


Fig. 2 | Evaluation of the current PWM models using the SNP-SELEX data. a, Scatter plot with PBS on the *y*-axis and ΔPWM scores on the *x*-axis. The red line denotes a linear regression of PBS as a function of ΔPWM. PCC and *P* value calculated by two-sided *t*-test are shown. **b**, Examples of the precision–recall curve show the variation of performance of different PWM models in predicting pbSNPs. **c**, Scatter plot ranking the predictive performance of 129 PWMs. AUPRC of only 24 transcription factors exceeded 0.75. Transcription factors shown in **b** are highlighted with red dots. **d**, Correlation of allelic biases of DNA binding detected from ChIP–seq experiments in HepG2 cells and those predicted by PBS (blue) and ΔPWM (yellow). PCC and *P* value calculated by two-sided *t*-test are shown. The allelic binding ratio is computed as $\log_{10}$(odds ratio) over input. One hundred and ninety-three transcription factor–SNP pairs involving 147 unique SNPs and 6 transcription factors (ATF2, FOXA2, HLF, MAFG, YBX1 and FOXA1) are shown. **e**, Comparison of PBS and ΔPWM in predicting the effect of SNPs on differential enhancer activity. SNPs were categorized into five quantiles according to their effect size on transcription factor binding on the basis of PBS (blue) or ΔPWM (yellow); quantile 1 includes the SNPs that have the largest effect size on differential transcription factor binding.

cycles and control variations among experiments. We estimated the significance of OBS for each pair of oligonucleotide and transcription factor, finding that 89,171 oligonucleotides displayed significant binding to at least one transcription factor ($P < 0.05$ by Monte Carlo randomization, $n = 250{,}000$) (Extended Data Fig. 2c, d). To describe the differential transcription factor binding between the reference and alternative alleles of each SNP, we next defined the preferential binding score (PBS) by computing the difference between OBSs of two alleles to each transcription factor (Fig. 1b, Extended Data Fig. 2e). A total of 11,079 SNPs exhibited significantly differential binding to at least one transcription factor (Monte Carlo randomization $P < 0.01$, $n = 250{,}000$) (Fig. 1b, Supplementary Table 3, Supplementary File 2). We termed these SNPs preferential binding SNPs (pbSNPs). Among the 270 transcription factors that passed quality control in SNP-SELEX, 250 exhibited preferential binding to at least one pbSNP. Overall, each transcription factor bound differentially to a median of 53 pbSNPs (Fig. 1c), and each pbSNP showed differential binding to one transcription factor on average (Fig. 1d).

Several lines of evidence support the reliability of SNP-SELEX results. First, both OBS and PBS were highly reproducible between independent replicative experiments (Extended Data Fig. 3a–c). Second, PBS and OBS of the full-length transcription factors matched very well with those of the corresponding DNA-binding domains (DBDs), to a similar degree between replicates (Extended Data Fig. 3d, e), consistent with a previous notion that the DBD adequately determined a transcription factor's DNA sequence specificity[4]. Third, the correlation between different transcription factors within the same structural family was significantly higher than that between randomly selected pairs of transcription factors[4] (Wilcoxon test, $P < 2 \times 10^{-16}$) (Extended Data Fig. 3d, e). The majority of transcription factors from the same family—except those from the C2H2 zinc finger family—tended to share similar pbSNPs, as previously noted[4] (Fig. 1e). Overall, our results suggest that SNP-SELEX is a cost-effective and highly reproducible platform for analysis of differential transcription factor binding to noncoding variants in vitro.

The performance of position weight matrices (PWMs) in predicting differential binding of a transcription factor to sequence variants has not been systematically evaluated. To address this, we first derived PWMs for transcription factors using HT-SELEX experimental data involving 40-mer random sequences[3]. Then, we compared PBSs of transcription factors to ΔPWM scores (differential PWM scores between alleles) for 255 out of the 549 transcription factors characterized to date. These PWMs were originally derived from HT-SELEX using a multinomial algorithm[3]. To avoid systematic bias caused by the choice of motif-generating algorithm, we also derived independent PWMs from the same set of data, but using the BEESEM[5] algorithm, which relies on binding energy models of protein–DNA interactions. We found that PBS and ΔPWM scores for the 70,402 SNPs with both types of estimation available are moderately correlated (Pearson's $r = 0.534$) (Fig. 2a). The ΔPWM-based prediction and SNP-SELEX experimental analysis agreed in more than 80% of cases (339,961 transcription factor–SNP pairs). However, in a substantial fraction of cases (17.85%), ΔPWM predictions did not match SNP-SELEX results (73,876 transcription factor–SNP pairs) (Extended Data Fig. 4a). These discordant cases frequently corresponded with low-affinity transcription factor–DNA binding events (Extended Data Fig. 4b). Notably, some common genetic diseases are believed to be attributable to a large number of common SNPs with small effect sizes; it is thus crucial to comprehensively characterize

these variants. In line with the role of sequence variations at weak binding sites in common diseases, suboptimal transcription factor binding sites are of particular importance in regulation of developmental genes[6].

When ΔPWM predictions of individual transcription factors were tested against pbSNPs to predict differential transcription factor binding, ΔPWM of many transcription factors—for example, IRF3—performed poorly (Fig. 2b). Out of the 129 transcription factors with more than 40 pbSNPs that have sufficient statistic power for evaluation, ΔPWM-based prediction of only 24 transcription factors achieved a satisfactory performance (area under the precision–recall curve (AUPRC) ≥ 0.75) (Fig. 2c). The performance of different PWM models varied markedly among different structural families of transcription factors. For example, PWMs of TFAP family transcription factors generally had outstanding predictive power, whereas E2F family transcription factors showed poor performance, despite the similar information content of their PWM models (Extended Data Fig. 4c, d).

When ΔPWM predictions differed from the PBSs derived from SNP-SELEX experiments, we found that the SNP-SELEX experiments could more accurately predict the effects of SNPs on transcription factor binding in vivo. First, we examined 12 publicly available or in-house chromatin immunoprecipitation followed by high-throughput sequencing (ChIP–seq) datasets corresponding to 10 transcription factors in either HepG2 hepatocytes or GM12878 lymphoblast cells[7] (Supplementary Table 4). Among the 86 pbSNPs, the ratios between allelic ChIP–seq signals in HepG2 were significantly correlated with PBS for that factor (*t*-test $P = 5.17 \times 10^{-5}$, Pearson's $r = 0.409$) whereas the correlation with ΔPWM was not significant (*t*-test $P = 0.792$, Pearson's $r = 0.027$) (Fig. 2d). The same trend was observed in ChIP–seq from GM12878 cells (Extended Data Fig. 4e). Second, using a high-throughput reporter assay STARR-seq[8], we examined the enhancer activity of 2,246 pbSNPs and 1,697 non-pbSNPs-containing genomic fragments in HepG2 and HEK 293T human embryonic kidney cells (Extended Data Fig. 5a, b, Supplementary Table 5), and found that 424 (in HepG2 cells) and 527 (in HEK 293T cells) pbSNP-containing genomic fragments showed significant enhancer activity (Extended Data Fig. 5c; empirical false discovery rate (FDR) <0.05); 200 of these SNPs displayed allelic bias on enhancer activity in HepG2 cells and 206 displayed allelic bias on enhancer activity in HEK 293T cells (FDR <0.05)—we designated these preferentially active SNPs (paSNPs) (Supplementary Table 6). pbSNPs were more likely to be associated with allelic enhancer activity than non-pbSNPs (Fisher's exact test $P = 0.027$, odds ratio = 1.57) (Extended Data Fig. 5d). Notably, the more allelic bias there was for a paSNP, the higher the PBS for the pbSNP (Fig. 2e). By contrast, significantly fewer paSNPs were identified by ΔPWM. SNPs predicted by ΔPWM to be differentially bound by transcription factors were not associated with the degrees of differential enhancer activities (Fisher's exact test $P = 0.465$, odds ratio = 1.23) (Extended Data Fig. 5e). These results strongly suggest that SNP-SELEX results are more reliable than ΔPWM scores for predicting the effects of noncoding variants on transcription factor binding in vivo.

The number of SNPs tested in SNP-SELEX remains much smaller than the number of known noncoding SNPs in the human genome[9]. Aiming to study differential DNA binding by transcription factors to any genetic variants, we used the deltaSVM[10] framework, which used changes in gapped *k*-mers support vector machine (gkm-SVM) scores to quantify effects of variants. We derived deltaSVM models for 533 transcription factors with previously published HT-SELEX data[3] (Extended Data Fig. 6a). The deltaSVM scores computed between the reference and alternative allele-containing genomic fragments were highly correlated with PBS values (Fig. 3a)—notably better than the correlation between PBS and ΔPWM scores (Fig. 2a). We then used pbSNPs from SNP-SELEX as a gold standard for comparing the performance between deltaSVM and ΔPWM in predicting effects of SNPs on transcription factor binding. To ensure sufficient statistical power, we included only

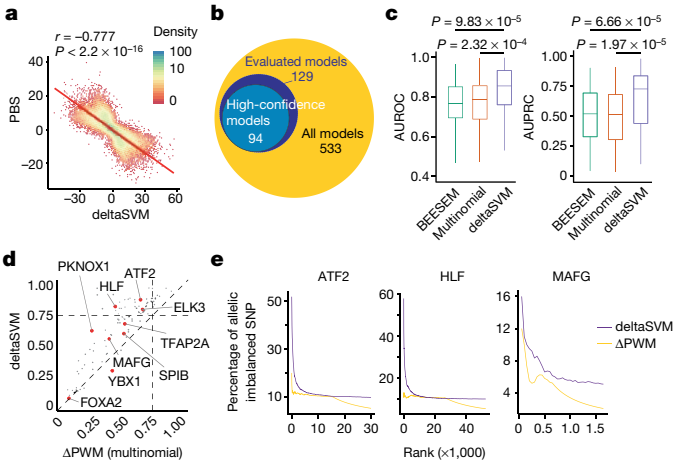


Fig. 3 | DeltaSVM models outperform ΔPWM in predicting differential transcription factor binding to noncoding variants in vitro and in vivo. **a**, Correlation between PBS and deltaSVM scores. The red line denotes a linear regression of the two scores. PCC and *P* value calculated by two-sided *t*-test are shown. Each dot represents one transcription factor–SNP pair. **b**, Venn diagram showing the number of transcription factors for which there are differential DNA binding models or experimental data defined by deltaSVM. **c**, Box plots comparing the performance of deltaSVM, PWM originally derived in HT-SELEX (multinomial), or derived with the BEESEM algorithm in predicting pbSNPs in the novel SNP-SELEX batch for 87 transcription factors. Two statistical evaluation methods were used, including area under the receiver operating curve (AUROC) (left) and AUPRC (right). *P* values by two-sided Wilcoxon test are shown. In box plot, the horizontal line shows median, hinges represent 25th and 75th percentile, and whiskers extend to the most extreme value no further than 1.5× the interquartile range. **d**, Comparison of performance of deltaSVM (*y*-axis) and multinomial-generated ΔPWM (*x*-axis) in predicting pbSNPs identified in the novel batch of SNP-SELEX. Both axes show AUPRC values. **e**, Elbow plots show that the allelic SNPs top-ranked by deltaSVM models were mostly allelic transcription factor-binding SNPs in vivo identified by ChIP–seq in HepG2 cells (purple). For allelic SNPs predicted by ΔPWM, only a very small fraction showed allelic binding in vivo (yellow).

129 transcription factors with 40 or more pbSNPs (Fig. 3b). In fivefold cross validation, deltaSVM substantially outperformed the ΔPWM models developed with either multinomial[3] or BEESEM[5] algorithms (Extended Data Fig. 6b–d; Supplementary Table 7).

To further evaluate the performance of deltaSVM models against ΔPWM, we conducted an independent set of SNP-SELEX experiments using 61,020 previously uncharacterized SNPs and 487 transcription factors (Extended Data Fig. 6e; Methods for the SNP selection). We identified an additional 21,299 pbSNPs ($P < 0.01$ by Monte Carlo randomization, $n = 250{,}000$) (Supplementary Table 8). When this list of pbSNPs was used as the gold standard, deltaSVM models continued to outperform ΔPWM models (Fig. 3c), with a median AUPRC value for deltaSVM of 0.728, compared to 0.513 and 0.521 for multinomial- and BEESEM-derived ΔPWM models, respectively (Fig. 3c, d, Extended Data Fig. 6f, Supplementary Table 7).

We reasoned that the poor performance of many ΔPWMs was probably because they did not take into account dinucleotide interdependency in transcription factor–DNA interactions and the influence of flanking DNA sequences[11,12]. Previous studies have shown that dinucleotide interdependency exists for some transcription factor dimers[4]. For example, according to the PWM model, the SNP rs79124498—located within a binding site of HLF, a bZIP family transcription factor that binds DNA as homodimers—would have little effect on HLF binding. However, SNP-SELEX indicated that the G allele bound more strongly than the T allele to HLF. This could be caused by the dinucleotide interdependency between position 2 (the SNP position) and position 10 in the binding site (Fisher's exact test $P < 2.2 \times 10^{-16}$, odds ratio = 3.34)

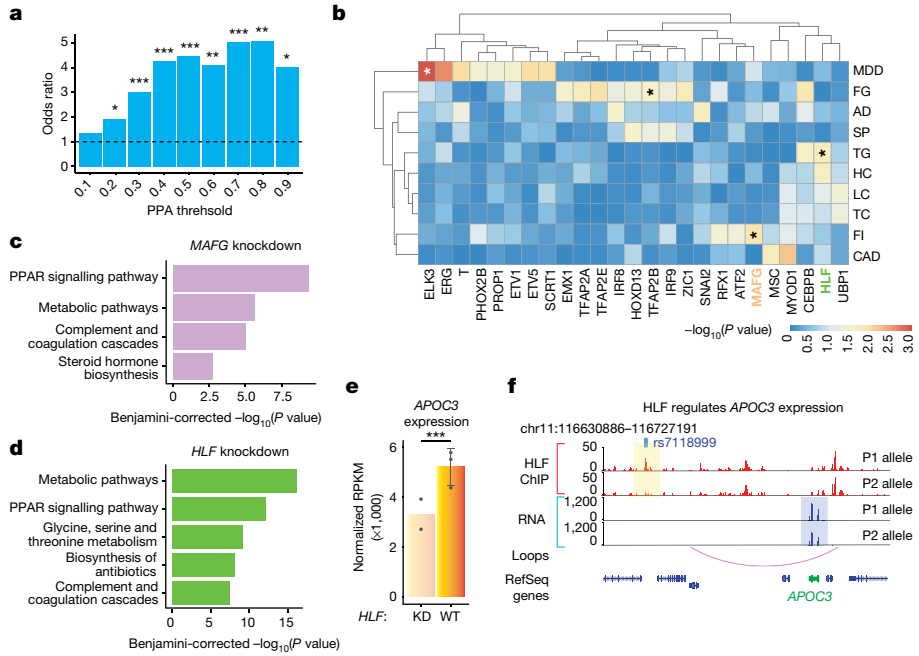


Fig. 4 | deltaSVM models predict TFs probably involved in complex traits and diseases. a, Enrichment of pbSNPs in reported candidate T2D-causing SNPs[12]. The level of association is categorized according to the posterior probability of association threshold. *P* values for the enrichment by Fisher's exact test are indicated; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. **b**, Heat map shows the significance of enrichment of SNPs with differential binding to transcription factors among traits- or disease-associated SNPs. Asterisks indicate transcription factor–trait pairs mentioned in the current study. MDD, major depression disorder; FG, fasting glucose; AD, Alzheimer's disease; SP, schizophrenia; TG, triglycerides; HC, high density lipoprotein cholesterol; LC, low density lipoprotein cholesterol; TC, total cholesterol; FI, fasting insulin; CAD, coronary artery disease. **c**, **d**, Enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways significantly affected by knockdown of *MAFG* (**c**) or *HLF* (**d**) in HepG2 cells. The *y*-axis shows –log Benjamini–Hochberg-corrected *P* values. **e**, Normalized *APOC3* gene expression in *HLF*-knockdown (KD) and wild-type (WT) HepG2 cells. $***P = 6.67 \times 10^{-5}$, as computed by DESeq2. Expression values are presented as mean ± s.d. reads per kilobase of transcript, per million mapped reads (RPKM). Three biological replicates were performed. **f**, Genome browser display showing that differential HLF binding to rs7118999 is linked to allelic gene expression of *APOC3*, which is predicted to be targeted by the SNP locus on the basis of a chromatin loop in HepG2 cells (purple curve). Top two tracks (red) show HLF binding according to ChIP–seq, with two alleles separated by haplotypes. Bottom two tracks (blue) show allelic expression of nearby genes. Stronger binding of HLF in P1 alleles corresponds to higher expression of *APOC3* on the same allele.

(Extended Data Fig. 6g). Such dinucleotide interdependency information is not embedded in regular PWM models, but could be captured by deltaSVM models.

We also found that ΔPWM performed poorly for SNPs located in low-affinity binding sites of transcription factors (Extended Data Fig. 4b). However, this limitation could be overcome by using deltaSVM. When we categorized SNPs into five quantiles on the basis of their binding affinities as measured by OBS, and assessed the performance of ΔPWM and deltaSVM in predicting their allelic binding by fivefold cross-validation or using the novel batch of SNP-SELEX experimental results (Extended Data Fig. 7), deltaSVM outperformed ΔPWM in all quantiles, particularly in the lower quantiles corresponding to weak transcription factor binding sites.

The above results demonstrate that deltaSVM models built from HT-SELEX datasets are superior to ΔPWM for predicting the effect of SNPs in transcription factor binding. We subsequently focused on the 94 high-confidence deltaSVM models (AUPRC >0.75) for genome-wide prediction and analysis (Fig. 3b, Supplementary File 3). These deltaSVM models outperformed ΔPWM scores in predicting differential transcription factor binding to SNPs (Extended Data Fig. 8a). In analysis of the allelic transcription factor–DNA binding in HepG2 cells from ChIP–seq datasets, deltaSVM models accounted for twice as many SNPs with allelic DNA binding as ΔPWM (Extended Data Fig. 8b). If we ranked the SNPs on the basis of their deltaSVM scores, the top-ranked SNPs recovered the most allelic-imbalanced SNPs identified by ChIP–seq. By contrast, ΔPWM predictions did not show such a trend (Fig. 3e). Similarly, deltaSVM models could explain a significantly higher percentage of allelic DNA binding for ATF2, PKNOX1 and NR2F1 in GM12878 cells than ΔPWM scores (Extended Data Fig. 8c, d).

If noncoding variants contribute to diseases by affecting transcription factor binding to *cis*-regulatory sequences, the causal SNPs should be enriched for pbSNPs discovered in the current study. Indeed, we found that pbSNPs were highly enriched in the set of likely causal SNPs reported for T2D from two independent studies[13,14] (Fig. 4a, Extended Data Fig. 9a). Notably, the enrichment of pbSNPs was even stronger in the set of variants with higher likelihood of causatively increasing the risk of the disease. When we performed similar analysis on the same dataset but using SNPs with allelic transcription factor binding predicted by ΔPWM scores, the likely causal SNPs were no longer enriched (Extended Data Fig. 9b), supporting the utility of pbSNPs in interpretation of the noncoding risk variants.

One such example is the SNP rs7578326, located in a candidate enhancer bearing histone H3 lysine 27 acetylation (H3K27ac). The SNP was found to affect binding of the liver-specific transcription factor CEBPB in our analysis (Extended Data Fig. 9c). This candidate enhancer is linked to the *IRS1* gene, located approximately 500 kb downstream, by long-range chromatin interactions in HepG2 cells. To confirm its regulatory role, we used CRISPR interference to silence this candidate enhancer in HepG2 and HEK 293T cells. Upon silencing, significant reduction of IRS1 was observed in HepG2 cells, which expressed a high level of CEBPB protein, but not in HEK 293T cells, where the expression of CEBPB was much lower (Extended Data Fig. 9d). This result, together with the findings that rs7578326 is an expression quantitative trait locus (eQTL) of IRS1 in liver and adipose tissues[15] (Extended Data Fig. 9e),

and is associated with fasting insulin levels and insulin sensitivity[16,17], suggest that the SNP is probably causal in T2D pathogenesis, acting through regulation of insulin sensitivity[18,19].

To further determine whether binding of any specific transcription factors was disproportionately affected by noncoding variants associated with T2D-related metabolic traits and other human diseases, we focused on the 94 high-confidence deltaSVM models and performed stratified linkage disequilibrium score regression (S-LDSC) to test the enrichment of SNPs affecting transcription factor binding in the set of variants identified from GWAS of these traits (Supplementary Table 9). As expected, transcription factors previously known to be associated with some metabolic traits showed strong enrichment among transcription factors that could be affected by the risk SNPs and those in linkage disequilibrium[20,21] (Fig. 4b). Notably, most of the trait-associating transcription factors, particularly those known key factors discussed above, would not be recovered if we performed enrichment analysis merely for the presence of trait-associated SNPs in transcription factor binding sites (Extended Data Fig. 10a).

We identified candidate transcription factors associated with additional human traits and diseases. For instance, our analysis predicted that MAFG has a role in regulating fasting insulin levels (Fig. 4b), an indicator of insulin sensitivity[16]. To validate this prediction, we examined the genes differentially expressed in HepG2 cells following knockdown of *MAFG* expression, and found that genes in the peroxisome proliferator activated receptor (PPAR) signalling pathway were most affected (Fig. 4c, Extended Data Fig. 10b, c). The PPAR signalling pathway is key to regulation of the insulin signalling cascade[22].

Our analysis also predicted that HLF could be associated with levels of circulating triglycerides (Fig. 4b). Supporting this prediction, knockdown of HLF in HepG2 cells affected many genes involved in metabolic and PPAR signalling pathways (Fig. 4d, Extended Data Fig. 10d, e), which are important regulators of blood triglyceride levels[23]. The gene for apolipoprotein C3 (*APOC3*)—a known regulator of triglyceride-rich lipoprotein metabolism[24,25]—is among those most affected after knockdown of *HLF* (Fig. 4e). Using ChIP–seq, we showed that HLF was bound to a putative enhancer containing the SNP rs7118999, located approximately 70 kb upstream of the *APOC3* promoter, and looped back to the *APOC3* promoter (Fig. 4f). Allelic binding of HLF to the heterozygous rs7118999 was accompanied by allelic expression of *APOC3* in HepG2 cells, where stronger binding of HLF correlated with higher expression of *APOC3* in *cis* (Fig. 4f). These combined results suggest that HLF can regulate *APOC3* expression and, in turn, the abundance of triglyceride-rich lipoprotein (VLDL) in blood, a major risk factor for coronary artery disease[25,26]. APOC3 is a drug target for reducing the risk of coronary artery disease in clinical studies[27]; our analysis therefore raises the possibility of targeting HLF for therapeutic intervention for coronary artery disease.

The SNP-SELEX study design is currently limited to a small fraction of the SNPs in the human genome and is slightly skewed towards T2D-associated risk loci. Future SNP-SELEX experiments designed to cover a broader range of SNPs may further improve the performance of deltaSVM models. Additionally, the list of transcription factors with validated deltaSVM models is expected to grow with increasing availability of recombinant transcription factor proteins and combinations of heterodimeric transcription factors[28] for SNP-SELEX experiments. We propose that this high-throughput approach and the resources described here will provide insights into the roles of noncoding risk variants in human diseases and uncover new therapeutic targets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03211-0.

1. Buniello, A. et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47** (D1), D1005–D1012 (2019).
2. Weirauch, M. T. et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31**, 126–134 (2013).
3. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
4. Jolma, A. et al. DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
5. Ruan, S., Swamidass, S. J. & Stormo, G. D. BEESEM: estimation of binding energy models using HT-SELEX data. *Bioinformatics* **33**, 2288–2295 (2017). https://doi.org/10.1093/bioinformatics/btx191.
6. Farley, E. K. et al. Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015).
7. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
8. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
9. Altshuler, D. M. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
10. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
11. Rohs, R. et al. The role of DNA shape in protein–DNA recognition. *Nature* **461**, 1248–1253 (2009).
12. Morgunova, E. et al. Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. *eLife* **7**, e32963 (2018).
13. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
14. Greenwald, W. W. et al. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nat. Commun.* **10**, 2078 (2019).
15. Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
16. Olefsky, J., Farquhar, J. W. & Reaven, G. Relationship between fasting plasma insulin level and resistance to insulin-mediated glucose uptake in normal and diabetic subjects. *Diabetes* **22**, 507–513 (1973).
17. Soyal, S. M. et al. Associations of haplotypes upstream of *IRS1* with insulin resistance, type 2 diabetes, dyslipidemia, preclinical atherosclerosis, and skeletal muscle *LOC646736* mRNA levels. *J. Diabetes Res.* **2015**, 405371 (2015).
18. Manning, A. K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
19. Scott, R. A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
20. Nordquist, N. et al. The transcription factor TFAP2B is associated with insulin resistance and adiposity in healthy adolescents. *Obesity* **17**, 1762–1767 (2009).
21. Apazoglou, K. et al. Antidepressive effects of targeting ELK-1 signal transduction. *Nat. Med.* **24**, 591–597 (2018).
22. Leonardini, A., Laviola, L., Perrini, S., Natalicchio, A. & Giorgino, F. Cross-talk between PPARγ and insulin signaling and modulation of insulin sensitivity. *PPAR Res.* **2009**, 818945 (2009).
23. Fruchart, J. C., Duriez, P. & Staels, B. Peroxisome proliferator-activated receptor-alpha activators regulate genes governing lipoprotein metabolism, vascular inflammation and atherosclerosis. *Curr. Opin. Lipidol.* **10**, 245–257 (1999).
24. Shachter, N. S. Apolipoproteins C-I and C-III as important modulators of lipoprotein metabolism. *Curr. Opin. Lipidol.* **12**, 297–304 (2001).
25. Crosby, J. et al. Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease. *N. Engl. J. Med.* **371**, 22–31 (2014).
26. Gotto, A. M., Jr. Triglyceride as a risk factor for coronary artery disease. *Am. J. Cardiol.* **82** (9A), 22Q–25Q (1998).
27. Khetarpal, S. A., Qamar, A., Millar, J. S. & Rader, D. J. Targeting ApoC-III to reduce coronary disease risk. *Curr. Atheroscler. Rep.* **18**, 54 (2016).
28. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

SNP selection

In total, 110 leading SNPs were selected from previous T2D GWAS[29,30]. Common SNPs (minor allele frequency >1%) within 500 kb of the 110 leading SNPs were extracted from the 1000 Genome Project from all available populations, resulting in 379,895 unique SNPs. From these SNPs, 6,724 SNPs were selected in linkage disequilibrium with leading SNPs in East Asian and Caucasian populations ($r^2 \geq 0.8$) from 1000 Genome Project Pilot 1[31], and 89,162 SNPs were selected on the basis of their distance (≤500 kb) to the accessible chromatin regions in ENCODE DHS sites[32] or FANTOME 5[33] permissive enhancers for all cell and tissue types. Altogether, 95,886 SNPs were included in the current study (Supplementary Table 1).

Experimental procedure

Oligonucleotide design was adapted to an Illumina TruSeq dual-index system (Extended Data Fig. 1a) and synthesized by CustomArray. The oligonucleotides were amplified using 20 cycles of PCR and sequenced with Illumina HiSeq 2500 to verify the identities. cDNAs encoding transcription factor proteins were cloned to pET20a plasmids[3] and expressed using Rosetta (DE3) pLysS strains (amino acid sequences of the transcription factors are presented in Supplementary Table 2) as previously described[4].

The SELEX experiments were performed essentially as previously described[4]. In each SNP-SELEX experiment, the double-stranded DNA library was incubated with a recombinant transcription factor protein. The bound DNA molecules were eluted, PCR-amplified and sequenced, while an aliquot was used as input for the next round of the SNP-SELEX experiment. The binding–washing–sequencing cycle was repeated a total of six times. Because the binding reaction is competitive and the washing steps are sufficiently long, the read counts for each 40-mer sequence can be assumed to be proportional to its binding affinity to the assayed transcription factors.

In brief, the 6× His-tagged transcription factor proteins were expressed in *Escherichia coli* and immobilized on Ni-sepharose beads (GE Healthcare, 17-5318-01) in Promega binding buffer (10 mM Tris pH7.5, 50 mM NaCl, 1 mM $MgCl_2$, 4% glycerol, 0.5mM EDTA, 5 μg ml^{-1} poly(deoxyinosinic-deoxycytidylic) acid. Oligonucleotides from input or the previous HT-SELEX cycles were added into the protein beads mixture and incubated at ambient temperature for 30 min. After binding, the beads were consecutively washed for 12 times with the Promega binding buffer. After the final wash, TE (10mM Tris pH 8.0, 1mM EDTA) was used to resuspend the beads and for PCR amplification. The PCR products from each HT-SELEX cycle were purified (Qiagen, 28004) and sequenced with Illumina HiSeq 2500. An aliquot of the PCR products was used for the next cycle of SELEX.

SNP-SELEX data analysis

Sequencing data from each SELEX cycle was aligned to the oligonucleotide library using BWA[34]. Several filters were applied to aligned reads after alignments: (1) reads of low quality, containing ambiguous bases, unaligned to reference and aligned outside of the oligonucleotide boundaries were filtered out and experiments with less than 10,000 reads were excluded from further analysis; (2) to control for PCR-duplication bias, the frequency of all PCR duplication bias control (PDC) sequences (256 combinations) of each cycle were compared to the input library (cycle 0) using a linear regression model. PDCs whose difference between expected and observed frequency exceeded 30% of the observed values were considered biased and all reads containing the biased PDC were removed.

De novo motif discovery was then conducted using the cycle six reads with the Homer toolset[35] (Supplementary File 1). Motifs were then compared to JASPAR 2016 non-redundant vertebrates motifs[36] and SELEX models to examine quality of the experiments[3]. Only SNP-SELEX experiments whose motif models match either its transcription factor or transcription factors of same structural family[4,37] were retained for further analysis (Supplementary Table 2). The frequencies of reads supporting each SNP oligonucleotide and its alleles were obtained from the remaining dataset. After quality control, 360 experiments passed this quality control step. In total, we obtained in total 828,455,040 measurement of transcription factor–DNA interactions for 95,886 oligonucleotides with six cycles and four possible nucleotides per oligonucleotides (360 experiments × 95,886 oligonucleotides × 4 possible bases × 6 cycles = 828,455,040 measurements).

Aiming to quantify the transcription factor binding to genomic oligonucleotides, the OBS was defined as the AUC of the logarithmic odds ratio curve along the HT-SELEX cycles to estimate the relative binding affinity of the 40-mer sequence to the transcription factor (Extended Data Fig. 2b). We first estimated odds ratio of observing an oligonucleotide at cycle i as $OR_{oligo,i}$, in which $P_{oligo,i}$ is the proportion of oligonucleotide at cycle i and $OR_{oligo,i}$ is the odds of observing oligonucleotide at cycle i in relation to all other oligonucleotides (equation (1)). We then compared odds ratios for each oligonucleotide at each cycle to cycle 0, namely the input library, to calculate the relative odds ratio at each cycle as $LOR_{oligo,i}$ (equation (2)). OBS was then computed as AUC of $LOR_{oligo,i}$ over six HT-SELEX cycles (equation (3)).

$$OR_{oligo,i} = \frac{P_{oligo,i}}{1 - P_{oligo,i}} \quad (1)$$

$$LOR_{oligo,i} = \log_{10}\left(\frac{OR_{oligo,i}}{OR_{oligo,0}}\right) \quad (2)$$

$$\begin{aligned} OBS &= \int LOR_{oligo,i}\,di \\ &= \frac{1}{2} \times \sum (LOR_{oligo,i} + LOR_{oligo,i+1}),\ \text{for } i \\ &= 0 \text{ to } 5 \end{aligned} \quad (3)$$

Likewise, PBS was introduced to quantify allele preferential binding for each SNP as difference of OBS between reference and alternative alleles in terms of logarithmic odds ratio along HT-SELEX cycles to estimate the difference of relative binding affinities between the two alleles to the transcription factor (Extended Data Fig. 2e). We first calculated odds ratios for each allele at each cycle comparing to cycle 0 as $LOR_{allele\,a,i}$ in a similar manner for oligonucleotides (equations (4) and (5)). We then compared two alleles for each SNP to calculate relative logarithmic odds ratio as $\Delta LOR_{snp,i}$ (equation (6)). PBS was then computed as AUC of ΔLOR_c over six HT-SELEX cycles (equation (7)).

$$OR_{allele\ a,i} = \frac{P_{allele\ a,i}}{1 - P_{allele\ a,i}} \quad (4)$$

$$LOR_{allele\ a,i} = \log_{10}\left(\frac{OR_{allele\ a,i}}{OR_{allele\ a,0}}\right) \quad (5)$$

$$\Delta LOR_{snp,i} = LOR_{allele\ reference,i} - LOR_{allele\ reference,i} \quad (6)$$

$$PBS = \int \Delta LOR_{snp,i}\,di = \frac{1}{2} \times \sum (\Delta LOR_{snp,i} + \Delta LOR_{snp,i+1}),\ \text{for } i = 0 \text{ to } 5 \quad (7)$$

The statistical significance of both PBS and OBS in each experiment was measured by Monte Carlo randomization, in which the

oligonucleotide and allele read counts were shuffled within each cycle and the scores were recomputed 250,000 times. Oligonucleotides were considered significantly bound to the transcription factor for OBS *P* value <0.05. oligonucleotides were considered significantly preferentially bound for SNPs for PBS *P* value <0.01 and OBS *P* value <0.05.

Novel batch of SNP-SELEX experiments

To generate a completely independent dataset to benchmark deltaSVM models, we performed additional novel batch of SNP-SELEX experiments. Variants tested in the novel batch included 32,289 SNPs within known T2D loci (lead variants and variants in linkage disequilibrium with $r^2 \geq 0.6$ in EUR and non EUR, and credible variants from fine mapping studies), 58,184 SNPs within islet enhancers (defined using assay for transposase-accessible chromatin using sequencing (ATAC-seq) and H3K27ac ChIP–seq data from human islets), and 8,000 negative control SNPs randomly chosen from the genome. The SNP-SELEX experiments were performed exactly as the first batch, with only four cycles.

The following filters were applied before calculating preferential binding: (1) each of the 768 experiments were done in replicate. If the replicates did not correlate ($r < 0.5$) with each other (the fraction of reads aligning to each oligonucleotide) and did not show motif enrichment, the experiment was excluded; (2) for each experiment, only variants covered by at least 8 read pairs for SNPs, or 4 reads pairs for indels, in all five cycles (0–4) were retained; (3) for each experiment, only variants with at least 2 read pairs in the input for both the reference and alternate alleles and composing 5% of the total reads in the pool were retained; (4) experiments with less than 25 variants remaining after the above two filtering steps were excluded.

PBS values were calculated as described in the previous section, and *P* values were computed using 250,000 Monte Carlo randomizations of allele read counts within cycles 1 to 4, while keeping reads at cycle 0 fixed. Results of the two replicates for each experiment were combined using meta-analysis of *P* values and average of effect sizes. Experimental replicates of the same transcription factor protein were meta-analysed to obtain a unique value for each transcription factor.

In summary, 1,048,486 transcription factor–SNP pairs including 66,329 SNPs (61,020 SNPs different from the first batch) and 487 transcription factors were tested. Out of these, there were 23,262 pbSNPs (*P* value <0.01).

STARR-seq experiments

Design of oligonucleotides. To directly evaluate the effect of pbSNP on enhancer activities, STARR-seq[8] was conducted with HEK 293T and HepG2 cells. In total, we tested 11,961 genomic sequences containing 2,246 pbSNPs and 1,697 non-pbSNPs from SNPs either located in the human islet ATAC-seq peaks[14] or displayed significant OBS scores in SNP-SELEX. In addition, we included 37 true positive controls which are known enhancers and 2,998 negative controls that correspond to random yeast open reading frames (ORFs) sequences (Supplementary Table 5).

Oligonucleotide design was adapted from the previously published STARR-seq work[8] (Extended Data Fig. 5a) and synthesized by Agilent. In brief, each oligonucleotide contains 190 bp of genomic sequence enclosing the SNP and 20 bp constant flanking sequences (upstream: 5′- ACACGACGCTCTTCCGATCT; downstream: AGATCGGAAGAGCACACGTC-3′) on both ends, which were used for amplification and cloning. The generic PCR primers including Illumina Truseq adaptor sequences and different indexes were used to amplify the oligonucleotide pool and cloned into the human STARR-seq plasmid (a gift from the Stark laboratory, Research Institute of Molecular Pathology (IMP)). PCR amplification from the plasmids was performed and sequenced with 2 × 100 paired-end cycles with Illumina HiSeq 4000 sequencer as input control.

Cell culture and transfection. The plasmid pool was transfected into HEK 293T or HepG2 cell lines using Fugene HD. The HEK 293T (ATCC, CRL-3216) and HepG2 (ATCC, HB-8065) cells were cultured under normal conditions with 5% CO_2 at 37 °C. Fugene HD (Promega, E2311) was used for plasmid transfection. Specifically, 2 μg of STARR-seq plasmids were mixed with 5 μl of transfection reagents for transfection into 300,000 cells cultured in a single well of a 6-well plate.

mRNA extraction and sequencing. Forty-eight hours after transfection, total RNA was extracted with RNeasy kit (Qiagen, 74104) and mRNA was enriched with poly(dT)$_{25}$ Dynabeads (Invitrogen, 61002). First-strand cDNA was synthesized using a specific primer (5′-CAAACTCATCAATGTATCTTATCATG) with high High-Capacity cDNA Reverse Transcription kit (ThermoFisher Scientific, 4368814). Nested PCR was used to amplify the SNP-specific fragments from cDNA, first using two reporter-specific PCR primers (5′-GGGCCAGCTGTTGGGGTGTCCAC & 5′-CTTATCATGTCTGCTCGAAGC) and then generic primers used in HT-SELEX. DNA was purified with AMPure beads and sequenced with 2 × 100 paired-end cycles with illumina HiSeq 2500 sequencer. In total, three biological replicates were performed with two technical replicates each for both HepG2 and HEK 293T cells.

STARR-seq data analysis. STARR-seq reads were aligned to the oligonucleotide libraries using BWA[38] with default parameters. Read counts for each oligonucleotides were then counted. Counts for technical replicates were merged. Oligonucleotides covered by more than 25 reads in the input library (the synthesized oligonucleotide pool) and more than five reads in at least three libraries were kept for downstream analysis.

We first identified oligonucleotide that were enriched compared with the input library. Enriched oligonucleotides were determined by a negative binomial regression from the R package edgeR[39]. Common biological dispersion was estimated using only yeast oligonucleotides where no real variation is expected. The resulting *P* values were adjusted by the Benjamini–Hochberg procedure, and the significance cut-off for enriched oligonucleotides was set to limit the rate of enriched yeast oligonucleotides to 5%.

We then focused on the SNPs for which at least one allele was significantly enriched, and calculated the difference of log fold-change activity between the two alleles using a paired *t*-test from R package limma[40], shrinking the variance with an empirical Bayesian method. The *P* values were adjusted by Benjamini–Hochberg procedure and SNPs were considered significant with adjusted *P* <0.01.

In situ Hi-C experiments to predict target genes of non-coding SNPs

The in situ Hi-C was performed according to a previously described protocol[41] with slight modifications. In brief, the HepG2 cells were trypsinized and washed with PBS. The chromatin was cross-linked with 1% formaldehyde (Sigma) at ambient temperature for 10 min and quenched with 125 mM glycine for 5 min. PBS washed tissue was homogenized with a loose-fitting Doune with 30 strokes before centrifugation to isolate the nuclei.

Nuclei were isolated and directly applied for digestion using 4 cutter restriction enzyme MboI (NEB) at 37 °C overnight. The single-strand overhang was filled with biotinylated-14-ATP (Life Technologies) using Klenow DNA polymerase (NEB). Different from traditional Hi-C, the ligation was performed when the nuclear membrane was still intact. DNA was ligated for 4 h at 16 °C using T4 ligase (NEB). Protein was degraded by proteinase K (NEB) treatment at 55 °C for 30 min. The crosslinking was reversed with 500 mM of NaCl and heated at 68 °C overnight. DNA was purified and sonicated to 300–700-bp small fragments. Biotinylated DNA was selected with Dynabeads My One T1 Streptavidin beads (Life Technologies). A sequencing library was prepared on beads and intensive washes were performed between reactions. Libraries

were checked with Agilent TapeStation and quantified using Qubit (Life Technologies). Libraries were sequenced with Illumina HiSeq 4000 with 100 cycles of paired-end reads.

Hi-C data was processed as previously described[42]. In brief, each end of read pairs was aligned separately using BWA MEM to the hg19 reference genome with default parameters. Chimeric read ends were further processed to keep only the five-prime alignment. Read ends with low mapping quality (mapq <10) were removed, and remaining read ends were paired using custom scripts. PCR duplicates were removed using the Picard tool (http://broadinstitute.github.io/picard). Resulting read alignments were stored as bam files using samtools. Aligned reads were further transformed to the juicer format and processed into hic format using juicebox tool[43]. Chromatin loops were called using HiCCUPS with default parameters.

To assign potential target genes for SNPs, two approaches were taken: (1) SNPs within 2 kb upstream region of a TSS were assigned to the TSS; (2) SNPs overlapping one anchor of chromatin loops (within a 25-kb window) were assigned to the TSS overlapping the other anchor (within a 25-kb window). Similar approaches were used to connect transcription factor binding sites to target genes.

Determination of allele imbalance of transctiption factor binding from ChIP–seq data

The ChIP–seq experiment was carried out using an established protocol[44]. In brief, the cells were crossed linked with 1% formaldehyde at ambient temperature for 10 min. The reaction was quenched by 125 mM glycine for 5 min at room temperature. Cells were washed with PBS and treated with hypotonic buffer (20 mM Hepes pH 7.9, 10 mM KCl, 1 mM EDTA, 10% glycerol and 1 mM DTT with additional protease inhibitor (Roche)) to isolate nuclei. The nuclei were suspended with RIPA buffer (10 mM Tris-HCl pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate with protease inhibitor) and sonicated using a Covaris S220 Focused-ultrasonicator. Fragmented chromatin was pre-cleared with protein G-conjugated sepharose beads (GE Healthcare).

Antibodies against HLF (Santa Cruz, sc-134359, 5 µg antibody was applied to 1 ml cell lysate per ChIP), MAFG (Santa Cruz, sc-166548 X, 5 µg antibody was applied to 1 ml cell lysate per ChIP), histone H3K4me1 (Abcam, ab8895, 5 µg antibody was applied to 1 ml cell lysate per ChIP), H3K4me3 (Abcam, ab8580, 5 µg antibody was applied to 1 ml cell lysate per ChIP), H3K27ac (Abcam, ab4729, 5 µg antibody was applied to 1 ml cell lysate per ChIP), and CTCF (Santa Cruz, sc-15914 X, 5 µg antibody was applied to 1 ml cell lysate per ChIP) were used to pull down the respective proteins and their associated chromatin. Washes with different concentration of NaCl were performed. The enriched protein–DNA complexes were reverse crosslinked at 65 °C overnight with proteinase K (NEB). DNA was purified with a Qiagen MinElute kit.

A sequencing library was prepared using an in-house kit, including end-repair, A addition and adaptor ligation. The library was sequenced with Illumina HiSeq 4000 for 50-bp single reads or 100-bp pair-end reads.

Reads were aligned using BWA MEM[38] with either single-end or pair-end model to the hg19 reference genome. Reads with low mapping quality (mapq <10) were filtered out, and PCR duplicates were removed using Picard tool (http://broadinstitute.github.io/picard/). MACS2[45] were then used to call peaks and generate signal tracks to view in the genome browser.

In addition to ChIP–seq performed in this study, ChIP–seq results for additional transcription factors were also collected from the ENCODE project (Supplementary Table 4). For allelic analysis, reads were aligned using WASP mapping pipeline to control potential allelic mapping bias[46]. Specifically, heterozygous SNPs called using whole-genome sequencing (WGS) data were used for HepG2 cells, and heterozygous SNPs from the 1000 Genome Project were used for GM12878 cells. Allelic read counts for each phased heterozygous SNP within the 300-bp window in transcription factor ChIP–seq data and corresponding control data were counted. Specifically, only reads with high mapping quality (mapq >10) and base pairs with high accuracy (base call quality >13) were used. To remove sampling biases, SNPs that are covered by less than 20 reads in either the treatment or the control were filtered out. Odds ratios were then computed for each SNPs comparing allelic counts between the treatment and control to measure allelic imbalance. SNPs were tested for allelic imbalance using binomial test using background ratio derived from control data. SNPs with Benjamini–Hochberg adjusted P value <0.05 were considered as allelic imbalanced.

Genotyping and haplotype phasing of HepG2 cells

The genomic DNA was extracted using Qiagen kit (69506). The DNA was then fragmented with Covaris S220 ultrasonicator to length of 300–500 bp. A sequencing library was then prepared using the same in-house kit as for ChIP–seq, including end-repair, A addition and adaptor ligation. The library was sequenced with Illumina HiSeq 4000 sequenced for 100-bp paired-end reads to achieve an average coverage of 30–40 times of the human genome.

Reads from WGS were aligned using BWA MEM[38] in pair-end model with default parameters. PCR duplicates were removed using Picard tools (http://broadinstitute.github.io/picard). Variants were then called according to the GATK best practice pipeline using GATK 3.6-0[47–49]. In brief, reads were realigned locally, and base pair qualities were recalibrated. Variants were then called using HaplotypeCaller with default parameters. Variants were then recalibrated based on known gold-standard variants. Only variants that passed filters were used in the downstream analysis.

To obtain haplotypes, aligned Hi-C bam files were processed through the GATK realignment pipeline as for the WGS data described above. Two filters were applied to SNPs so that bi-allelic SNPs and heterozygous SNPs with high genotype quality (GQ >20) were kept. WGS and Hi-C data were then parsed to extract informative fragments with extractHAIRs[50] using filtered SNPs. The fragments from Hi-C and WGS data were combined, and HAPCUT2[50] was used to derive haplotypes. Results from HAPCUT2 were then paired with SNPs in 1000 Genome Project phase 3 data, and Beagle 4.1[51] was used to impute haplotypes for SNPs that were not phased by HAPCUT2. We obtained chromosome-span haplotypes for all auto chromosomes except for chr22 (Supplementary Table 10). Phasing quality was further examined by computing fraction of homologous trans (h-trans) reads in RNA-seq data from HepG2 cells. Specifically, h-trans reads are read pairs that contain SNPs from both haplotypes. Chromosome-span haplotypes with high accuracy were obtained (Supplementary Table 10).

Differential gene-expression analysis

HepG2 (ATCC) cells were cultured under normal conditions with 5% CO_2 at 37 °C. For siRNA transfection, HiPerfect transfection was used following the manufacturer's guidance. For each experiment, 50 nM of siRNA was used with 5 µl of HiPerfect reagent to make the transfection complex for $1–3 \times 10^4$ cells. Cells were continued to be cultured for 72 h. The siRNAs targeting human *HLF* (cat. no. GS3131) and *MAFG* (cat. no. GS4097) were commercially available from Qiagen. Silencer negative control siRNA was manufactured by Thermo Fisher (cat. no. AM4635).

Total RNA was isolated using Qiagen RNeasy mini kit. The sequencing library was prepared using the Illumina Truseq RNA Library Prep Kit v2 (cat. no. RS-122-2001). The library was sequenced using Illumina HiSeq 4000 for 100-bp paired-end reads.

Reads were aligned to the hg19 reference genome using STAR 2.4.2a[52] with default parameters in paired-end mode. Only uniquely aligned reads were kept for further analysis. Cufflinks 2.2.1[53] was used to compute FPKM for each gene.

For allelic gene-expression analysis, reads were aligned to the hg19 reference genome using STAR and WASP[46] pipeline to control allelic mapping bias. The same set of SNPs and haplotypes were used for

RNA-seq as ChIP–seq as described above in HepG2 cells. Allelic counts for each gene were generated using htseq-count 0.6.0[54]. Genes with at least 10 allelic reads were tested for allelic imbalance using the binomial test using background ratio derived from WGS data. Genes with Benjamini–Hochberg adjusted *P* value <0.1 were considered allelic imbalanced.

For differential gene-expression analysis, read counts for each gene were obtained using htseq-count[54] using GENCODE human annotation release 24 as reference[55]. DESeq2[56] was used to identify differentially expressed genes using default parameters. Genes with Benjamini–Hochberg adjusted *P* value <0.2 were considered as differentially expressed. KEGG pathway enrichment analysis was performed with DAVID[57].

Enhancer perturbation using CRISPRi

CRISPR–dCas9 fused with KRAB domain (Addgene 71236) was introduced to the genomic locus enclosing the SNP rs7578326 using sgRNA (targeting sequence TCCGTTGGTGACACAGTTGG) in HepG2 cells. CRISPR–dCas9 with the same sgRNA was used as negative control. Similarly, both plasmids were transfected in HEK 293T cells as control. RNA was extracted using Qiagen RNeasy kit and reverse transcribed using High-Capacity cDNA Reverse Transcription Kit (Thermo). Quantitative PCR was performed to measure the expression of *IRS1* gene using pre-designed primers (Qiagen QT00074144) and beta actin for internal control (Qiagen QT00095431). Triplicates were carried out for each experiment for the *t*-test.

Determination of transcription factor binding preference using PWMs

Using motifs from the previous HT-SELEX study[3], the score for reference and alternative genomic oligonucleotide sequences was measured for 255 distinct transcription factors with SNP-SELEX data. In particular, we used the pssm function from Biopython[58] to obtain position specific scoring matrices (PSSM) for each motif. PWM score of each sequence was then obtained by computing the maximum motif score of a sliding window over sequence in both forward and reverse strand. For each position, the calculate function from Biopython was used to calculate PWM scores. PWM scores for two alleles were calculated separately and ΔPWM scores were then computed as the difference of PWM score between reference allele *r* and alternative allele *a*.

To assign significance for ΔPWM scores, we used atSNP[59] to calculate *P* values for each SNP–transcription factor pair. In brief, atSNP estimates random distribution for each motif and used the random distribution to calculate *P* values. The same *P* value cut-off ($P < 0.01$) was used to select SNPs with allelic transcription factor binding predicted by ΔPWM scores.

Development of deltaSVM models

Training of deltaSVM models. Fastq files of 533 transcription factors from a previous HT-SELEX study[3] were used to build deltaSVM models. For each transcription factor, each sequence retained after every SELEX cycle was used as positives and the sequences only present in cycle 0 as negatives (Extended Data Fig. 6a). Both positive and negative sequences were randomly down sampled to 20,000 sequences due to computing capacity. The gkm-SVM models were trained using lsgkm[60] with two *k*-mer sizes, using parameters '-l 10 -k 6 -d 3' and '-l 8 -k 5 -d 3', respectively.

We then calculated deltaSVM scores using trained gkm-SVM models as described[10] using 40-bp sequences with SNP at the centre for each transcription factor–SNP pair. In brief, scores for each 10-mer were pre-computed using aforementioned gkm-SVM models via the gkmpredict command. Therefore, scores for any SNP-containing 10-mer genomic sequences can be assigned, regardless of the position of the SNP within the 10-mer. When defining delta, we calculated the sum of subtractions between two alleles in all 10-mers overlapping the SNP (that is, (summed SVM scores of all 10-mers containing reference allele) – (summed SVM scores of all 10-mers containing alternative allele)). Specifically, we used deltasvm.pl script from http://www.beerlab.org/deltasvm/. We used 10-mer as the default parameter in deltaSVM without testing additional parameters, although it is possible other lengths of *k*-mer may lead to even better performance.

For each transcription factor, we trained gkm-SVM models for two parameters and all six SELEX cycles. We then select best models among them for each transcription factor as described below.

Validation of deltaSVM models by cross-validation. To validate deltaSVM models, we performed fivefold cross-validation. Specifically, pbSNPs ($P < 0.01$) and non-pbSNPs ($P > 0.5$) from SNP-SELEX experiments were used as positives and negatives respectively. SNPs were then divided into for five equal-sized fractions for each transcription factor, while ensuring the equal number of pbSNPs and non-pbSNPs within each fraction. We then selected best model using fourfolds of SNPs based on AUPRC and then tested performance of the model on the remaining fold. The same procedure was performed for each fold. To ensure the quality of data, only transcription factors with more than 40 pbSNPs were used in testing.

The AUROC and AUPRC for each model were computed using R package PPROC[61].

Validation of deltaSVM models in the novel batch SNP-SELEX experiments. To fully avoid over-fitting issues, we performed another novel batch of SNP-SELEX experiments as described in previous sections. For each transcription factor, the best model was selected based on AUPRC calculated on the entire set of pbSNPs and non-pbSNPs in the first batch of SNP-SELEX experiments. The models were then used to calculated deltaSVM scores for each SNP tested in the novel batch SNP-SELEX experiments.

After removing 5,309 SNPs with the first batch, there are 959,367 transcription factor–SNP pairs including 61,020 SNPs and 487 transcription factors. Among them, there are 21,299 unique pbSNPs (Supplementary Table 8). Among them, only 87 transcription factors with >40 pbSNPs and for which both PWM models and deltaSVM models are available were included for comparison.

Comparison of PWM models with deltaSVM models. To compare the performance comparison of PWM and deltaSVM models, two methods were used to calculate ΔPWM scores. For multi-nominal models, ΔPWM scores were calculated as described in the previous section for all transcription factor–SNP pairs. For BEESEM[5] models, beesem.py from https://github.com/sx-ruan/BEESEM was used to generate PWM models with default parameters. The BEESEM-derived PWM models were then used to calculate ΔPWM scores as described in the previous section. Both PWM models were applied to the same set of SNPs as deltaSVM models to compare performance.

For cross-validation, exactly the same set of SNPs were used in each fold for each transcription factor to ensure a fair comparison. Similar to deltaSVM models, best models were selected using SNPs in fourfolds based on AUPRC and SNPs in the remaining fold were used to compare performance. AUPRC and AUROC were also calculated using R package PPROC[61]. Only 129 transcription factors for which both PWM models and deltaSVM models are available were included for comparison.

For the novel batch of SNP-SELEX experiments, the same set of SNPs for each transcription factor were used to compare performance. Best models were selected based on AUPRC using all SNPs in the first batch SNP-SELEX experiments for each transcription factor. Then ΔPWM scores were calculated for each SNP using the selected models for multinomial and BEEMSEM models respectively. AUPRC and AUROC were then calculated for each transcription factor to compare performance. Only 87 transcription factors with >40 pbSNPs and for which

both PWM models and deltaSVM models are available were included for comparison.

Prediction of the effect of each SNP on transcription factor binding. To predict the effect of each SNP on transcription factor binding, a pair of 40-bp genomic sequences from the hg19 reference genome were selected, with the SNP to test located in the centre of the oligonucleotide. We first scored both sequences using gkm models and determined if at least one of oligonucleotides can be bound by the transcription factor. The threshold was determined based on bound oligonucleotides identified using SNP-SELEX experiments. Specifically, we computed gkm scores for all bound oligonucleotides and used the medium of the scores for the bound oligonucleotides for each transcription factor as the threshold to determine transcription factor binding. Only bound oligonucleotides were further predicted for allelic transcription factor binding.

The deltasvm.pl script was used to predict preferential binding of the transcription factor to the oligonucleotide sequences with the reference allele and alternative allele. We computed deltaSVM scores for all pbSNPs and used the medium of pbSNPs' scores for each transcription factor as the threshold to determine allelic transcription factor binding.

Validation of the predicted SNPs effect on transcription factor binding using ChIP–seq data

We made predictions for heterozygous SNPs covered by at least 20 allelic reads in ChIP–seq experiments in HepG2 and GM12878 cells[7], respectively. For each transcription factor ChIP–seq experiment, we computed the percentage of allelic imbalanced SNPs in predicted pbSNPs and non-pbSNPs. Confidence intervals for fraction of allelic imbalanced SNPs were calculated using binom.confint function in the R package binom. Allelic imbalanced SNPs were determined as described in the previous section. For ΔPWM models, predicted pbSNPs were determined similarly using the median ΔPWM score for the bound oligonucleotides and pbSNPs, respectively.

Prediction of the transcription factors implicated in complex traits and diseases

To predict potential transcriptional regulators that may contribute to complex traits and disease, we applied S-LDSC[62] to examine whether SNPs affecting certain transcription factor binding are enriched in GWAS signals of complex traits and disease[18,63–71]. In brief, S-LDSC models the casual effect of each SNP for a given trait as a linear additive contribution by a list of annotations and then estimates per-SNP heritability for each annotation as regression coefficient considering not only the SNP to test but also all SNPs in linkage disequilibrium. Then, the *P* value was computed to test if regression coefficient for annotation *i* is positive, which means annotation *i* explains additional heritability on top of other annotations. In other words, annotation *i* are enriched for SNPs associated with the trait.

We made predictions for 94 transcription factors with excellent deltaSVM models (AUPRC >0.75) for all common SNPs in 1000 Genome Project phase 3 for European populations as mentioned above. The list of SNPs was obtained from website https://data.broadinstitute.org/alkesgroup/LDSCORE/. The SNPs predicted to have an effect for each transcription factor within the accessible chromatin regions in ENCODE DHS sites[32] or FANTOME 5[33] permissive enhancers for all cell and tissue types were then used as annotation to estimate annotation-specific linkage disequilibrium scores for each transcription factor. We then ran LDSC using these SNPs for each transcription factor along with 53 baseline models including genic regions, enhancer regions and conserved regions. In many cases, the SNP does not affect transcription factor binding, even though the transcription factor binds to the SNP. To rule out this scenario, we also included predictions for SNPs bound by the transcription factor in the regression model. In summary, we ran LDSC using 55 annotations including predicted SNPs with allelic transcription factor binding, binding SNP prediction, and 53 baseline models, and *P* values for regression coefficient for each transcription factor were used to measure whether predicted SNPs with allelic transcription factor binding explains additional heritability. The *P* values for the term of binding SNP prediction were used in Extended Data Fig. 10a.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sequencing data generated in this study can be accessed via Gene Expression Omnibus (GEO) under accession number GSE118725. The raw sequencing data for transcription factor ChIP–seq of GM12878 is extracted from the ENCODE portal (https://www.encodeproject.org). The specific transcription factor data can be accessed by searching the accession numbers listed in Supplementary Table 4. The web portal (http://renlab.sdsc.edu/GVATdb/) provides a searchable interface for SNPs and transcription factors tested in the current study. Enriched motifs for SNP-SELEX experiments using Homer are presented in Supplementary File 1. Scores for all tested SNP–transcription factor pairs from SNP-SELEX experiments are shown in Supplementary File 2. The data for high-confidence allelic binding of 94 transcription factors to all common SNPs in the human genome predicted by deltaSVM models are presented in Supplementary File 3.

Code availability

Custom codes used to process and generate the results described in the current study were deposited to GitHub (https://github.com/ren-lab/snp-selex).

29. Kato, N. Insights into the genetic basis of type 2 diabetes. *J. Diabetes Investig.* **4**, 233–244 (2013).
30. Mahajan, A. et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
31. Johnson, A. D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
32. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
33. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
34. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
35. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
36. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44** (D1), D110–D115 (2016).
37. Nitta, K. R. et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4**, (2015).
38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. Zhou, X., Lindsay, H. & Robinson, M. D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* **42**, e91 (2014).
40. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
41. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
42. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
43. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
44. Yan, J. et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801–813 (2013).
45. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
46. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
47. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
48. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

49. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.3 (2013).
50. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
51. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
52. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
53. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
54. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
55. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).
56. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
57. Dennis, G., Jr et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, 3 (2003).
58. Cock, P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
59. Zuo, C., Shin, S. & Keleş, S. atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* **31**, 3353–3355 (2015).
60. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).
61. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
62. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
63. Dubois, P. C. et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
64. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
65. Lambert, J. P. et al. Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nat. Methods* **10**, 1239–1245 (2013).
66. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
67. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
68. Bentham, J. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).
69. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
70. Nelson, C. P. et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
71. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

Acknowledgements We thank S. Preissl and S. A. Chen for insightful comments during manuscript preparation; and S. Kuan, Z. Liu and B. Li for technical assistance. This work was supported by the Ludwig Institute for Cancer Research (B.R.), NIDDK (U01 DK105541 to B.R., M.S., and K.F.), Vetenskapsrådet Sweden (537-2014-6796 to J.Y.), and a CAPES foundation fellowship (BEX 5304/15-6 to A.M.R.S.).

Author contributions B.R., M.S., K.J.G., K.A.F., J.T. and J.Y. conceived the project. J.Y., Y.Y., X.L., N.N., and N.V. carried out experiments. Y.Q., A.M.R.d.S., Y.E.L., A.R., S.F., P.B., F.C. and J.C. performed data analysis. J.Y., Y.Q., A.M.R.d.S., J.T. and B.R. wrote the manuscript with input from all co-authors.

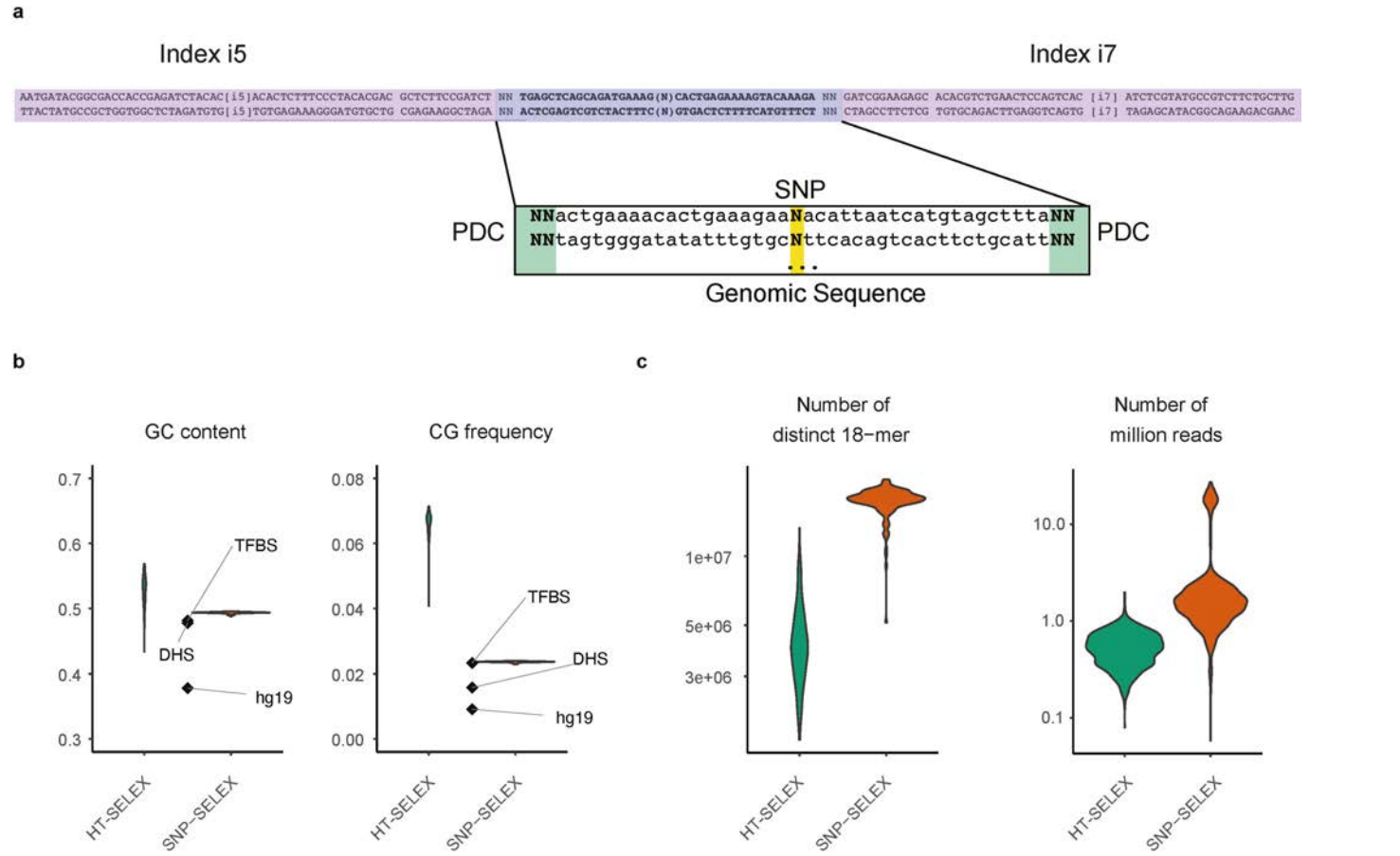
Competing interests B.R. is a co-founder and consultant for Arima Genomics and a co-founder of Epigenome Technologies.

Additional information

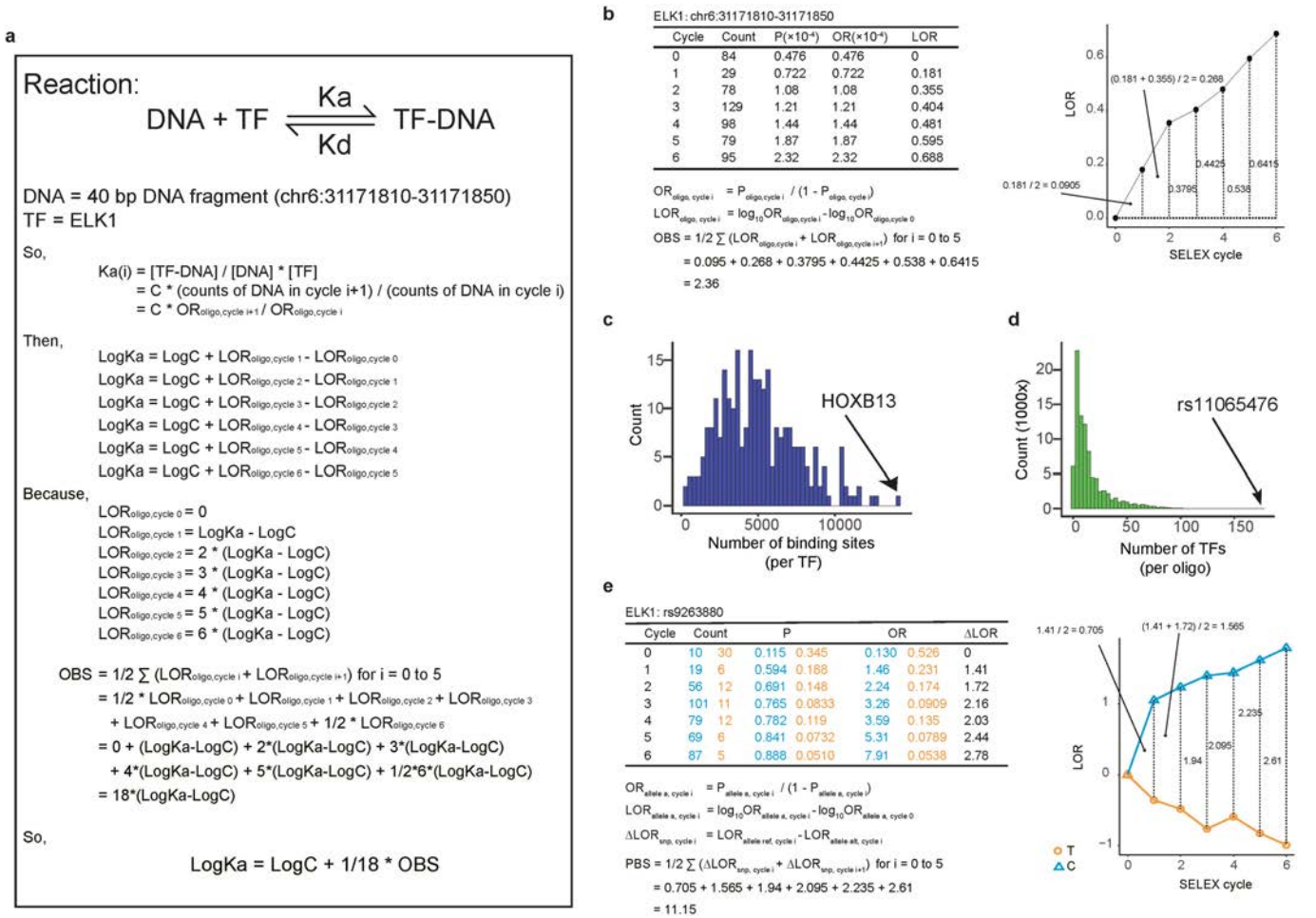
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03211-0.

Correspondence and requests for materials should be addressed to J.Y., J.T. or B.R.

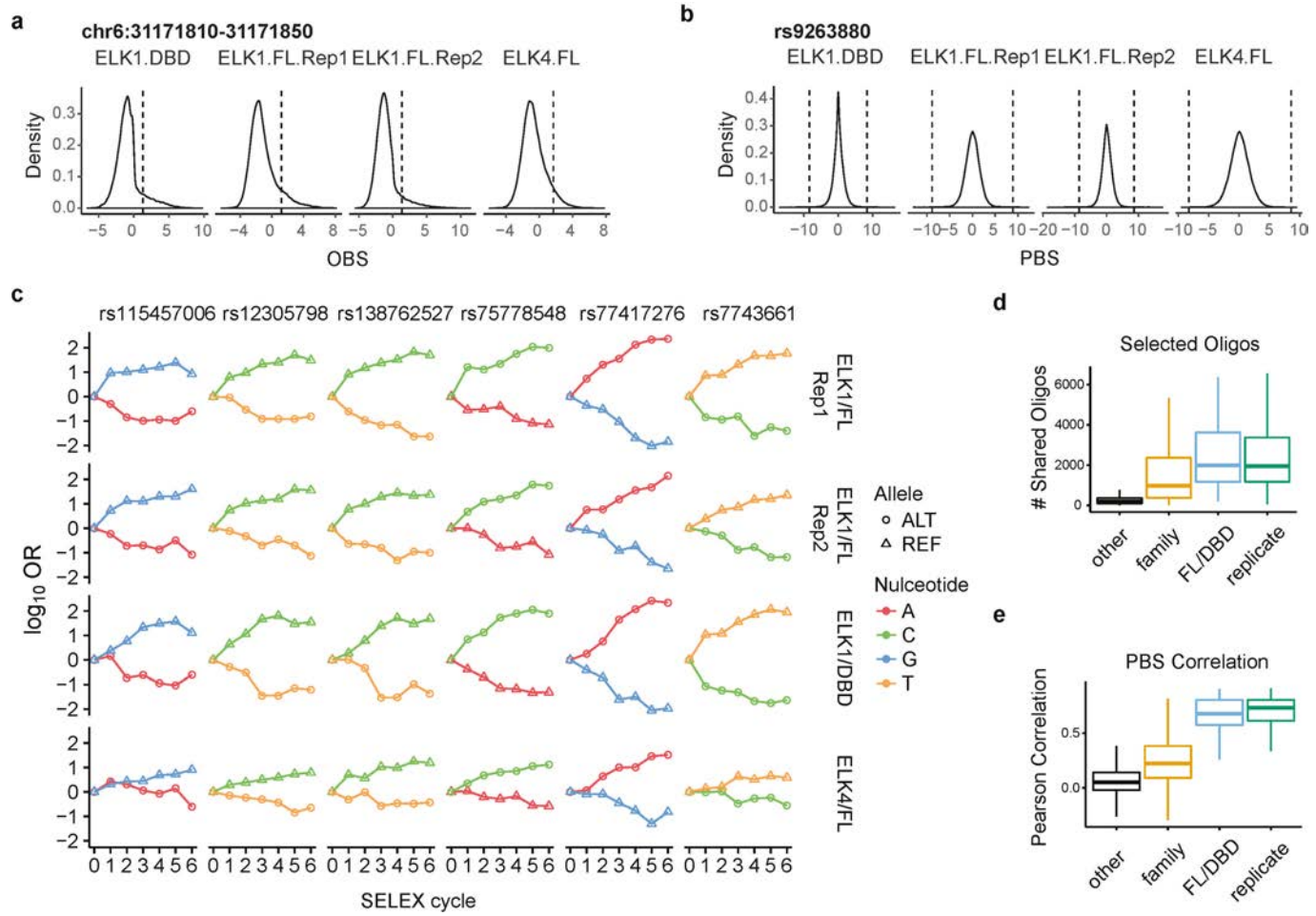
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work.



Extended Data Fig. 1 | The sequence features of input oligonucleotides. **a**, An example of the oligo design for SNP-SELEX. Two random nucleotides were added to each end of the oligos as unique molecule identifiers (UMIs) to remove over-represented PCR duplicates. Illumina TruSeq dual-index system was adapted for oligo design. **b**, The GC content (left) and CpG frequency (right) of SNP-SELEX input were more similar to those of TF binding sites in the human genome (TFBS), open chromatin (DHS) and the entire human genome in general (hg19) than random sequences used in HT-SELEX. **c**, Comparison of *k*-mer coverage (left) and sequencing depth (right) of libraries between SNP-SELEX and HT-SELEX.



Extended Data Fig. 2 | Derivation of OBS and PBS. a, Equations demonstrate the relationships between OBS and the association constant (Ka) of TF-DNA interactions. **b**, An example of how oligonucleotides were evolutionarily selected during SNP-SELEX. Table of counts for oligonucleotide chr6:31171810-31171850 is shown at left and the OBS curve is shown on the right. **c**, **d**, Histograms show the number of oligonucleotide sequence bound by each TF (**c**), the number of binding TFs for each oligonucleotide sequence (**d**). **e**, An example of how the abundance of SNPs varies in the course of a SNP-SELEX experiment. The table of counts for SNP rs9263880 is shown at the left and PBS curve is shown on the right. The orange line inside the black boxes indicates the reads of T-allele-containing fragment and the blue line shows the reads of C-allele-containing fragment.

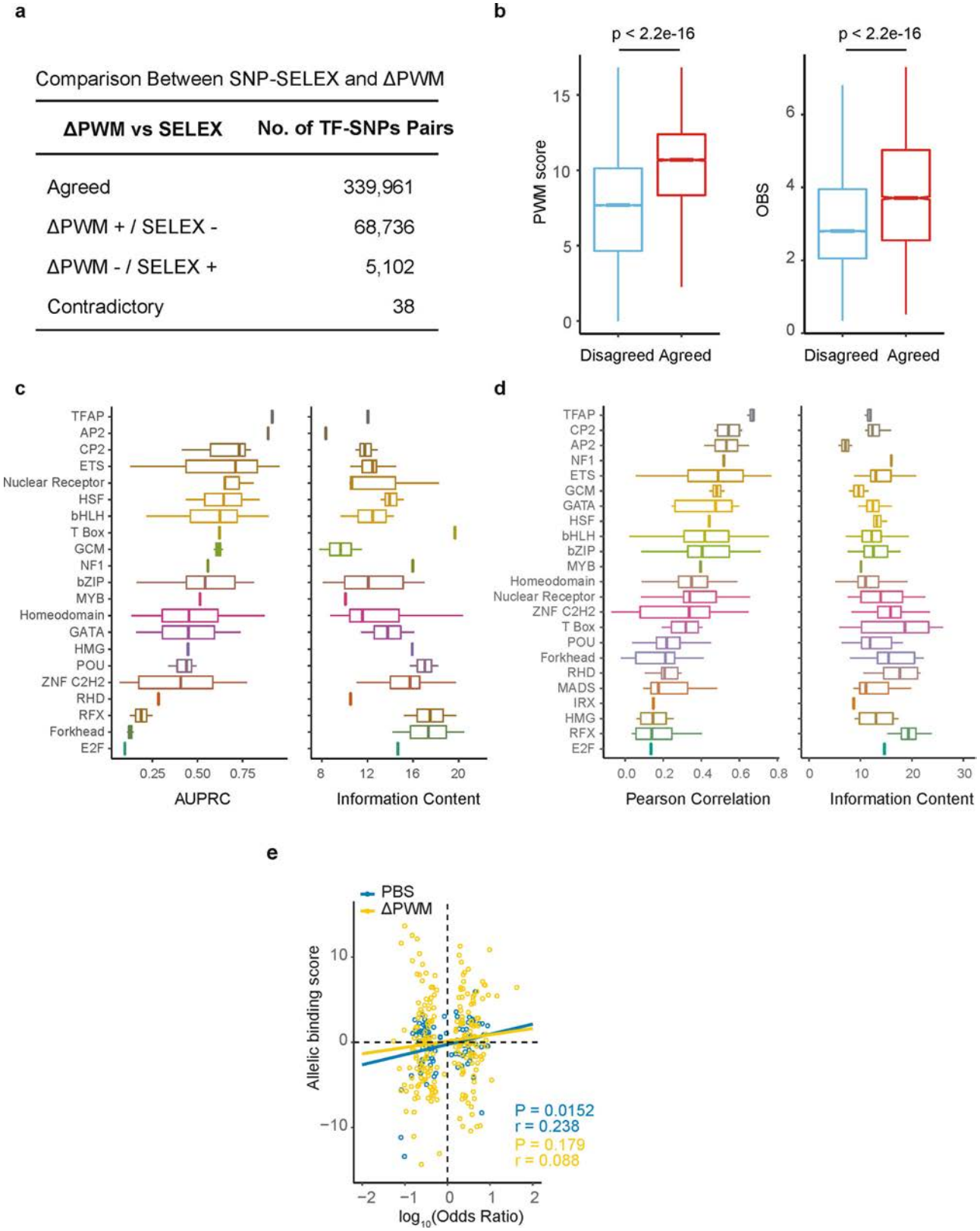


Extended Data Fig. 3 | Reproducibility of SNP-SELEX data. a, Density plots show an example of the distribution of OBS of all oligos assayed in ELK SNP-SELEX replicative experiments. Vertical dashed lines indicate the cut-off for significant binding sequences ($P = 0.05$ by Monte Carlo randomization). The 40-bp genomic sequences with OBS that is over the indicated values are recognized as significant binding sites of ELK1 or ELK4. DBD: DNA binding domain. FL: full-length protein. **b**, Density plots show an example of the distribution of PBS of all oligos assayed in ELK SNP-SELEX replicative experiments. Vertical dashed lines indicate the cut-off for significantly differential binding ($P = 0.01$ by Monte Carlo randomization). The 40-bp SNP-containing genomic sequences with PBS over the indicated values are recognized as significantly differential (allelic) binding sites of ELK1 or ELK4. DBD: DNA binding domain. FL: full-length protein. **c**, An example illustrating differential DNA binding at six SNPs, in four SNP-SELEX experiments, including (i) two full-length ELK1 replicates, on the first two lines; (ii) one DNA binding domain (DBD) ELK1, on the third line; and one full-length ELK4 TF which belongs to the same structure family, on the last line. Each panel represents the logarithmic odds-ratio (y-axis) of observing the reference allele (REF), represented by a triangle, and the alternative allele (ALT), represented by a circle, over SNP-SELEX cycles (x-axis). The two alleles of each SNP are coloured according to their nucleotides, where A is red, C is green, G is blue, and T is yellow. The figure shows that SNP-SELEX experiments of both replicates, full-length, DBD, and same structure TF family presents the same allelic preference. **d**, **e**, Comparison of oligonucleotide enrichment (**d**) and allele preference (**e**) between different biological replicates (replicates), full-length (FL), and DNA Binding Domain (DBD), members of the same structural family (family), and random pairs (others). For each pair of experiments, we compared the oligonucleotides that display binding in both experiments for binding oligonucleotides and compared PCC between the PBS from each experiment. Horizontal line is median; hinges are 25th and 75th percentile; whiskers are most extreme value no further than 1.5× interquartile range (IQR).

a

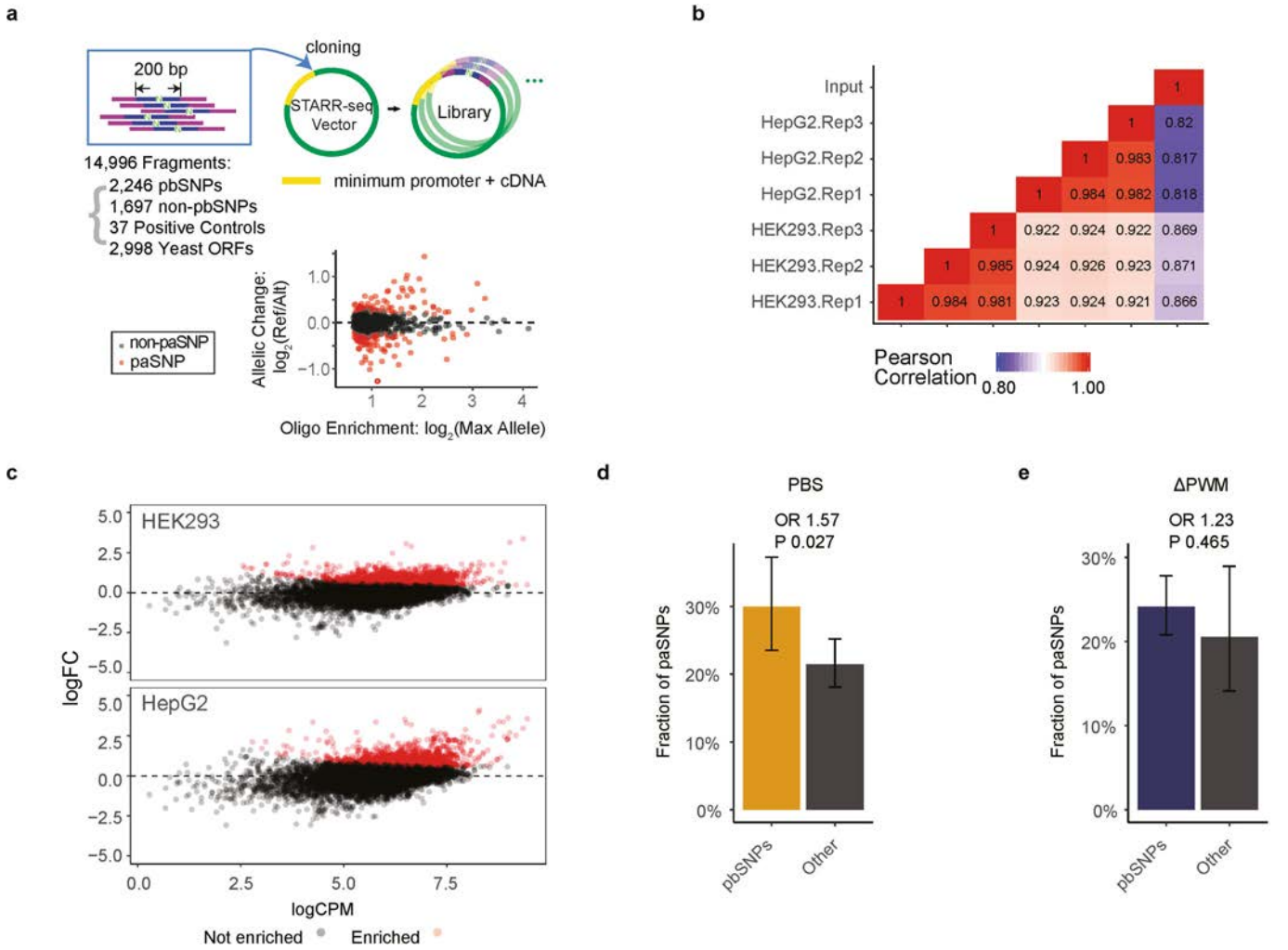
Comparison Between SNP-SELEX and ΔPWM

ΔPWM vs SELEX	No. of TF-SNPs Pairs
Agreed	339,961
ΔPWM + / SELEX -	68,736
ΔPWM - / SELEX +	5,102
Contradictory	38

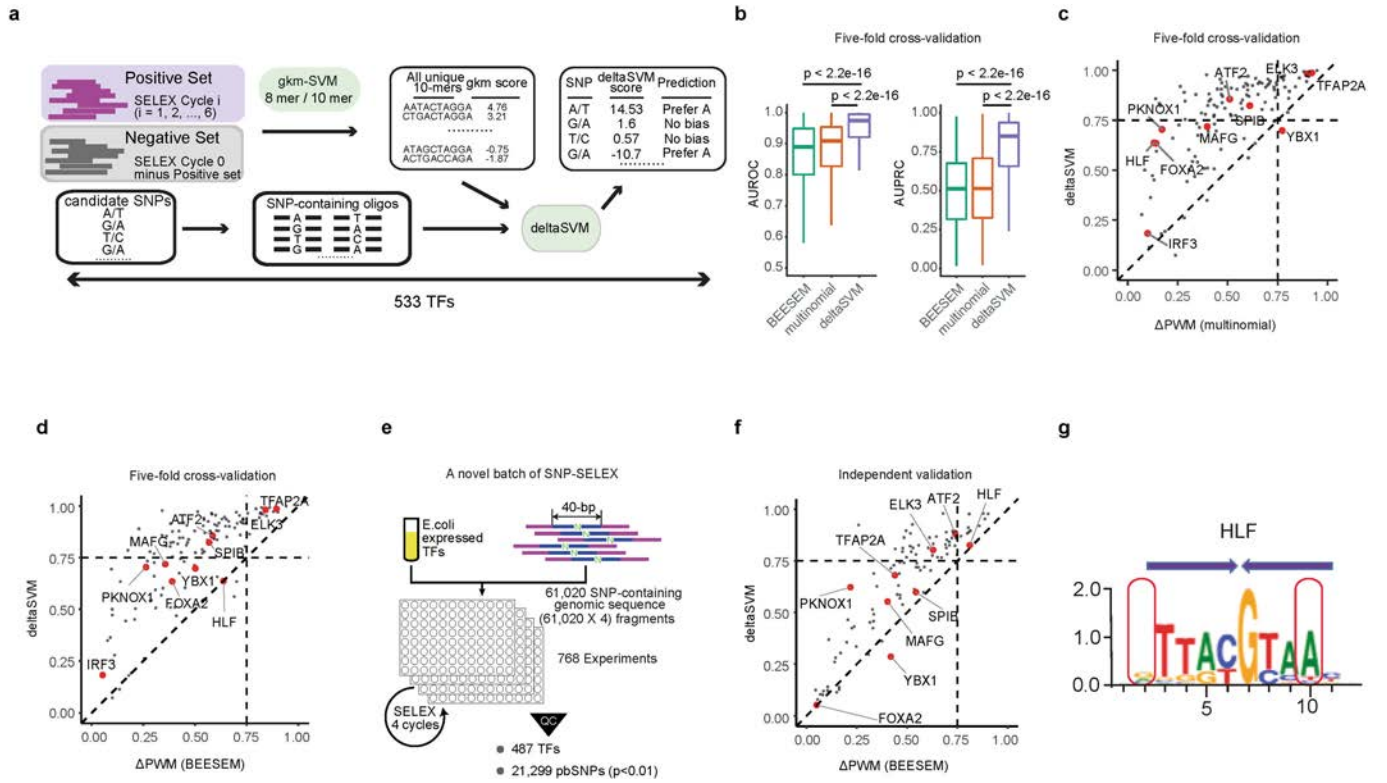


Extended Data Fig. 4 | See next page for caption.

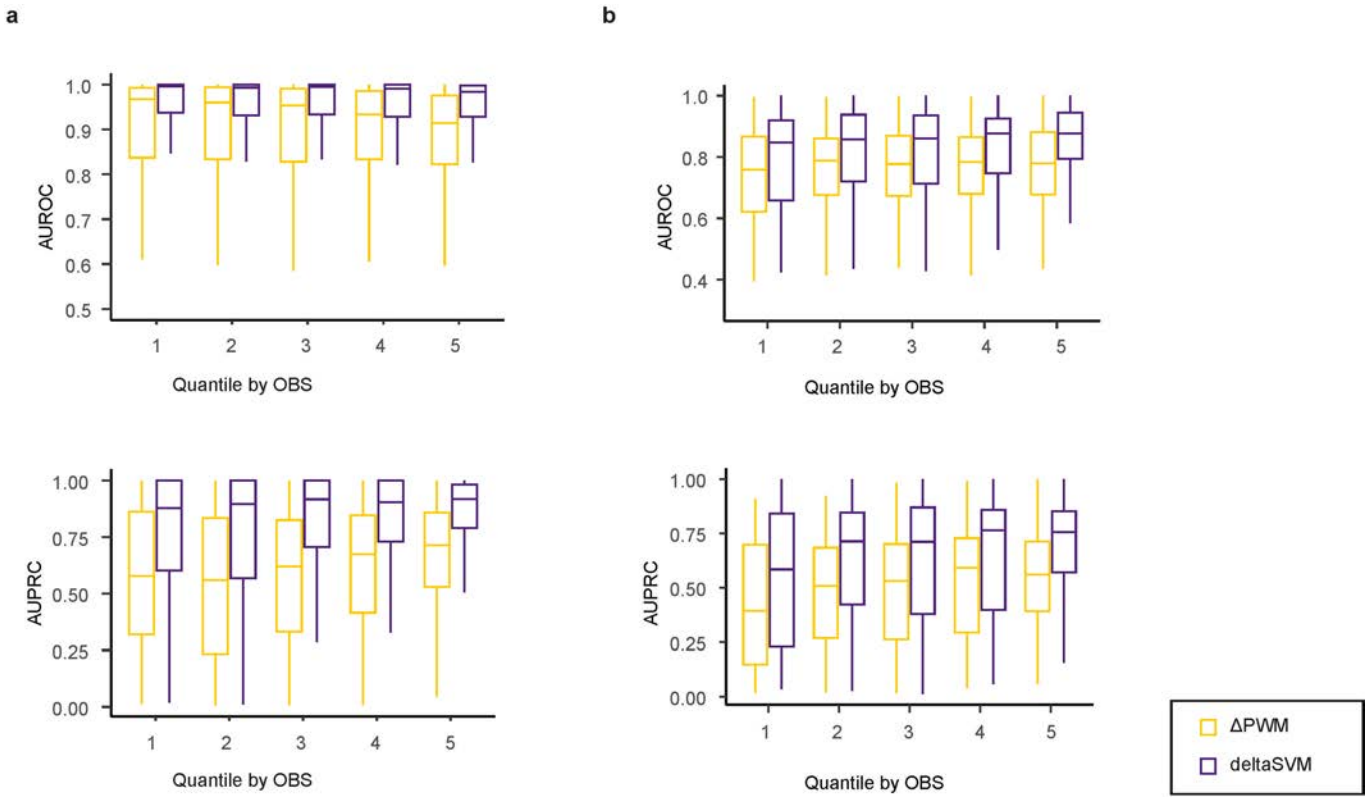
Extended Data Fig. 4 | SNP-SELEX results are correlated with TF binding in vitro and in vivo. a, Comparison of the SNPs with differential TF binding determined by SNP-SELEX and ΔPWM. An error matrix table showing the number of SNPs for which the same allele was identified as the preferred allele by both methods (Agreed), SNPs for which one allele was determined as preferential substrate by one method but no allele was called by the other (PWM+/ SNP-SELEX− and PWM−/ SNP-SELEX+), and SNPs where different alleles were called as preferential bound by each method (Contradictory). Note that the vast majority of the results agreed, with the most disagreement coming from PWM+/ SNP-SELEX−. **b**, Comparison of the PWM scores (left) and the OBS scores (right) between SNPs with concordant and discordant predictions. Note that discordant predictions mostly come from weak binding sites with low PWM scores and low OBS scores. Two-sided Mann–Whitney *U* test *P* value is shown on the top. Horizontal line is median; hinges are 25th and 75th percentile; whiskers are most extreme value no further than 1.5× IQR. **c**, Box plots show performance of ΔPWM in predicting pbSNPs grouped by DNA binding domain structural families (left) and information content of motifs for each corresponding TF family (right). AUPRC is used to evaluate the performance of ΔPWM. Horizontal line is median; hinges are 25th and 75th percentile; whiskers are most extreme value no further than 1.5× IQR. **d**, Box plots show PCC between PBS and ΔPWM (left) and information content (right) for each TF family. PCCs for some TF families are higher than others, independent of the information content (IC) of corresponding PWM models. Horizontal line is median; hinges are 25th and 75th percentile; whiskers are most extreme value no further than 1.5× IQR. **e**, A scatter plot shows the correlation of PBS and allelic binding ratio derived from SNP-SELEX and ChIP–seq in GM12878 cells respectively. The PCCs and *P* values calculated based on *t*-test are shown on the lower right corner. The allelic binding ratio is computed as the $\log_{10}$ odd ratio over input (see Methods for details). In total, 341 TF-SNP pairs including 269 unique SNPs and six TFs were plotted. TFs used include ATF2, PKNOX1, IRF3, NR2F1, YBX1, and TBX21.



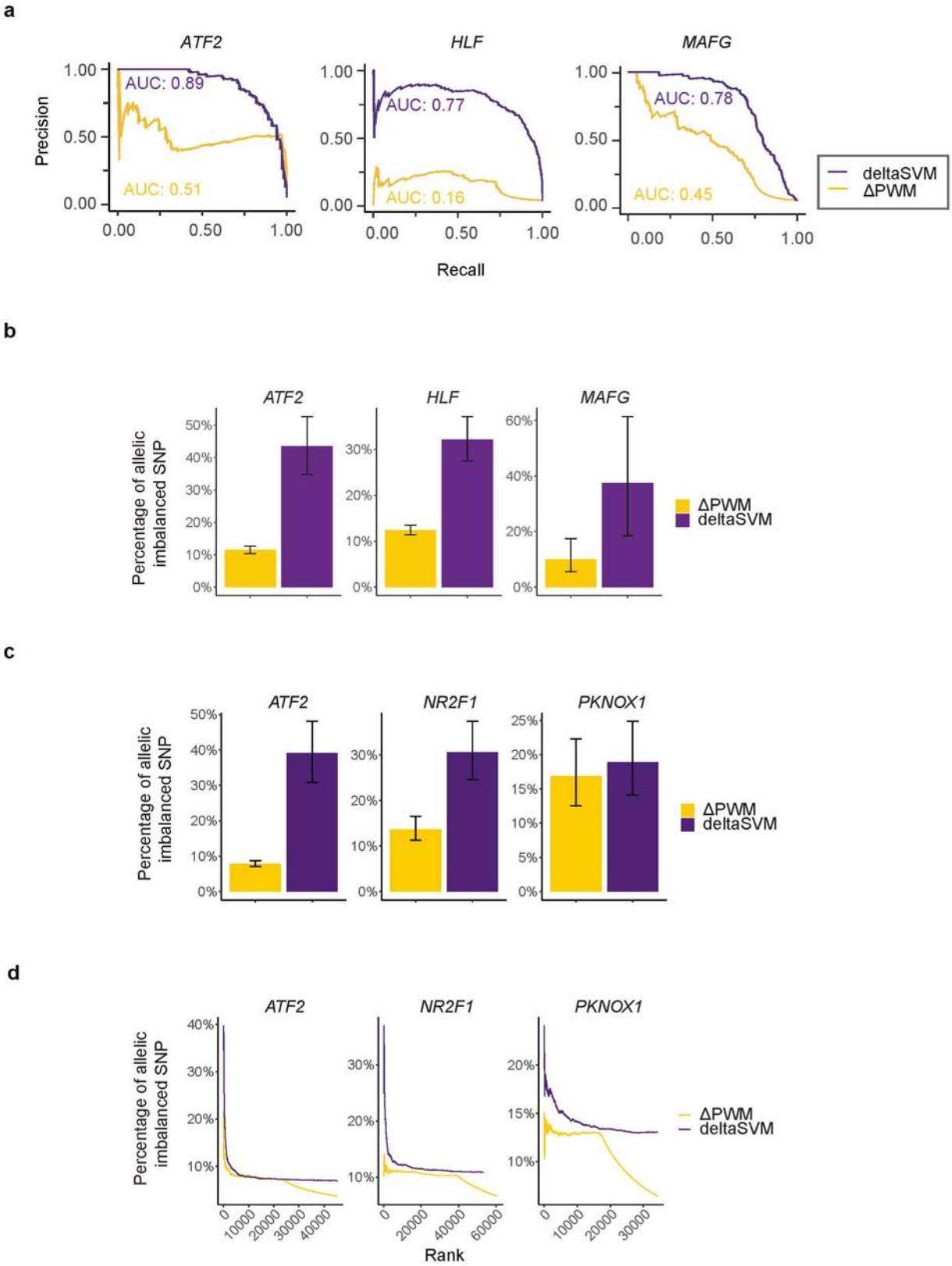
Extended Data Fig. 5 | SNP-SELEX results are correlated with allelic enhancer activities detected using high-throughput reporter assays. **a**, A schematic diagram shows the strategy of using STARR-seq to assess the effect of SNPs in enhancer activity in HepG2 and HEK 293T cells. **b**, Heat map shows pair-wise PCCs calculated among STARR-seq datasets. The read counts of each SNP in the starting reporter library, in the mRNA pools in three HepG2 replicates, and three HEK 293T replicates were used for PCC calculation. **c**, MA plot of the logarithmic fold-change (*y*-axis) of read counts of SNP-containing mRNA over that of the input library expressed as logarithmic counts per million (CPM) (*x*-axis) for HEK 293T, on the top panel, and HepG2, on the bottom panel. Each dot on the plot corresponds to an oligonucleotide, and the oligonucleotides showing enrichment (empirical FDR <0.05) are colored in red. **d**, Bar plots comparing the fractions of paSNPs determined using STARR-seq in pbSNPs and non-pbSNPs by SNP-SELEX. Odds Ratio (OR) is shown between imbalanced and balanced SNPs, and the *P* value is calculated by Fisher exact test. Error bars denote the 95% confidence interval calculated by Wilson method (Methods). Only pbSNPs corresponding to the highly expressed TFs (RPKM >3) in the cell lines are considered for the analysis. $n = 167$ SNP-cell pairs for pbSNPs; $n = 509$ SNP-cell pairs for non-pbSNPs. **e**, Bar plots comparing the fractions of paSNPs determined using STARR-seq in pbSNPs and non-pbSNPs predicted by ΔPWM. SNPs with $P < 0.01$ by atSNP were considered as pbSNPs. Odds Ratio (OR) is shown between imbalanced and balanced SNPs, and the *P* value is calculated by Fisher exact test. Error bars denote the 95% confidence interval calculated by Wilson method (Methods). Only pbSNPs by highly expressed TFs (RPKM >3) in the corresponding cell lines are considered for the analysis. $n = 564$ SNP-cell pairs for pbSNPs; $n = 112$ SNP-cell pairs for non-pbSNPs.



Extended Data Fig. 6 | deltaSVM more accurately predicts effects of noncoding variants on TF binding in vivo than ΔPWM. a, A schematic graph for the training of deltaSVM models for 533 TFs. Data from previously reported HT-SELEX experiments using random DNA oligonucleotide sequences were used to derive these models. To develop deltaSVM models for each TF, the reads in each HT-SELEX cycle beyond cycle 0 reads were used as positive training sets, and the reads not enriched were used as negative training sets. All unique 10-mers were scored using gapped *k*-mer models to compute weights for deltaSVM. The two alleles of the 40-bp SELEX oligos were then scored using deltaSVM models to generate deltaSVM scores. **b**, Box plots compare the performance of deltaSVM, PWM derived from HT-SELEX with the multinomial or BEESEM algorithms in predicting pbSNPs for 129 TFs. The results from fivefold cross-validation were shown. Two statistical evaluations were used, including AUROC (left) and AUPRC (right). *P* values by two-sided Mann–Whitney *U* test are shown on the top. Horizontal line is median; hinges are 25th and 75th percentile; whiskers are most extreme value no further than 1.5× IQR. **c**, **d**, Scatter plots compare the performance between deltaSVM (*y*-axis) and ΔPWM (*x*-axis) derived by multinomial models (**c**) and BEESEM models (**d**) by in predicting allelic binding of 129 TFs for which both models were available. Results from fivefold cross-validation were shown. The values in both axes were AUPRC. **e**, An overview of the SNP-SELEX experimental procedure describing the novel batch of SNP-SELEX. **f**, A scatter plot compares the performance between deltaSVM (*y*-axis) and BEESEM-generated ΔPWM (*x*-axis) in predicting allelic binding of 87 TFs for which both models are available by the novel batch of SNP-SELEX. The values in both axes are AUPRC. **g**, The logo describes the PWM model of a homodimeric binding pattern of TF HLF, with the monomeric half-site indicated by the purple arrows. The red boxes indicate the positions at which the SNP rs79124498 is located (left) and its co-dependent base position (right). The *y*-axis corresponds to the information content at each position of the PWM (*x*-axis).

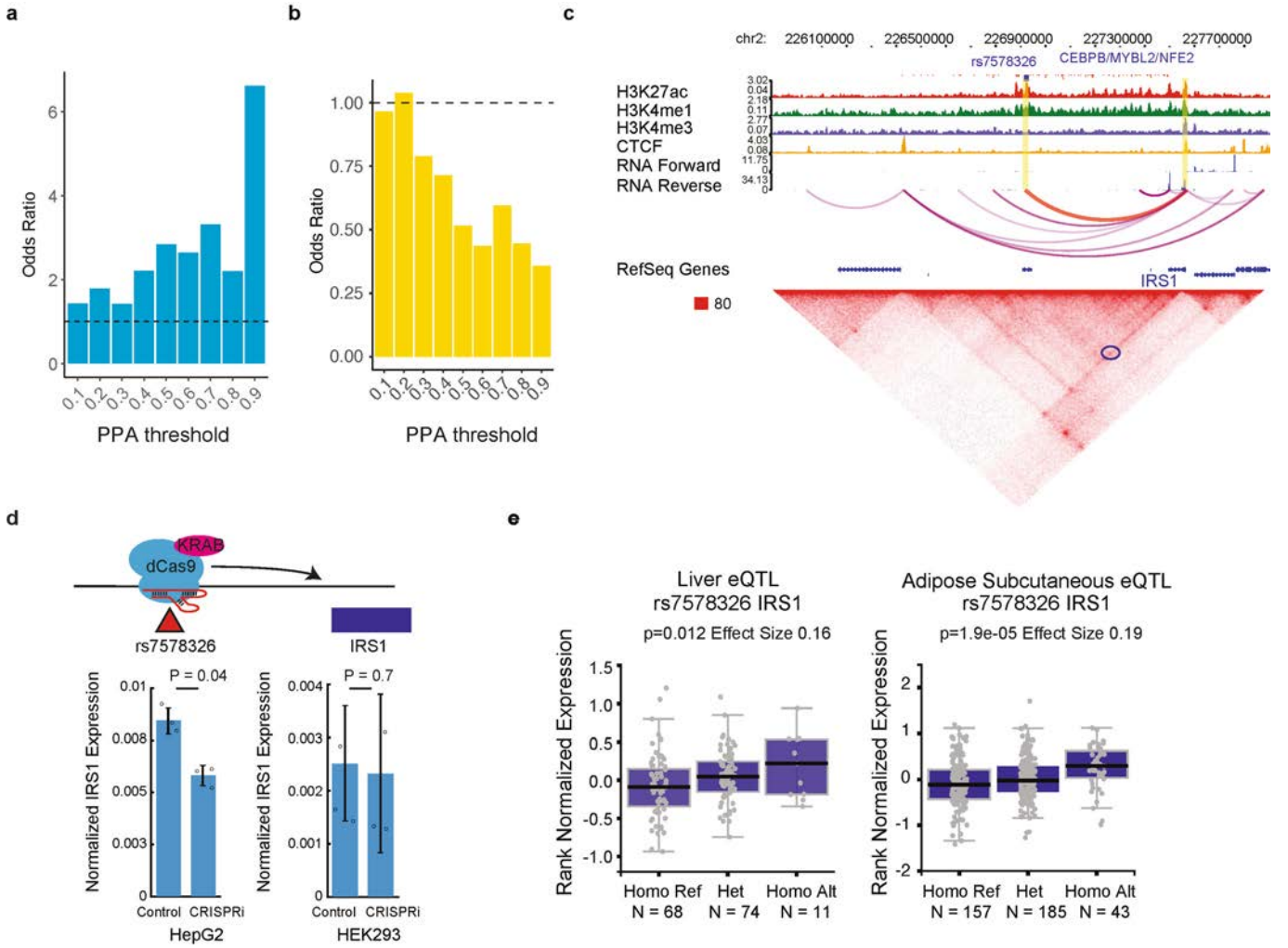


Extended Data Fig. 7 | Comparison of deltaSVM models and ΔPWM in predicting allelic TF binding in weak and strong TF binding sites. SNPs are categorized into five quantiles based on the OBS of the 40-bp DNA fragments. The performance of ΔPWM (green) and deltaSVM (orange) in predicting allelic binding of TFs was evaluated for SNPs in each category. **a**, **b**, Two batches of pbSNPs were used as gold standards for performance assessment: the pbSNPs from the initial SNP-SELEX experiments, with fivefold cross-validation (**a**) and the novel batch SNP-SELEX data (**b**). Both AUROC (upper) and AUPRC (lower) are shown for statistic assessment of the model performance. The first quantile represents SNPs with the weakest binding strength and the fifth quantile represents SNPs with the strongest binding strength. Horizontal line is median; hinges are 25th and 75th percentile; whiskers are most extreme value no further than 1.5× IQR.

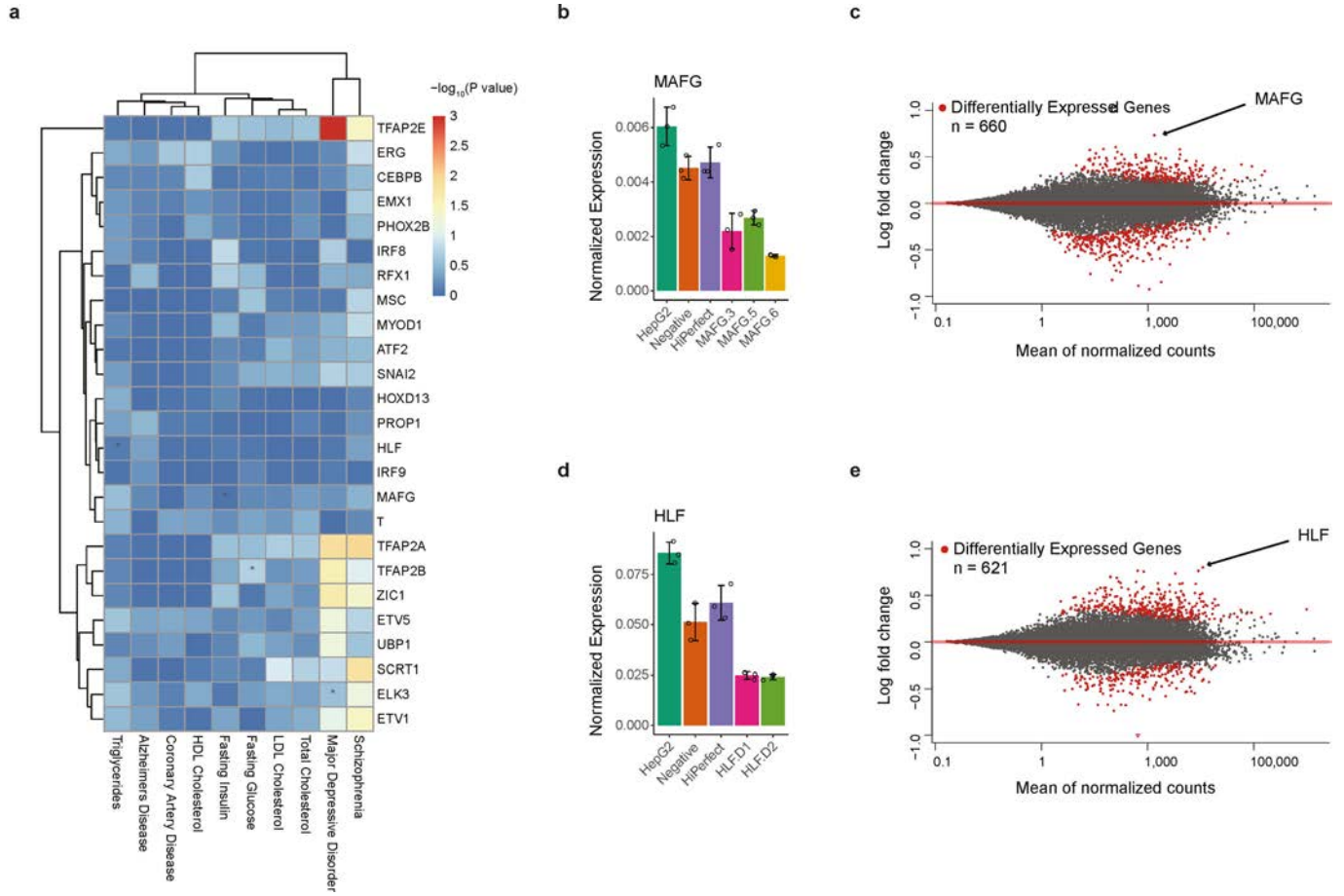


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | DeltaSVM models predict more accurately the noncoding variants affecting TF binding in vivo than ΔPWM. a, DeltaSVM models outperform ΔPWM in predicting differential DNA binding in vitro. Precision-Recall curves were used to assess the performance of either model in predicting allelic binding events identified in SNP-SELEX for three TFs, including ATF2, HLF, and MAFG. In all three cases, the performance of deltaSVM models (purple) was much better than that of ΔPWM (yellow). The AUC used for quantitative comparison was shown within each plot. **b**, Bar plots show the fractions of pbSNPs exhibiting allelic imbalance in TF ChIP–seq assays in HepG2 cells among all SNPs that were predicted to be differentially bound by a TF according to the deltaSVM models (purple) or the ΔPWM (yellow). The same datasets as in Fig. 3e were used. Only SNPs that were predicted to be bound by the TF were used in the comparison. The threshold for oligonucleotide binding and for the predicted pbSNPs was determined as the median score for the bound oligonucleotides and pbSNPs respectively. Error bars centred with mean percentage denote the 95% confidence interval calculated by Wilson method (Methods). For ΔPWM, $n = 2872$(ATF2); $n = 4134$(HLF); $n = 100$(MAFG). For deltaSVM, $n = 115$(ATF2); $n = 355$(HLF); $n = 16$(MAFG). **c**, Bar plots show the fractions of pbSNPs exhibiting allelic imbalance in TF ChIP–seq assays in GM12878 cells among all SNPs that were predicted as differentially bound by a TF according to the deltaSVM models (purple) or the ΔPWM (yellow). Three TFs were included in the analyses, ATF2, NR2F1, and PKNOX1. Only SNPs that were predicted to be bound by the TF were used in the comparison. The threshold for oligonucleotide binding and the predicted pbSNPs was determined as the median scores for the bound oligos and pbSNPs respectively. Error bars centred with mean percentage denote the 95% confidence interval calculated by Wilson method (Methods). For ΔPWM, $n = 4318$(ATF2); $n = 673$(NR2F1); $n = 225$(PKNOX1). For deltaSVM, $n = 142$(ATF2); $n = 229$(NR2F1); $n = 142$(PKNOX1). **d**, Similar to Fig. 3e, deltaSVM models outperform ΔPWM in predicting differential DNA binding in vivo. Three TF ChIP–seq datasets from GM12878 cells were used for the comparison, including the same dataset as shown in **b**. Elbow plots show that for each TF, the top-ranked allelic SNPs predicted by deltaSVM models were found to have allelic imbalance in ChIP–seq assays performed in GM12878 cells (purple). By contrast, for allelic SNPs predicted by ΔPWM, only a small fraction showed allelic imbalance in vivo (yellow).



Extended Data Fig. 9 | T2D risk SNPs are enriched for pbSNPs. a, Bar plots show the enrichment of pbSNPs in T2D risk SNPs identified from an independent study[14]. The levels of enrichment were displayed for different groups risk SNPs categorized based on the PPA (Posterior Probability of Association). Note that SNPs with stronger PPAs and thus higher likelihood of being causal for T2D are more likely to be pbSNPs. **b**, Bar plots show the enrichment of T2D risk SNPs in allelic TF binding SNPs predicted by PWM models using the same credible sets as Fig. 4a (ref. [13]). Specifically, SNPs with $P<0.01$ by atSNP were used as allelic TF binding SNPs. The level of association is categorized according to PPA as in **a**. Note that the likely causal SNPs with stronger T2D risk association no longer display higher enrichment for ΔPWM-predicted allelic SNPs. **c**, A T2D GWAS leading SNP rs7578326 and a pbSNP differentially bound by TFs CEBPB, CEBPE, MYBL2, and NFE2, is predicted to target the *IRS1* gene based on Hi-C analysis (circled in blue in bottom panel) in HepG2 cells. The locus around the SNP is enriched for H3K27ac and H3K4me1. **d**, CRISPRi using dCas9 fused with repressive KRAB domain and guide RNA targeting the locus of SNP rs7578326 (upper) leads to reduced expression of *IRS1* gene in HepG2 but not in HEK 293T cells. qPCR results from three biological replicates in HepG2 (left) and HEK 293 (right) cells are plotted in the bottom panel. Y-axis shows the power transformed values of expression presented as mean ± s.d. Raw data are shown as small black circles for clarification. *P* values computed using two-sided *t*-test are noted on the top. **e**, SNP rs7578326 is an eQTL in liver and adipose tissues. Normalized expression value from GTEx project for *IRS1* gene is grouped based on individuals' genotype of SNP rs7578326. Linear regression *P* values and effect sizes are noted on the top. Horizontal line is median; hinges are 25th and 75th percentile; whiskers are most extreme value no further than 1.5× IQR.



Extended Data Fig. 10 | Candidate TFs involved in complex traits and diseases identified by enrichment of TF binding alone. a, A heat map shows the significant enrichment of SNPs predicted to be located within TF-DNA binding sites among traits- or disease-associated SNP. The colour key is shown, and the value represents the -$\log_{10}$ *P* value. TF-trait pairs mentioned in the text were marked with *. Note that the SNPs here do not necessarily affect TF binding affinity. The candidate regulator we observed and validated (Fig. 4b) could not be identified here if we only use the presence of SNPs at the binding sites without taking into account the effect of SNP on binding affinity. **b**, **d**, qPCR results from three biological replicates of MAFG (**b**) and HLF (**d**) in WT (HepG2), Control (Negative and HiPerfect), and cells treated with different siRNAs. Expression values are presented as mean ± s.d. **c**, **e**, MA-plot showing differentially expressed genes comparing MAFG knockdown (**c**) and HLF knockdown (**e**) versus controls. Significant differentially expressed genes (FDR <0.2) were marked in red.

nature research

Corresponding author(s): Jian Yan, Jussi Taipale and Bing Ren

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No code was used for data collection.
Data analysis	Softwares used in this study are listed below. More detailed information can be found in Methods. bwa 0.7.8-r455 picard 1.131 samtools 1.3 GATK 3.6-017-19 macs2 2.1.1.20160309 STAR 2.4.2a Cufflinks 2.2.1 htseq-count 0.6.0 Beagle 4.1 Juicer tools 1.6.2 HapCUT2 ldsc v1.0.0 edgeR 3.14.0 DESeq2 1.12.4 BEESEM 1.0.1

binom 1.1-1
PRROC 1.3.1
homer v4.7.2
atSNP 1.0.0

Custom codes used to process and generate the results described in the current study were deposited into GitHub at [https://github.com/ren-lab/snp-selex].

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data generated in this study can be accessed via Gene Expression Omnibus (GEO) under accession number GSE118725. The raw sequencing data of TF ChIP-seq of GM12878 is extracted from the ENCODE portal [https://www.encodeproject.org]. The specific TF data can be accessed by searching the accession ID listed in Supplementary Table 4.

The web portal [http://renlab.sdsc.edu/GVATdb/] provides a searchable interface for SNPs and TFs tested in the current study.

Enriched motifs for SNP-SELEX experiments using Homer are available in Supplementary File S1. Scores for all tested SNP-TF pairs by SNP-SELEX experiments are available in Supplementary File S2. The 94 high-confidence deltaSVM models predicted allelic binding of all common SNPs in the human genome are available in Supplementary File S3.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes (n) were chosen to provide sufficient confidence to conduct statistical tests and were stated in respective figure legnds. The sequencing depth of SNP-SELEX libraries was set to ensure that at least 10,000 unique reads are available for each TF experiment.
Data exclusions	No exclusion was applied to the raw data uploaded in GEO. The failed SNP-SELEX experiments were excluded according to the QC criteria (see also Methods): Sequencing data of each SELEX cycle was aligned to the oligo library using BWA. Several filters were applied to aligned reads after alignments: 1) Reads of low quality, containing ambiguous bases, unaligned to reference and aligned outside of the oligo boundaries were filtered out and experiments with less than 10,000 reads were excluded from further analysis; 2) To control for PCR-duplication bias, the frequency of all PCR bias control (PDC) sequences (256 combinations) of each cycle were compared to the input library (cycle 0) using a linear regression model. PDC whose difference between expected and observed frequency exceeded 30% of the observed values were considered biased and all reads containing the biased PDC were removed. De novo motif discovery was then conducted using the cycle six reads with Homer toolset (Supplementary File S1). Motifs were then compared to JASPAR 2016 non-redundant vertebrates' motifs and SELEX models to examine quality of the experiments. Only SNP-SELEX experiments whose motif models match either its TF or TF of same structural family were kept for further analysis. No exclusion was applied for histone ChIP-seq. For TF ChIP-seq, peaks were called using MACS2 and motif enrichment was conducted with Homer toolset. Only experiments with peaks matching known motifs were used.
Replication	Biological or technical replicates were performed to ensure replication for SNP-SELEX, STARR-seq, RNA-seq, Hi-C and CRISPRi experiments. Results for SNP-SELEX were shown to be reproducible between replicates and DBD/FL from the same TF. For RNA-seq, and STARR-seq, we have conducted three independent replicative experiments and Pearson correlation were calculated between replicates to ensure high reproducibility. For Hi-C and HLF Knock-down RNA-seq, two independent replicative experiments were conducted.
Randomization	Samples were analyzed directly and individually, and not randomized to experimental groups.
Blinding	All analyses were performed using computational algorithms. Investigators were not blinded.

Reporting for specific materials, systems and methods

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
☒	☐ Unique biological materials	☐	☒ ChIP-seq
☐	☒ Antibodies	☒	☐ Flow cytometry
☐	☒ Eukaryotic cell lines	☒	☐ MRI-based neuroimaging
☒	☐ Palaeontology		
☒	☐ Animals and other organisms		
☒	☐ Human research participants		

Antibodies

Antibodies used: HLF (Santa Cruz, sc-134359), MAFG (Santa Cruz, sc-166548 X), Histone H3K4me1 (Abcam, ab8895), H3K4me3 (Abcam, ab8580), H3K27ac (Abcam, ab4729), and CTCF (Santa Cruz, sc-15914 X).

Validation: All the antibodies were commercially available and validated by the manufacturers respectively.
1. HLF, Santa Cruz, sc-13459, https://www.scbt.com/p/hlf-antibody-4d8
2. MAFG, Santa Cruz, sc-166548 X, https://www.scbt.com/p/maff-g-k-antibody-d-12
3. Histone H3K4me1, Abcam, ab8895, https://www.abcam.com/histone-h3-mono-methyl-k4-antibody-chip-grade-ab8895.html
4. H3K4me3, Abcam, ab8580, https://www.abcam.com/histone-h3-tri-methyl-k4-antibody-chip-grade-ab8580.html
5. H3K27ac, Abcam, ab4729, https://www.abcam.com/histone-h3-acetyl-k27-antibody-chip-grade-ab4729.html
6. CTCF, Santa Cruz, sc-15914 X, https://www.scbt.com/p/ctcf-antibody-c-20

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s): ATCC (Hep G2 and HEK293T)

Authentication: Neither of them were authenticated.

Mycoplasma contamination: All were tested negative.

Commonly misidentified lines (See ICLAC register): No commonly misidentified cell lines.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

All sequencing datasets used in this study have been deposited on Gene Expression Omnibus with the accession number GSE118725. (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118725).

The web portal [http://renlab.sdsc.edu/GVATdb/] provides a searchable interface for SNPs and TFs tested in the current study.

Files in database submission: List of datasets used is in Supplementary Table 4.

Genome browser session (e.g. UCSC): http://epigenomegateway.wustl.edu/legacy/?genome=hg19&session=PJW8rY7chj&statusId=1526985429

Methodology

Replicates: ChIP-seq for TF and histone marks was performed w/o replicates.

Sequencing depth: Number of reads for each experiment are listed in Supplementary Table 4.

Antibodies: HLF (Santa Cruz, sc-134359), MAFG (Santa Cruz, sc-166548 X), Histone H3K4me1 (Abcam, ab8895), H3K4me3 (Abcam,

Antibodies	ab8580), H3K27ac (Abcam, ab4729), and CTCF (Santa Cruz, sc-15914 X).
Peak calling parameters	Reads were aligned using BWA MEM with either single-end or pair-end model to the hg19 reference genome. Reads with low mapping quality (mapq<10) were filtered out, and PCR duplicates were removed using Picard tool (http://broadinstitute.github.io/picard/). MACS2 were then applied to call peaks and generate signal tracks to view in the genome browser with default parameters.
Data quality	For TF ChIP-seq, motif were called using homer and compared to known motifs. For histone ChIP-seq, sufficient sequencing depth were achieved according to ENCODE standards. Number of peaks for each antibody was compared to published data of corresponding antibody.
Software	BWA, samtools, picard, MACS2, homer.