Genetic effects on liver chromatin accessibility identify disease regulatory variants

Kevin W. Currin,[1] Michael R. Erdos,[2] Narisu Narisu,[2] Vivek Rai,[3] Swarooparani Vadlamudi,[1] Hannah J. Perrin,[1] Jacqueline R. Idol,[2] Tingfen Yan,[2] Ricardo D'Oliveira Albanus,[3] K. Alaine Broadaway,[1] Amy S. Etheridge,[4] Lori L. Bonnycastle,[2] Peter Orchard,[3] John P. Didion,[2] Amarjit S. Chaudhry,[5] NISC Comparative Sequencing Program,[2] Federico Innocenti,[4,7] Erin G. Schuetz,[5] Laura J. Scott,[6] Stephen C.J. Parker,[3,8] Francis S. Collins,[2,9] and Karen L. Mohlke[1,9,*]

Summary

Identifying the molecular mechanisms by which genome-wide association study (GWAS) loci influence traits remains challenging. Chromatin accessibility quantitative trait loci (caQTLs) help identify GWAS loci that may alter GWAS traits by modulating chromatin structure, but caQTLs have been identified in a limited set of human tissues. Here we mapped caQTLs in human liver tissue in 20 liver samples and identified 3,123 caQTLs. The caQTL variants are enriched in liver tissue promoter and enhancer states and frequently disrupt binding motifs of transcription factors expressed in liver. We predicted target genes for 861 caQTL peaks using proximity, chromatin interactions, correlation with promoter accessibility or gene expression, and colocalization with expression QTLs. Using GWAS signals for 19 liver function and/or cardiometabolic traits, we identified 110 colocalized caQTLs and GWAS signals, 56 of which contained a predicted caPeak target gene. At the *LITAF* LDL-cholesterol GWAS locus, we validated that a caQTL variant showed allelic differences in protein binding and transcriptional activity. These caQTLs contribute to the epigenomic characterization of human liver and help identify molecular mechanisms and genes at GWAS loci.

Introduction

Genome-wide association studies (GWASs) have identified thousands of loci associated with complex traits, but the vast majority of variants fall outside the coding region. As a consequence, the causal variants, molecular mechanisms, target genes, and tissues of action for most loci have not been characterized. Studies of gene expression quantitative trait loci (eQTLs) have been instrumental in identifying plausible target genes and tissues for GWAS loci.[1] Chromatin conformation capture techniques, such as Hi-C, have identified variants at GWAS loci that physically interact with gene promoters.[2] However, additional approaches are needed to further pinpoint functional variants and to identify how these variants alter gene expression.

Variants at GWAS loci are enriched in transcriptional regulatory elements, which are typically marked by chromatin accessibility, in trait-relevant tissues.[3] Recent studies have identified chromatin accessibility QTLs (caQTLs), many of which overlap transcription factor (TF) binding sites and motifs.[4–9] A subset of caQTLs are colocalized with eQTLs and GWAS loci, suggesting that variants at these loci influence gene expression and GWAS traits by altering chromatin accessibility.[4–9] However, caQTLs have been mapped in a limited set of human tissues. Mapping caQTLs in additional tissues and cell types is valuable to characterize the transcriptional regulatory mechanisms for a larger set of GWAS loci.

Liver is involved in numerous processes, including lipid metabolism, glucose storage, drug metabolism, and immune response.[10] Several studies have mapped eQTLs in liver tissue, and liver eQTLs are colocalized with GWAS loci for lipid, drug response, and other traits.[11–13] Lipid GWAS loci are enriched in regulatory chromatin states, including enhancers and promoters, in HepG2 hepatocytes.[14] QTLs for the active regulatory element histone marks H3K27ac and H3K4me3 have been identified in liver tissue, including a subset colocalized with liver eQTLs and GWAS loci.[12] Chromatin accessibility marks active regions containing H3K4me3 and H3K27ac, as well as poised promoters and enhancers that often do not display these histone marks.[15,16] Consequently, mapping caQTLs in liver tissue can help functionally characterize GWAS loci that act by altering gene expression in liver.

In this study, we jointly mapped genotypes, gene expression, and chromatin accessibility in liver tissue from 20 organ donors and identified caQTLs in liver tissue. We predicted the impact of caQTL variants on TF binding and predicted caQTL target genes using four approaches. Finally,

[1]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; [2]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA; [3]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA; [4]Eshelman School of Pharmacy and Center for Pharmacogenomics and Individualized Therapy, University of North Carolina, Chapel Hill, NC 27599, USA; [5]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; [6]Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA; [7]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA; [8]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA

[9]These authors contributed equally

*Correspondence: mohlke@med.unc.edu

https://doi.org/10.1016/j.ajhg.2021.05.001.

we used caQTLs, TF binding motifs, and target gene links to predict mechanisms at GWAS loci for multiple traits.

Material and methods

Liver tissue samples

Healthy human liver tissue was collected from 20 deceased organ donors through the National Institutes of Health Liver Tissue Cell Distribution System (LTCDS). Tissue was obtained from LTCDS and approved for use in this study as non-human subjects research by the Institutional Review Boards (IRBs) at St Jude Children's Research Hospital (Memphis, TN) and the University of North Carolina (Chapel Hill, NC).

Genotyping and imputation

We genotyped more than 2.5 million variants using the Infinium Omni2.5Exome-8 BeadChip array v1.3 (Illumina) at the NHGRI Genomics Core facility. Overall genotyping call rates ranged from 99.0%–99.6%. We mapped the Illumina array probe sequences to the hg19 genome assembly[17] using novoalign (see web resources), excluding variants with ambiguous probe alignments and variants with 1000 Genomes (1000G) phase 3 minor allele frequency (MAF) $> .01$ within 7 bp of the 3′ end of probes. No individuals were related at a 3rd-degree relationship threshold using KING v.1.4.[18] Prior to performing genotype principal component analysis (PCA), we removed variants with minor allele count < 4 and that were found within regions of unusually high linkage disequilibrium (LD, see web resources) using VCFtools v.0.1.14,[19] and selected distinct ($r^2 < 0.2$) variants using PLINK v.1.9.[20] We performed PCA of 59,674 genotypes using PLINK v.1.9[20] and found that each principal component (PC) explained essentially the same amount of variation (5%), and no PC explained a disproportionate amount of variation. Therefore, we did not include any genotype PCs as covariates when identifying caQTLs.

Prior to genotype imputation, we combined the genotypes of the samples in this study with genotypes from 177 samples from a separate study genotyped on similar chips and removed variants that met the following criteria: allele frequency difference > 20% with 1000G phase 3 Europeans, palindromic variants with MAF $> .2$, genotype missingness > 2.5%, and deviation from Hardy-Weinberg equilibrium ($p < 1 \times 10^{-4}$). Using the Michigan Imputation Server,[21] we phased 1,789,889 autosomal variants using Eagle v.2.3[22] and imputed missing genotypes using minimac3[21] with the Haplotype Reference Consortium (hrc.r1.1.2016) panel.[23] We retained variants with imputation $r^2 > .3$ for downstream analyses.

RNA-seq library preparation, read alignment, and selection of expressed genes

We extracted and purified total RNA from 20 frozen liver tissue samples using Trizol as previously described.[24] Paired-end, strand-specific, poly(A) RNA sequencing (RNA-seq) was performed on an Illumina NovaSeq 6000 with 2× 151 bp cycles. RNA-seq reads were trimmed using Trimmomatic[25] and aligned to the hg19 genome assembly[17] using STAR v.2.53[26] with default parameters. Using verifyBamID v.1.1.1,[27] we found no evidence of library contamination or sample swaps. Expression levels of GENCODE v.19[28] genes were quantified using QoRTs v.1.2.42.[29] We classified genes as expressed if the median transcripts per million (TPM) across the 20 individuals was at least 1. We performed principal component analysis on gene counts normalized by library size and variance-stabilized using DESeq2.[30] Principal components (PCs) were correlated against technical factors to identify covariates in downstream analyses (see "Correlation of caPeaks with promoter peaks and gene expression").

ATAC-seq library preparation

Nuclei were isolated as previously described[31] with the following modifications. We pulverized 50-mg pieces of frozen human liver tissue in liquid nitrogen using a Cell Crusher (CellCrusher), homogenized the tissue powder in ice-cold nuclei isolation buffer (NIB: 20 mM Tris-HCl, 50 mM EDTA, 5 mM spermidine, 0.15 mM spermine, 0.1% mercaptoethanol, 40% glycerol [pH 7.5]) using a 1-mL dounce for 40 strokes, and rotated for 5 min at 4°C. We filtered the solution through a Miracloth (Calbiochem), centrifuged at 1,100 × *g* for 10 min at 4°C, washed the pellet with 250 μL NIB containing 0.5% Triton-X, centrifuged at 500 × *g* for 5 min at 4°C, and resuspended the pellet in 250 μL of resuspension buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM $MgCl_2$ [pH 7.4]). After counting isolated nuclei, we pelleted 50,000 nuclei at 500 × *g* for 5 min at 4°C for each of three replicate ATAC-seq libraries per sample. Libraries were prepared using Nextera kits (Illumina) as previously described.[32]

ATAC-seq read alignment and identification of consensus peaks

We trimmed ATAC-seq reads to a uniform length of 126 bp using cutadapt[33] and aligned reads as previously described.[34] Briefly, we trimmed sequencing adapters using CTA (see web resources) and aligned reads to the hg19 human genome[17] using BWA-MEM (see web resources). We selected properly paired autosomal alignments with high mapping quality (mapq > 30) with samtools[35] and removed duplicate alignments using Picard (see web resources). We used ataqv[36] to generate ATAC-seq quality metrics and confirmed ATAC-seq libraries corresponded to the correct genotypes using verifyBamID.[27]

To assess reproducibility of libraries from the same individual, we called narrow peaks separately for each library using MACS2[37] with parameters –nomodel –shift -100 –extsize 200, then merged peaks across all individuals and replicates using BEDTools merge,[38] and selected peaks present in at least 3 libraries. We counted the number of reads overlapping each peak using featureCounts[39] and performed library size normalization and variance-stabilization using DESeq2.[30] We computed pairwise Pearson correlations of normalized counts for all peaks and for the 10,000 most variable peaks between libraries and visualized the results using the heatmap.2 function in the gplots R package[40] (see web resources). Libraries from the same individual were highly correlated, so we merged the alignment .bam files across libraries for each individual using SAMtools.[35]

To identify consensus peaks, we converted the merged .bam files for each individual to .bed files using BEDTools,[38] called narrow peaks for each individual using MACS2[37] with parameters –nomodel –shift -100 –extsize 200 –keep-dup all, and removed peaks overlapping blacklisted regions.[38,41] We then merged peaks across individuals using BEDTools[38] and defined consensus peaks as merged peaks that shared at least 1 base with a peak present in samples from at least 3 individuals.

Overlap of consensus peaks with roadmap chromatin states

We computed overlap of ATAC-seq consensus peaks with chromatin states in adult liver tissue from the Roadmap Epigenomics

Consortium.[3] We defined the following states: promoter (1_TssA, 2_TssFlnk, 3_TssFlnkU, 4_TssFlnkD, 14_TssBiv), transcribed (5_Tx, 6_TxWk), enhancer (7_EnhG1, 8_EnhG2, 9_EnhA1, 10_EnhA2, 11_EnhWk, 15_EnhBiv), polycomb (16_ReprPC, 17_ReprPCWk), heterochromatin (13_Het), ZNF repeats (12_ZNF/Rpts), and quiescent (18_Quies). For each consensus ATAC peak, we computed the fraction of bases that overlapped each chromatin state in liver tissue (Roadmap epigenome ID E066) using BEDTools coverage.[38] We assigned each peak to the chromatin state with which it shared the most bases, except for the quiescent state; we only assigned a peak to a quiescent state if all bases of a peak were found within a quiescent state. If a peak shared most, but not all, of its bases with a quiescent state, we assigned the peak to the state with the second highest coverage.

Selection of transcription factor motifs

We obtained transcription factor (TF) binding motifs from Cis-BP v.1.02,[42] selected all directly determined motifs per TF or the best inferred motif when a TF did not have a directly determined motif (TF_Information.txt dataset from Cis-BP), and restricted to motifs for TFs expressed in liver tissue from GTEx v.8 (median transcripts per million ≥ 1). We performed clustering to remove redundant motifs using RSAT matrix-clustering[43] with parameters -hclust_method average -calc sum -metric_build_tree Ncor -lth w 5 -lth cor 0.8 -lth Ncor 0.8 -quick, resulting in 516 motif clusters. For each motif cluster, we defined the representative TF as the TF with the highest expression in liver tissue from GTEx v.8 (measured in median TPM) and the representative motif as the motif assigned to the representative TF. If multiple motifs existed for the representative TF in a given cluster, we selected the motif with the highest information content. Although we often use the representative TF name to refer to motif clusters for convenience, any TF in the cluster may bind at a given locus. Therefore, we listed all expressed TFs in the cluster in supplemental tables. Some TFs were assigned as the representative TF for multiple clusters, potentially representing distinct binding profiles for the same TF. We retained all of these clusters unless otherwise noted.

Enrichment of TF motifs and ChIP-seq binding sites in ATAC peaks

We tested for enrichment of 286 non-redundant transcription factor (TF) motifs in consensus ATAC peaks using Analysis of Motif Enrichment (AME)[44] with parameters –control –shuffle– –kmer 2 –scoring max –hit-lo-fraction 0.75. We classified motifs with E-value $< 1 \times 10^{-100}$ as significantly enriched. We derived the 286 motifs from the set of 516 non-redundant motifs (see "Selection of transcription factor motifs") by selecting the motif with the highest information content per TF.

We downloaded liver tissue ChIP-seq peaks for 17 TFs[45] from the ENCODE portal[46] (sample accession ENCDO882MMZ) and defined binding sites as the summit of the ChIP-seq peaks. We computed the number of binding sites overlapping consensus ATAC-seq peaks for each TF using BEDTools intersect.[38] To determine whether the number of binding sites overlapping ATAC peaks was more than expected given their genomic frequency, we permuted binding sites across the genome 1,000 times excluding blacklisted regions[41] using BEDTools shuffle[38] and computed the number of overlaps for each permutation. We calculated an enrichment p value by determining the fraction of permuted overlaps that were equal to or greater than the observed number of overlaps.

Enrichment of heritability in ATAC peaks

Using stratified LD score regression as implemented in LDSC v.1.0.1,[47] we tested whether liver ATAC peaks were enriched for heritability of 13 GWAS traits: liver enzymes traits alanine aminotransferase (ALT),[48] alkaline phosphatase (ALP),[48] and gamma-glutamyl transferase (GGT);[48] cardiometabolic traits body mass index,[49] high-density lipoprotein cholesterol (HDL),[14] low-density lipoprotein cholesterol (LDL),[14] triglycerides,[14] total cholesterol,[14] coronary artery disease,[50] waist-hip ratio adjusted for body mass index (WHRadjBMI),[49] and type 2 diabetes;[51] and two negative control traits likely less relevant to liver, height[52] and rheumatoid arthritis[53] (see web resources). We computed LD scores for liver ATAC peaks using LDSC with 1000G phase 3 European LD and restricting to HapMap3 SNPs. We computed partitioned heritability of the ATAC peaks using LDSC correcting for the baseline v.1.2 model, which consists of 53 annotations.[47] We report heritability enrichment as the proportion of heritability explained by SNPs within ATAC peaks divided by the proportion of SNPs within ATAC peaks and classify enrichments with enrichment p value (enrichment_p) <0.05 as significant.

Chromatin accessibility QTL identification

We identified caQTLs using RASQUAL,[5] which jointly tests for association of genotype with peak accessibility across individuals and allelic imbalance in read counts at heterozygous variants within the same individual. We selected ~4 million genetic variants with MAF > 0.1 in the 20 individuals and within 100 kb of consensus peak centers and then restricted to variants present in 1000G phase 3 Europeans. To quantify peak accessibility across samples, we extended alignments 100 bp from either end of the 5′-most base using BEDTools[38] and counted the number of alignments overlapping each peak using featureCounts.[39] We did not use WASP[54] to remove reads exhibiting allelic mapping bias because RASQUAL models and accounts for allelic mapping bias.[5] We used DESeq2 size factors[30] to adjust for library size and the gcCor.R script provided with RASQUAL[5] to adjust for GC bias. To identify global variation between samples that may confound caQTL detection, we performed PCA on peak counts adjusted for library size and variance-stabilized by DESeq2.[30,40] We ran RASQUAL using differing numbers of PCs as covariates ranging from 0 to 10 in increments of 1 and selected 2 PCs to maximize the number of peaks with a caQTL at false discovery rate (FDR) of 5%. We performed multiple testing correction using the two-step eigenMT-BH procedure.[55] First, we used eigenMT[56] with the 1000G phase 3 European reference panel to adjust for the differing variant density around each peak, taking into account the LD between variants. Second, we selected the most significant eigenMT-adjusted p value for each peak and calculated FDR using the Benjamini-Hochberg (BH) procedure.[57] We selected significant caQTLs with FDR $< 5\%$ and correlation r^2 between prior and posterior genotypes > 0.8. We refer to peaks with a significant caQTL as caPeaks. We repeated the caQTL analysis using ~0.6 million variants within 1 kb of peak centers. Unless otherwise noted, all downstream analyses were performed using caQTLs identified using variants within 1 kb of peak centers.

Identification of caQTLs strongly influenced by one sample

To identify caQTLs strongly influenced by one sample, we separately removed each sample from the analysis and re-identified caQTLs in the 20 sets of 19 samples. We used the same caQTL parameters as for all 20 samples, except that we reduced the minimum MAF threshold to 0.05 to retain variants with MAF of 0.1 in the 20 samples. We restricted analyses to the lead variant-peak pairs detected in the 20-sample analysis. Given our small sample size, we would expect some caQTLs to no longer be significant when one sample is removed due to power even if no influential samples are present. Therefore, we defined caQTLs that are strongly influenced by one specific sample as caQTLs that no longer meet the FDR < 5% threshold (eigenMT-adjusted $p < 8.4 \times 10^{-4}$) only when one specific sample is removed, but remain significant when any other sample is removed.

ATAC-seq allelic imbalance and comparison to caQTL effect sizes

Instead of removing reads that exhibit allelic mapping bias, RASQUAL estimates and accounts for allelic mapping bias during QTL mapping.[5] To compare the RASQUAL results to another strategy, we used an alternative method to remove reads exhibiting allelic mapping bias and calculate allelic imbalance (AI). We removed ATAC-seq reads exhibiting allelic mapping bias using the WASP mapping pipeline[54] and counted the number of ATAC-seq reads mapping to each allele at heterozygous variants using ASEReadCounter[58] with the option –min-base-quality 30. We removed variants that had aligned bases other than the two genotyped alleles and included variants with >10 total reads, >3 reads per allele, and that were heterozygous in >3 individuals. After pooling reads across individuals, each variant had a minimum of 30 total reads and 9 reads per allele. The average reference allele fraction across all heterozygous sites for each sample ranged from 0.502 to 0.505, and the average reference allele fraction after combining samples was 0.503, indicating that little to no systematic allelic mapping bias remains. We fit allele counts to a beta-binomial distribution using the VGAM R package,[40,59] tested for AI using a two-tailed beta-binomial test, and adjusted for multiple testing using the BH procedure.

To compare effect sizes of AI variants and caQTL signals, we selected caQTLs that had at least one AI variant in strong LD ($r^2 > 0.8$, 1000G phase 3 Europeans) with the caQTL lead variant and that resided within the caPeak; LD was calculated using PLINK v.1.9.[20] For each caQTL with a linked AI variant, we selected the AI variant with the strongest evidence of imbalance (smallest beta-binomial p value). For both methods, we calculated an effect size by subtracting 0.5 from the estimated fraction of reads containing the alternate allele, which is the RASQUAL PI value for caQTLs. An alternate allele fraction of 0.5 corresponds to an equal number of reads for each allele, which is an effect size of 0. We then computed the Pearson correlation between the absolute value of effect sizes between the caQTLs and AI variants.

Colocalization of caQTL and H3K27ac QTL signals

We retrieved QTLs for 921 histone 3 lysine 27 acetylation (H3K27ac) peaks (termed H3K27ac QTLs, FDR < 5%, n = 18) from a recent report.[12] We only tested for colocalization between QTL signals where the caPeak and H3K27ac peak overlapped (defined as sharing at least one base). We calculated LD and haplotype phase between H3K27ac QTL and caQTL lead variants using PLINK[20] v.1.9 and classified signals as colocalized if these lead variants exhibited strong pairwise LD ($r^2 > 0.8$, 1000G phase 3 Europeans). We calculated effect sizes for caQTLs and H3K27ac QTLs by subtracting 0.5 from the RASQUAL PI values. We then computed the Pearson correlation between the absolute value of caQTL and H3K27ac QTL effect sizes.

caQTL enrichment in chromatin states

To identify which regulatory elements preferentially contain caPeaks, we compared the number of caPeaks (FDR < 5%) and non-caPeaks (eigenMT-adjusted $p > 0.5$) assigned to various liver tissue chromatin states from Roadmap.[3] We tested whether caQTL variants were enriched in specific liver tissue chromatin states relative to variants matched for MAF, number of LD proxies, and distance to nearest gene using the logistic regression model implemented in GARFIELD.[60] We defined caQTL variants as significantly enriched in a chromatin state if the p value for the logistic regression beta was less than the Bonferroni-corrected threshold (alpha of 0.05 for 7 chromatin states) of 7.1×10^{-3} and the odds ratio was greater than 1. We defined caQTL variants as significantly depleted in a chromatin state if $p < 7.1 \times 10^{-3}$ and odds ratio < 1.

Overlap of caPeaks with macrophage ATAC peaks

We retrieved a set of 296,220 ATAC peaks mapped across macrophages exposed to four experimental conditions: naive, IFNγ stimulation, *Salmonella* infection, and both exposures[4] (see web resources). To compare peak positions, we used liftOver[61] with the option -minMatch = 0.75 to convert the 3,123 liver caPeaks from GRCh37 (hg19) to GRCh38[17] coordinates. We identified liver caPeaks that overlapped (defined as sharing at least 1 base) with a macrophage peak using BEDTools intersect.[38] We also applied liftOver to the macrophage peaks and obtained the same results.

Transcription factor motif disruption by caQTL variants

We selected 5,378 caQTL variants that resided within a caPeak using BEDTools intersect[38] and that were in strong LD ($r^2 > 0.8$, calculated with PLINK[20]) with the caQTL lead variant. To ensure that each motif occurrence was disrupted by only one variant, we removed 793 variants located within 30 bp of another caQTL variant, resulting in 4,585 variants. For both alleles of each caQTL variant, we extracted the nucleotide sequence for the region containing the variant and the 30 nucleotides on either side of the variant using the BEDTools slop and getfasta tools.[38] We scanned these sequences for occurrences of 516 non-redundant TF motifs using Find Individual Motif Occurrences (FIMO)[62] with parameters –thresh 0.01–max-stored-scores 1000000–no-qvalue–skip-matched-sequence –text and only retained motif occurrences that overlapped caQTL variant positions. For each motif-variant pair, we selected the strongest motif match (smallest p value) per allele and only retained motif occurrences that matched strongly to at least one allele ($p < 1 \times 10^{-4}$). If different motifs for the same representative TF overlapped the same variant, we selected the motif with the strongest match.

Similar to a recent study,[63] we quantified the difference in motif match between alleles of a variant using the log ratio of FIMO p values. The FIMO p value for a given motif occurrence is the probability of observing a motif occurrence with the same or greater score, which inherently accounts for differences in score distributions between different motifs. For a given variant-motif pair, we define motif disruption as $\log_{10}(p_{aw}) - \log_{10}(p_{as})$, where p_{aw} and p_{as} are the FIMO p values for the alleles with the weaker and stronger motif match, respectively. As motif disruption is always positive, we classified a motif as disrupted if motif disruption was > 1, corresponding to a 10-fold difference in the FIMO p values between alleles.

We identified motifs whose disruption was associated with caQTL status using logistic regression. To generate a set of non-caQTL variants, we first selected peaks with no evidence of genetic regulation (caQTL eigenMT-adjusted $p > 0.5$), that overlapped at least one variant tested in the caQTL analysis and that were similar to caPeaks in GC content (±5%), peak width (±20%), and distance to nearest transcription start site (TSS) of a protein-coding gene in GENCODE[28] (±20%). We identified 10 non-caPeaks for >99% of the caPeaks used in the motif disruption analysis and defined non-caQTL variants as the 50,054 variants that were within non-caPeaks and were located more than 30 bp from the nearest variant. We tested these non-caQTL variants for TF motif disruption using the same procedure as for caQTL variants and restricted analysis to the 109 motifs with at least 20 disruptions by caQTL variants. For each representative TF, we selected the motif with the most disruptions by caQTL variants to ensure that we used only one motif per representative TF. We then regressed caQTL status (1 = caQTL, 0 = non-caQTL) against motif disruption status (1 = disrupted, 0 = not disrupted) for each motif-variant pair using logistic regression. We classified motif disruption as associated with caQTL status if the p value for the logistic regression beta was less than the Bonferroni-corrected threshold (alpha of 0.05 for 109 motifs) of 4.6×10^{-4}. Because residual differences may exist in peak GC content, width, and distance to nearest protein coding TSS, we performed logistic regression with and without these features as covariates and obtained the same set of significantly enriched motifs.

caPeak target gene identification

We used four methods to identify target genes for caPeaks: proximity to a gene's TSS, overlap of caPeaks with promoter-centered chromatin contacts, correlation of caPeaks with peaks at gene promoters or with gene expression, and colocalization of caQTLs and eQTLs. We excluded genes from the analysis if their Entrez ID did not map to exactly one Ensembl ID (eQTL data only) or if their symbol (common name) didn't map to exactly one Ensembl ID. When combining results across the four methods, we matched genes based on Ensembl ID.

TSS proximity

We classified a caPeak as TSS proximal if it was located within 2 kb upstream and 1 kb downstream of the TSS of any of the 13,782 expressed genes (median TPM > 1) in our 20 liver samples using BEDTools closest.[38]

Promoter-centered chromatin contacts

We obtained promoter-distal and promoter-promoter contacts mapped in liver tissue using promoter capture Hi-C from a recent study[2] (see web resources). Using described filtering criteria,[2] we selected contacts with p value < 0.01 and interaction frequency ≥ 5. We identified caPeaks overlapping distal ends of promoter-distal contacts or either end of promoter-promoter contacts using BEDTools intersect.[38]

Correlation of caPeaks with promoter peaks and gene expression

We classified an ATAC-seq peak as the promoter peak for an expressed gene if it was the closest peak to the TSS of the gene and it was within 2 kb upstream and 1 kb downstream of the TSS.[64] A promoter peak may or may not be a caPeak. We identified promoter peaks for 10,074 of 13,782 expressed genes. For each gene with a promoter peak, we identified caPeaks for correlation that were within 1 Mb of the gene's TSS but that were not TSS proximal. For peak and gene counts, we performed library size normalization and variance-stabilization using DESeq2[30] and GC bias-correction using RASQUAL.[5] We additionally adjusted peak counts by the percent of high-quality autosomal alignments (HQAA) in peaks (a measure of ATAC signal-to-noise), which was strongly correlated with the first ATAC-seq PC, and gene counts by the percent of reads mapping to the most expressed gene and the percent of reads mapping to the top 10 most expressed genes (geneDiversityProfile_top1pct and geneDiversityProfile_top10pct metrics from QoRTs[29]), which were strongly correlated with RNA-seq PCs 1 and 2, respectively, using the limma removeBatchEffects function.[65] We then computed the Spearman correlation between (1) gene expression and caPeaks and (2) promoter peaks and caPeaks using the cor.test function in R.[40] We adjusted for multiple testing using the BH procedure[57] and classified correlations with FDR < 5% as significant.

Colocalization of caQTLs and eQTLs

We obtained liver tissue expression quantitative trait loci (eQTLs) for 15,668 genes (FDR < 5%) from a meta-analysis of 1,183 individuals[11] and restricted to the 15,418 eQTLs on autosomes. We calculated LD and haplotype phase between eQTL and caQTL lead variants using PLINK[20] v.1.9 and classified signals as colocalized if these lead variants exhibited strong pairwise LD ($r^2 > 0.8$, 1000G phase 3 Europeans). To compare the direction of effect for colocalized caQTLs and eQTLs, we compared the sign of the caQTL effect size (RASQUAL pi statistic - 0.5) and the eQTL effect size (meta T statistic).

For the caQTL-eQTL colocalizations identified based on LD, we also assessed colocalization using the Bayesian approach implemented in coloc.[66] We ran coloc using p values and minor allele frequencies because regression coefficients and variances are not available from the RASQUAL model. coloc estimates five posterior probabilities (PP): no variant in the tested region affects either trait (PP0), a variant affects one trait but not the other (PP1 for caQTL and PP2 for eQTL), different variants affect each trait (PP3, no colocalization), and the same variant affects both traits (PP4, colocalization). We considered signals to show strong evidence of colocalization if PP4 > 0.8, suggestive evidence of colocalization if PP4 > 0.5, and evidence against colocalization if PP3 > 0.5. If the sum of PP0, PP1, and PP2 was > 0.5, we concluded that power was too low to assess colocalization. We note that coloc was designed to operate on results from linear regression or logistic regression[66] and may not be appropriate for the caQTL results generated from RASQUAL, which combines results from a negative binomial generalized linear model and tests of allelic imbalance.[5]

Colocalization of caQTL and GWAS signals

We downloaded the NHGRI-EBI GWAS catalog[67] on October 28, 2019, extracted only single variant associations, and converted variant genomic coordinates from GRCh38[17] to GRCh37 (hg19) using liftOver.[61] We extracted variants associated with 19 trait groups ($p < 5 \times 10^{-8}$) relevant to liver function and cardiometabolic diseases: liver enzymes, high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), total cholesterol (TC), triglycerides (TG), cardiovascular disease (CVD), hypertension/blood pressure (HTBP), type 2 diabetes (T2D), insulin, glucose, glycated albumin, serum albumin, glycated hemoglobin (HbA1c), C-reactive protein (CRP), bilirubin, body mass index (BMI), waist-hip ratio adjusted for BMI (WHRadjBMI), liver injury, and non-alcoholic fatty liver disease (NAFLD). We also included two negative control traits, height and rheumatoid arthritis, which presumably have less relevance to the liver. We extracted alleles for each variant from the dbSNP[68] build 151 common variant set (see web resources), restricting to bi-allelic variants. To select one variant per association signal, we performed LD clumping separately for each trait using swiss (see web resources); variants in strong LD ($r^2 > 0.8$, 1000G phase 3 Europeans) and within 1 Mb of a variant with a more significant p value were removed. We calculated LD between lead caQTL and GWAS variants using PLINK[20] v.1.9 and classified signals in high LD ($r^2 > 0.8$) as colocalized. We made LocusZoom plots for specific loci using LocusZoom v.1.4.[69]

To identify liver caQTL-GWAS colocalizations also observed in blood, we retrieved caQTLs mapped in macrophages exposed to four experimental conditions[4] and activated T cells[8] (see web resources). For macrophages, we downloaded the caQTL lead variant summary statistics and selected significant caQTLs at FDR < 10% using the same procedure described in the previous report,[4] and we converted the genomic coordinates from GRCh38 to hg19 using liftOver.[61] For T cells, we used the set of publicly available caQTLs at FDR < 5% mapped to hg19 coordinates.[8] For both datasets, we identified caQTL signals colocalized with GWAS signals using the procedure described above. We considered a liver caQTL-GWAS colocalization to be present in a blood cell type if the liver and blood caPeaks shared at least one base and if the lead variant of the blood caQTL was in strong LD ($r^2 > 0.8$) with the same GWAS variant as the liver caQTL. Blood caQTL lead variants were not tested for colocalization if variants were not in the 1000G LD reference panel.

Transcriptional activity reporter assays

HepG2 hepatocyte cells were cultured in MEM-alpha supplemented with 10% FBS and 1 mM sodium pyruvate, THP-1 monocyte cells were cultured in RPMI-1640 supplemented with 10% FBS, and both cell types were maintained at 37°C with 5% CO_2. To test haplotypic differences in transcriptional activity, we designed PCR primers (5′-TATGTTGCACAGGCTGGTCT-3′ and 5′-GGCAATAACGCCCACCTC-3′) to amplify a 666-bp DNA element (chr16:11,644,551–11,645,216) spanning the ATAC-seq peak and containing variants rs3784924, rs11644920, and rs57792815, and we generated PCR products using DNA from individuals homozygous for both haplotypes. We cloned the derived PCR products into luciferase reporter vector pGL4.23 (Promega) as described previously.[70] The day before transfection, we plated 120,000 HepG2 cells, and on the day of transfection, we plated 300,000 THP-1 cells. We transfected duplicate wells with four to five sequence-verified independent constructs for each haplotype. We co-transfected wells with phRL-TK Renilla reporter vector using lipofectamine 3000 (Life Technologies) following the manufacturer's protocol. To induce differentiation into macrophages,[71] we added 100 nM 1α,25-Dihydroxyvitamin D_3 (Sigma) to the THP-1 cells at the time of transfection. To obtain activated macrophages, we added 100 ng/mL lipopolysaccharides (Sigma) to vitamin D_3-treated cells 24 h after transfection and incubated cells for an additional 24 h. Firefly luciferase activity was measured 48 h post-transfection and normalized to Renilla activity to adjust for differences in transfection efficiency. Fold-changes in luciferase activity were calculated relative to an empty pGL4.23 vector, and statistical differences in activity were determined using two-tailed Student's t tests. We repeated transcriptional activity experiments on a separate day and obtained equivalent results.

Electrophoretic mobility shift assays (EMSAs)

We designed and annealed 3 biotin-labeled and unlabeled 17-bp complementary oligonucleotide probes centered on each of variants rs3784924, rs11644920, and rs57792815. We conducted EMSAs using the LightShift Chemiluminescent EMSA kit (Thermo Scientific) following the manufacturer's protocol. The binding reactions consisted of 6 μg HepG2 nuclear extract (NE-PER Kit, Thermo Fisher Scientific), 1 μg poly(dI-dC), 1x binding buffer, and 400 fmol biotinylated oligonucleotide as described previously.[70] To test the specificity of the protein complexes to each allele, we added 10-fold excess unlabeled probes. Protein-DNA complexes were resolved by gel electrophoresis and transferred and detected by chemiluminescence as described previously.[70] We repeated EMSA experiments on a separate day and obtained equivalent results.

Results

Joint profiling of gene expression and chromatin accessibility in human liver tissue

We obtained liver tissue from 20 deceased donors from the St. Jude liver bank (Table S1) and profiled gene expression using RNA-seq and chromatin accessibility using ATAC-seq[32] (Figure 1A). All RNA libraries had RNA integrity number (RIN) of at least 6.5, and the median RIN value was 8 (Table S2), indicating that we extracted high-quality RNA. We identified 13,782 expressed genes, 13,317 of which are on autosomal chromosomes (Tables S2–S4; Figure S1). By generating triplicate ATAC-seq libraries, we obtained an average of 204 million high-quality autosomal ATAC-seq alignments (HQAAs) per sample and all libraries had >13% of HQAAs within peaks and TSS enrichment > 4 (Tables S5–S7; Figure S2), indicating that we generated libraries from tissue with high signal-to-noise. We identified 223,265 consensus accessible chromatin regions (peaks) with median peak width of 617 base pairs (Figure 1B).

To predict the regulatory function of ATAC-seq peaks, we assigned peaks to liver tissue chromatin states from the Roadmap Epigenomics Project[3] and tested for enrichment of transcription factor (TF) binding sites and motifs in peaks. Among all 223,265 peaks, 34% were located in enhancers and 10% in promoters, and among the 50,000 most accessible peaks, ranked by median DESeq2 normalized count across individuals, 54% were located in enhancers and 38% in promoters (Figure 1C). These results indicate that the strongest peaks were mostly located in promoters and enhancers, as expected, but that weaker peaks observed in at least three individuals were located in less well-characterized regions. We found 90 TF motifs enriched in peaks (E-value $< 1 \times 10^{-100}$; Table S8), including motifs for HNF4G (MIM: 605966), FOXA family members (HNF3), CEBPB[72] (MIM: 189965), the multifaceted protein CTCF[73] (MIM: 604167), and KLF family members, which regulate numerous processes in liver.[74] Of 17 TFs with ChIP-seq data in liver tissue,[45] binding sites for all TFs were significantly enriched (permutation $p < 1 \times 10^{-3}$) in ATAC peaks (Table S9), and 11 TFs had over 90% of their binding sites within ATAC peaks (Table S9), similar

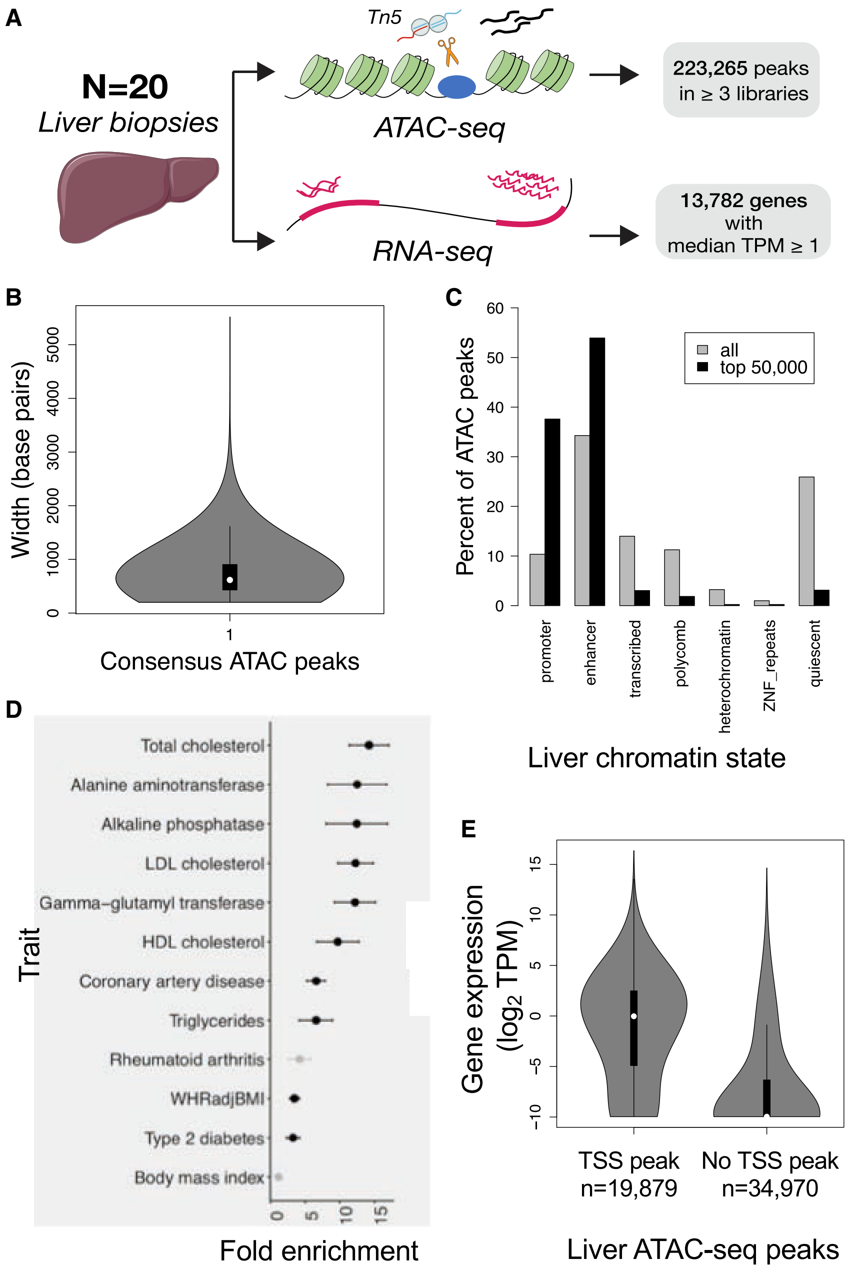


Figure 1. Joint profiling of gene expression and chromatin accessibility in human liver tissue

(A) RNA-seq and ATAC-seq was performed in liver samples from 20 donors.

(B) Distribution of consensus ATAC peak widths in base pairs.

(C) Percent of consensus ATAC peaks by chromatin state in liver tissue from the Roadmap Epigenomics Project. All peaks, gray; 50,000 most accessible consensus peaks, black; quiescent represents unannotated regions.

(D) Heritability enrichment of GWAS variants for multiple traits in all 223,265 liver ATAC peaks using stratified LD score regression. Points represent fold enrichment (proportion of heritability divided by proportion of SNPs within ATAC peaks) and error bars represent standard error. Significant enrichment (enrichment_p < 0.05), black; non-significant enrichment (enrichment_p > 0.05), gray.

(E) Comparison of the distribution of expression between genes with and without an ATAC peak overlapping the transcription start site (TSS).

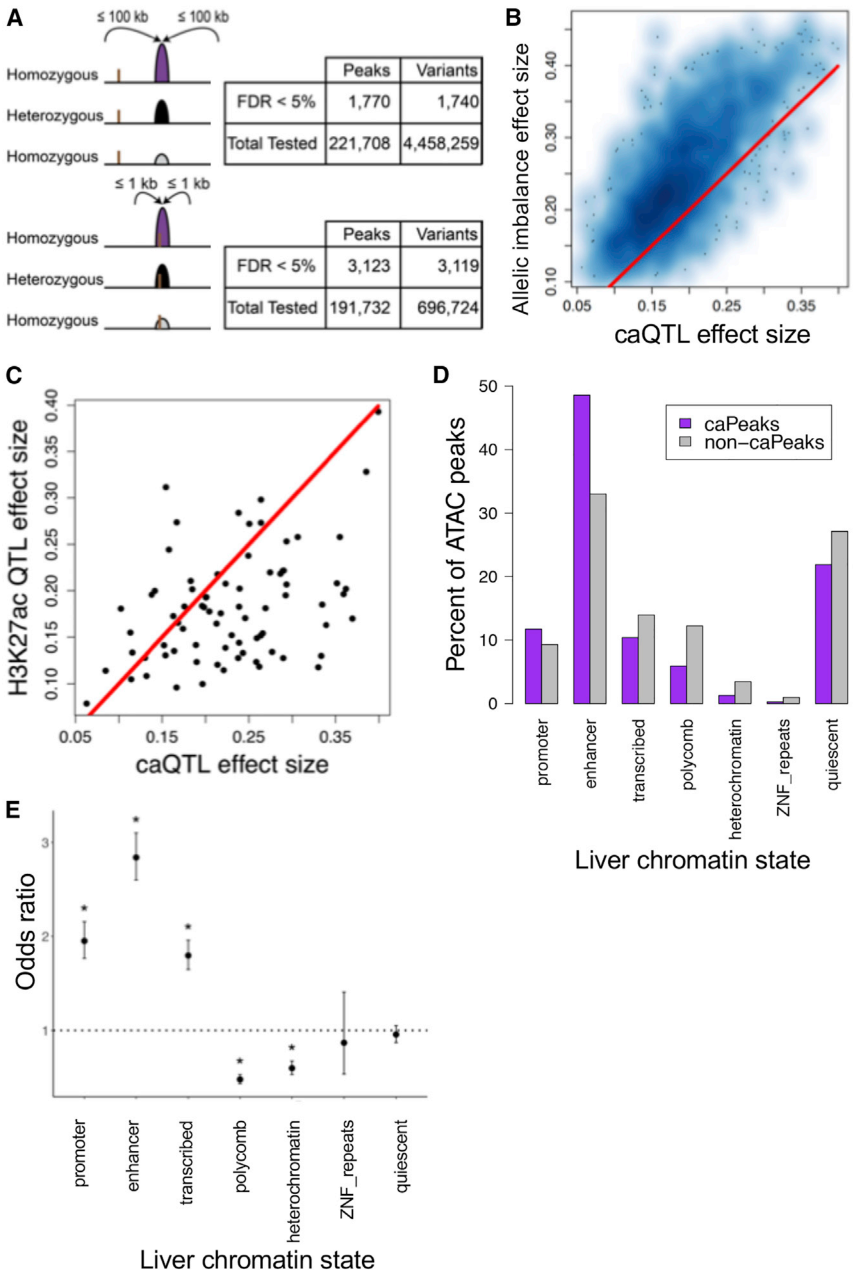
to previous findings.[15] Taken together, ATAC peaks marked previously annotated transcriptional regulatory elements and TF binding sites in liver tissue.

We tested whether liver ATAC peaks were enriched for heritability of liver-relevant traits using stratified LD score regression.[47] We observed significant heritability enrichment ($p < 0.05$) for 11 of 13 tested traits (Figure 1D), and total cholesterol displayed the strongest enrichment (enrichment = 14.2, $p = 7.2 \times 10^{-5}$). We also observed strong enrichments (fold enrichment > 10) for LDL cholesterol and the liver enzymes. Heritability enrichment for cholesterol traits in liver regulatory elements marked by H3K4me1 has been previously identified,[47] consistent with our results. As expected, we did not observe significant enrichment for rheumatoid arthritis and body mass index.[47] These results indicate that liver ATAC peaks are enriched for genetic variants associated with liver-relevant traits.

We next determined whether genes with ATAC peaks at their transcription start site (TSS) were more likely to be expressed compared to genes without TSS peaks. A larger proportion of expressed genes had an ATAC peak directly overlapping the TSS (9,904 of 13,317, 74%) compared to non-expressed genes (9,975 of 41,532, 24%). Similarly, genes with a peak at the TSS tended to have higher expression than genes without a peak at the TSS (Figure 1E; Kolmogorov-Smirnov test, $p < 2.2 \times 10^{-16}$). Together, the data provide high-quality gene expression and chromatin accessibility profiles in human liver tissue.

Identification of genetic variants associated with liver chromatin accessibility

We identified chromatin accessibility quantitative trait loci (caQTLs) using RASQUAL[5] and two distance thresholds: variants within 100 kilobases (kb) and within 1 kb of peak centers (Figures 2A and S3–S5; Tables S10–S12). Testing variants within 100 kb of peak centers, we identified significant caQTLs for 1,770 peaks (caPeaks), corresponding to 1,740 unique lead caQTL variants (Figure 2A; Table S11). For a substantial portion of caPeaks, the lead caQTL variant was within 1 kb of the caPeak center



	Peaks	Variants
FDR < 5%	1,770	1,740
Total Tested	221,708	4,458,259

	Peaks	Variants
FDR < 5%	3,123	3,119
Total Tested	191,732	696,724

Figure 2. Identification and characterization of caQTLs

(A) caQTLs identified using variants within 100 kb or 1 kb of peak centers.

(B) Comparison of effect sizes between caQTLs and simple allelic imbalance (Pearson's R = 0.75). The red line is the one-to-one line for caQTL effect sizes.

(C) Comparison of effect sizes between caQTLs and H3K27ac QTLs (Pearson's R = 0.40). The red line is the one-to-one line for caQTL effect sizes.

(D) Comparison of the number of caPeaks and non-caPeaks assigned to each chromatin state in liver tissue from the Roadmap Epigenomics Project. caPeaks, purple; non-caPeaks, gray; quiescent represents unannotated regions.

(E) Enrichment of caQTL variants in liver chromatin states. Error bars represent 95% confidence intervals. * indicates significant enrichment ($p < 0.0071$).

(n = 692, 39%, Figures S4B and S4C), and 654 of these 692 variants were within the caPeak. Testing variants within 1 kb of peak centers, we identified a significant caQTL for 3,123 peaks (Figure 2A; Table S12). We likely identified more caQTLs using a smaller window size because of a reduced multiple testing burden. We used this set of 3,123 caQTLs for all subsequent analyses unless noted otherwise.

We next tested whether any caQTLs were strongly influenced by a single sample. Of the 3,123 caQTLs, 355 were no longer significant when one specific sample was removed, but remained significant when any other sample was removed (Table S12). However, all but 6 remained nominally significant ($p < 0.05$). The most common influential sample (sample 459) accounted for only 48 of the 355 caQTLs (14%) and had the highest percent of HQAA within peaks (Table S12), indicating that this sample has high quality. Taken together, the vast majority of the caQTLs are not strongly influenced by one sample.

To compare the RASQUAL model to another method that accounts for allelic mapping bias, we used WASP to remove reads exhibiting allelic mapping bias[54] and then calculated AI. 1,912 (81%) caQTLs identified by RASQUAL exhibited nominal (beta-binomial $p < 0.05$) and 1,112 (47%) exhibited genome-wide AI (FDR < 5%), all with the same direction of effect as the caQTL (Table S13). Lead caQTL variants and representative AI variants exhibiting nominal AI showed strongly correlated effect sizes (Pearson's R = 0.75, Figure 2B). AI effect sizes tended to be larger than caQTL effect sizes (Figure 2B), possibly because AI was calculated using individual variants whereas caQTLs were identified using entire peaks. Therefore, we conclude that allelic mapping bias has no systematic effect on the caQTL results.

To determine the extent of shared genetic effects across different markers of transcriptional regulatory elements, we compared the 3,123 caQTLs to 921 H3K27ac QTLs from a recent report.[12] Of the 921 H3K27ac QTL peaks, 77 (8%) overlap a caPeak and have a lead variant in strong LD ($r^2 > 0.8$) with the caQTL lead (Table S14). The 77 colocalized caQTL-H3K27ac QTL signals all showed the same direction of effect, and their effect sizes were moderately correlated (Pearson's R = 0.40, Figure 2C). The largely distinct results may be due to the small sample sizes, analysis differences, and different genetic effects on the two epigenetic marks.

To predict the regulatory function of caPeaks, we compared caPeaks to liver tissue chromatin states from the Roadmap Epigenomics Consortium.[3] Relative to non-caPeaks (eigenMT-adjusted p > 0.5), caPeaks were more frequently located in enhancers (48.6% versus 33.0%) and promoters (11.7% versus 9.3%) (Figure 2D). caQTL variants were significantly enriched in enhancers (OR = 2.9), promoters (OR = 2.0), and transcribed regions (OR = 1.8) and depleted in polycomb (OR = 0.5) and heterochromatin (OR = 0.6) states, which are associated with gene repression and presumably inaccessible chromatin (Figure 2E, Table S15). Taken together, caQTLs showed strong overlap with active transcriptional regulatory elements, with particularly strong enrichment in enhancers.

To identify liver caQTLs that would not be identified in blood, we counted liver caPeaks that overlapped macrophage ATAC peaks,[4] using all macrophage ATAC peaks, not just caPeaks, due to limited sample sizes. Of the liver caPeaks, 1,268 (41%) overlapped a macrophage ATAC peak (Table S12), suggesting that 59% of liver caQTLs mark regulatory elements not present in macrophages. This estimate is likely conservative because we included macrophage ATAC peaks that do not have caQTLs and demonstrates the importance of mapping caQTLs in a diverse set of tissues.

Disruption of transcription factor binding motifs by caQTLs

One way genetic variants may alter chromatin accessibility is by disrupting TF binding sites.[5,6,8] Among 4,585 variants within a caPeak and in strong LD with the caQTL lead, 3,132 (68%) variants altered the binding affinity of a TF motif (Figure 3A; Table S16). Of the 2,793 caPeaks containing a variant, 2,249 (81%) contained at least one variant predicted to disrupt a motif, and 602 of these contained 2 or more predicted motif-disrupting variants. Motifs for many TFs were disrupted by multiple caQTL variants, with 109 TF motifs disrupted by 20 or more variants (Table S16). Disruption of motifs for 29 of these 109 TFs was significantly associated with caQTL status (log$_n$ OR > 0, p < 4.6 × 10^{-4}) (Figure 3B; Table S17), including TFs from the HNF, FOXA, and CEBP families,[72] CTCF, and ATF2 (MIM: 123811). FOXA and CEBP factors can act as pioneer factors by binding to inaccessible chromatin and initiating the establishment of accessible chromatin[75] and ATF2 can

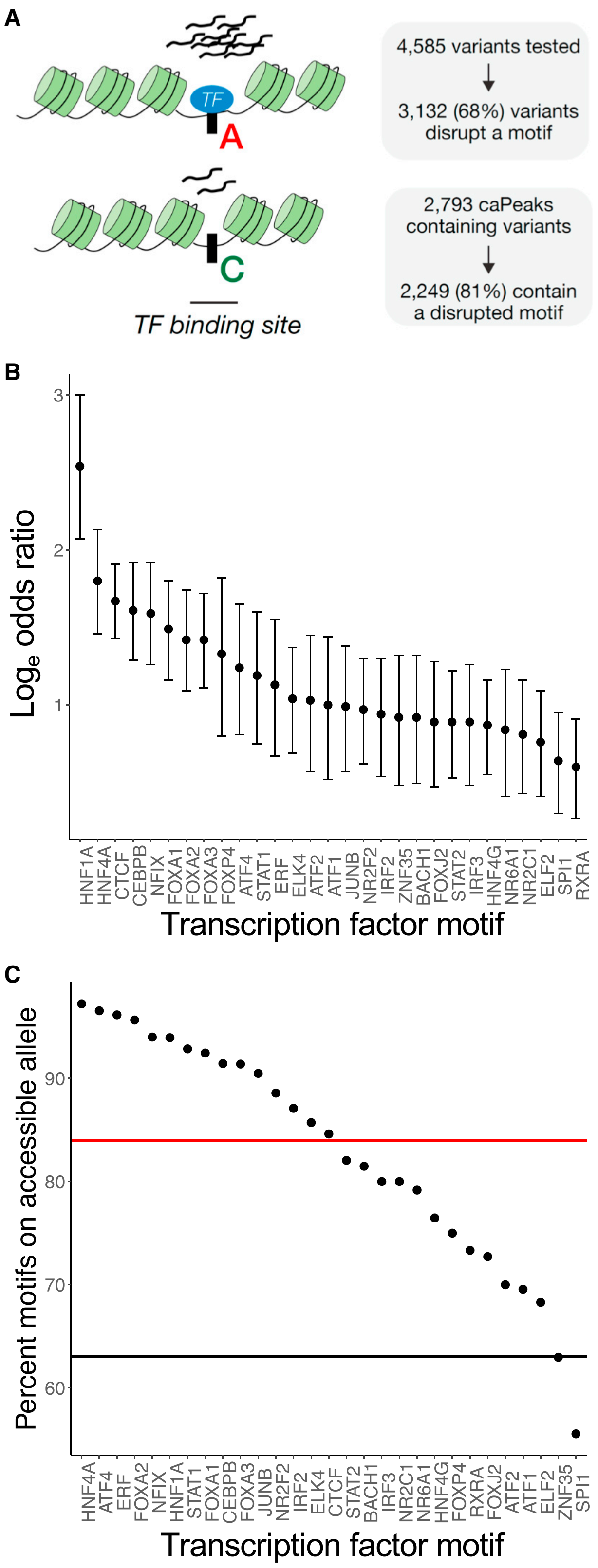
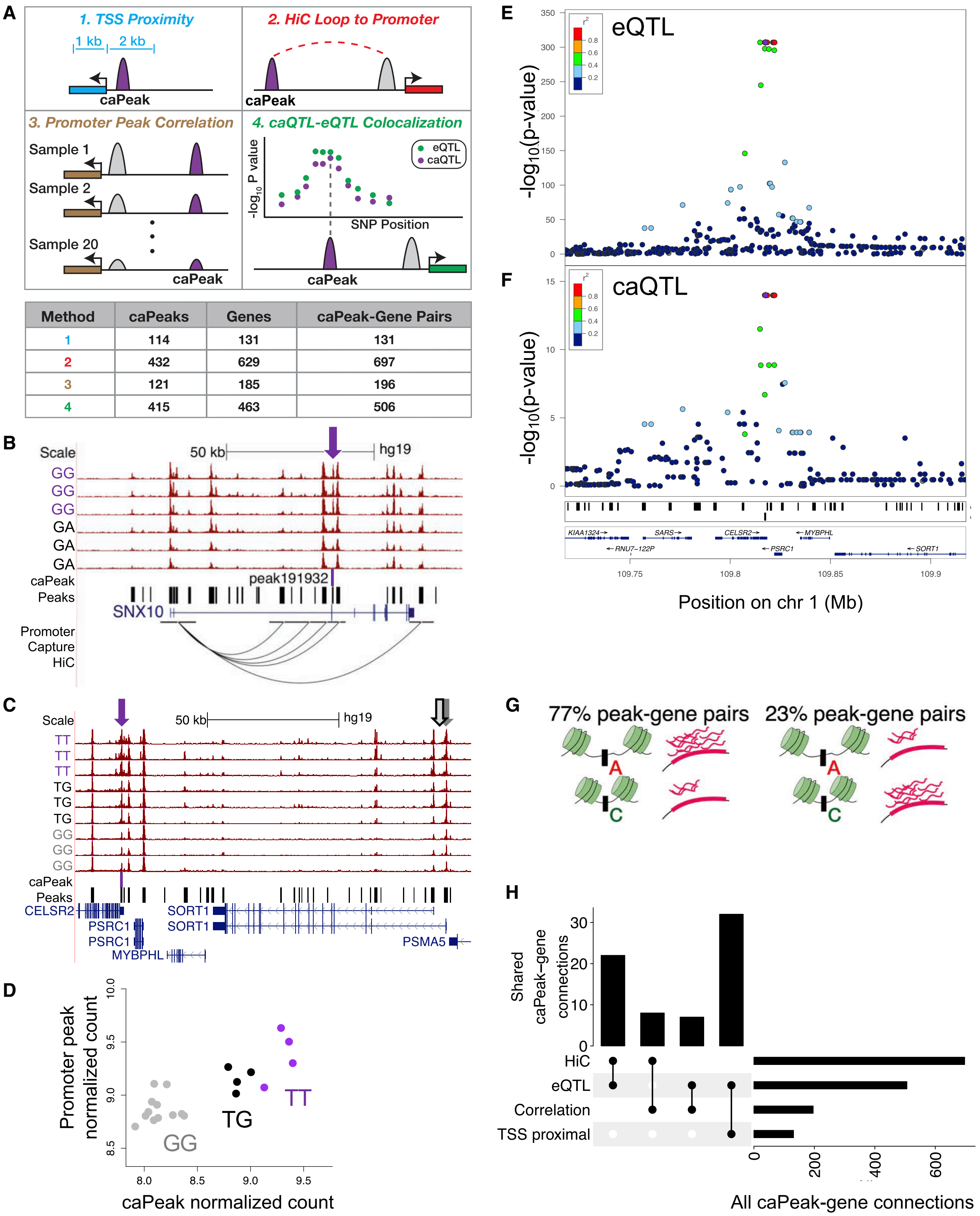


Figure 3. Disruption of TF binding motifs by caQTL variants
(A) Allele affinities for TF binding and chromatin accessibility for variants within caPeaks and in strong LD with the caQTL lead variant ($r^2 > 0.8$).
(B) Association of caQTL status with motif disruption status. Only the 109 TFs with at least 20 motifs disrupted by caQTL variants were included in the analysis, and only the 29 significant associations (p < 4.6 × 10^{-4}) are shown. Error bars indicate 95% confidence intervals.
(C) Percent of disrupted motifs for which the allele with higher chromatin accessibility matched the motif better. Percents are shown for the 29 TFs that had at least 20 motifs disrupted by caQTL variants. Black line, percent for all disrupted motifs across all tested TFs; red line, average percent across the 29 TFs.



Method	caPeaks	Genes	caPeak-Gene Pairs
1	114	131	131
2	432	629	697
3	121	185	196
4	415	463	506

Figure 4. Prediction of target genes for caPeaks using four approaches
(A) Illustrations of four approaches to predict caPeak target genes.
(B) Hi-C chromatin contact shown as an arc between caPeak191932 and the *SNX10* promoter. Selected ATAC-seq signal tracks are shown for each caQTL genotype of rs12534816. More accessible homozygotes, purple; heterozygotes, black.
(C) Genome browser image showing the correlation across rs12740374 genotypes of caPeak9372 and a peak at a *SORT1* promoter. The purple arrow indicates the caPeak and the gray arrow indicates the promoter peak.
(D) The same peak correlation with points representing normalized peak counts of individual samples colored by rs12740374 genotype.

(legend continued on next page)

alter chromatin structure to activate or repress transcription,[76] suggesting that this approach identifies TFs that may influence chromatin accessibility.

To investigate how often TFs bind the more accessible allele, we compared alleles associated with higher chromatin accessibility to the motifs. Among 7,629 motifs for all TFs, the more accessible allele matched the motif better for 4,770 motifs (63%, binomial $p < 4.1 \times 10^{-107}$). Similarly, among 3,132 motifs for the highest expressed TF at each variant, the more accessible allele matched the motif better for 1,953 motifs (62%, binomial $p < 8.0 \times 10^{-44}$). When restricting analysis to 993 observations of the 29 TFs for which motif disruption is associated with caQTL status, the more accessible allele matched the motif better for 834 motifs (84%, binomial $p < 5.1 \times 10^{-111}$). TFs exhibited variation in the percent of motifs that matched better to the more accessible allele (Figure 3C). For 11 TFs, including HNF4A (MIM: 600281), ATF4 (MIM: 604064), ERF (MIM: 611888), and FOXA2 (MIM: 600288), more than 90% of stronger motif matches corresponded to the more accessible allele, while for SPI1 (MIM: 165170) only 56% of stronger motif matches corresponded to the more accessible allele. These results suggest that TFs typically, but not always, bind to the more accessible allele.

Identifying putative target genes for caPeaks

Connecting caPeaks to their target genes is challenging, particularly when the caPeaks are distal to transcription start sites (TSSs). Individual approaches for identifying target genes have limitations and may not always show a direct regulatory relationship between a caPeak and gene. To address these challenges, we used four approaches to connect caPeaks to genes (Figure 4A).

First, we identified caPeaks proximal (−2 kb/+1 kb) to TSSs of genes expressed in liver. Of 3,123 total caPeaks, 114 (4%) were proximal to the TSS of at least 1 gene. Among these 114 caPeaks, 15 were proximal to the TSS of two or three genes (Table S18). This approach identified 131 unique caPeak-gene connections (Figure 4A).

Second, we used liver tissue promoter capture Hi-C[2] to identify caPeaks that physically interact with gene promoters. We identified 329 distal caPeaks (>15 kb from any promoter as defined in the Hi-C analysis) that interact with promoters for 451 genes (Table S18), including a caPeak that interacts with the promoter of *SNX10* (MIM: 614780; Figures 4B and S6A). The caPeak near *SNX10* was identified even though only two genotypes were observed in these samples, demonstrating that caQTL effect sizes can be large. Among caPeaks that overlapped the promoter of one gene and interact with the promoter of another gene, we identified an additional 104 caPeaks that interact with promoters of 190 genes. Combining promoter-distal and promoter-promoter interactions, we identified 697 caPeak-gene connections (Figure 4A; Table S18).

Third, we identified caPeak sizes that either correlated with expression level of nearby genes or with the size of ATAC peaks at promoters. More caPeaks were correlated with promoter ATAC peaks than with gene expression level; 120 caPeaks were significantly correlated (FDR < 5%) with promoter ATAC peaks while only 2 caPeaks were correlated with gene expression (FDR < 5%), resulting in 121 unique caPeaks because gene RP11-101E14.2 had both types of correlations (Table S18; Figure 4A). When using the same p value threshold for both analyses ($p < 2.9 \times 10^{-4}$), 5 additional caPeaks were correlated with gene expression. As an example at a regulatory element previously shown to regulate *SORT1*[77] (MIM: 602458), caPeak9372 is positively correlated with a peak proximal to a *SORT1* TSS (peak9400, Spearman rho = 0.76, $p < 1.6 \times 10^{-4}$; Figures 4C, 4D, and S6B; Table S18) and nominally correlated with *SORT1* expression (Spearman rho = 0.69, $p < 1.2 \times 10^{-3}$; Table S18). The vast majority of peak-peak correlations (167 of 173, 97%) are positive, suggesting that higher caPeak accessibility is usually associated with higher accessibility of connected promoter peaks (Table S18). Using either caPeak-promoter peak or caPeak-gene correlations, we identified 196 caPeak-gene connections (Figure 4A; Table S18).

Finally, we identified caQTLs for which the lead variant exhibited high LD ($r^2 > 0.8$) with an eQTL lead variant for 15,418 autosomal genes from a liver tissue eQTL meta-analysis of 1,183 individuals.[11] Of 3,119 unique caQTL lead variants, 414 (13%) were in strong LD with at least 1 eQTL lead variant (Table S18), which is similar to the percentage reported in a previous caQTL study.[6] Among caQTL lead variants, 71 were in strong LD with more than one eQTL lead variant, suggesting that some caPeaks may affect expression of multiple genes. In total, we identified 463 target genes for 415 caPeaks, representing 506 unique caPeak-gene connections (Figure 4A; Table S18). For example, we identified a caQTL signal with the same variants as an eQTL signal for *SORT1* (Figures 4E and 4F). At connected loci, the allele associated with higher chromatin accessibility was usually associated with higher gene expression (390 of 506 loci, 77%; Figure 4G), suggesting caPeaks frequently act as promoters or enhancers to gene expression. We obtained a similar result when restricting to caQTL variants associated with

(E and F) *SORT1* eQTL associations at the signal colocalized with the caQTL for caPeak9372 (E) and caQTL associations with caPeak9372 (F). In both plots, the caQTL lead variant within 1 kb of the peak center is indicated by a purple diamond and LD is based on 1000G phase 3 Europeans.

(G) Comparison of directions of effect among all colocalized caQTL and eQTL signals. The A allele represents the more accessible allele than C, and more red marks indicate higher gene expression.

(H) UpSet plot comparing the number of shared and unique caPeak-gene links identified by the four approaches. It is not possible for a caPeak-gene pair to be predicted using all four methods because if a caPeak is TSS proximal, it cannot form a Hi-C loop with the same gene and it cannot be a distal caPeak correlated with the promoter peak for the same gene.

only one peak and colocalized with eQTL variants associated with only one gene (273 of 337 loci, 81%). Of the 506 caQTL-eQTL signals colocalized based on LD, 28 showed strong evidence of colocalization using coloc[66] (PP4 > 0.8), and an additional 48 showed suggestive evidence of colocalization (PP4 > 0.5 but < 0.8) (Table S18). Of the 430 signals that did not show suggestive evidence of colocalization, 409 (95%) did not have sufficient power to detect colocalization (PP0+PP1+PP2 > 0.5) and no signals showed evidence of separate, but not colocalized signals (PP3 > 0.5). Therefore, we conclude that the study is underpowered to detect colocalizations using coloc.

Together the four methods identified a total of 1,461 caPeak-gene connections, although the approaches showed low overlap. Only 69 caPeak-gene connections were predicted by two methods, and no connections by three methods, likely due to the low power of many of the approaches (Figure 4H; Table S18). The 69 caPeak-gene associations consist of 67 unique caPeaks and 67 unique genes; two caPeaks had two target genes. It is not possible for a caPeak-gene pair to be predicted using all four methods because if a caPeak is TSS proximal, it cannot be found within the distal end of a Hi-C loop >15 kb from the same TSS and it cannot be a distal caPeak correlated with the promoter peak for the same gene. Thus, the only method that can corroborate TSS proximity is caQTL-eQTL colocalization. Of the 131 caPeak-gene connections identified by TSS proximity, 32 (24%) were supported by caQTL-eQTL colocalization. In addition, when considering peak correlations for which the distal caPeak was tested by other approaches, 98 of 307 (32%) caQTL-eQTL colocalizations and 108 of 436 (25%) Hi-C loops showed at least nominal ($p < 0.05$) evidence. These methods are limited by power and technical factors, suggesting that the 69 caPeak-gene connections identified by two methods may be a conservative estimate. This integrated approach predicted a target gene for 861 of 3,123 caPeaks (28%), suggesting that caPeaks frequently interact with genes.

Prediction of regulatory mechanisms at GWAS loci

To identify genetic variants that may influence disease by altering chromatin accessibility, we identified colocalized caQTL and GWAS signals, based on strong LD ($r^2 > 0.8$) between lead caQTLs and lead GWAS variants. Using GWAS variants for 19 traits relevant to liver function and cardiometabolic traits from the NHGRI-EBI GWAS catalog[67] (Table S19), we identified 110 potentially colocalized caQTL and GWAS signals, corresponding to 111 caPeaks, because one caQTL signal was associated with two caPeaks (Table S20). We identified at least one colocalized caQTL for 15 of the 19 traits, and of the GWAS signals for these traits, liver enzymes showed the highest percentage of potentially colocalized caQTLs (14 signals, 18%) (Table 1). For traits with at least 5 GWAS-caQTL signals, we identified a relatively high percentage of colocalized signals (>5%) for total cholesterol and LDL cholesterol, consistent with the involvement of liver in lipid metabolism.[10] As a negative control, we observed a relatively low percentage (<2%) of GWAS signals colocalized with liver caQTLs for height and rheumatoid arthritis (Table S21).

Only 26 of the 143 (18%) liver caQTL-GWAS colocalizations were observed using blood caQTL datasets (Table S20). For liver enzymes, total cholesterol, and LDL cholesterol, respectively, only 3 of 14, 3 of 18, and 2 of 13 liver caQTL-GWAS colocalizations were observed in blood (Table S21). GWAS signals for liver enzymes were colocalized with a higher percentage of liver caQTLs (0.51%) than each of the blood cell type caQTLs (0.06%–0.12%), whereas GWAS signals for rheumatoid arthritis were colocalized with a higher percentage of blood caQTLs (0.09%–0.21%) than liver caQTLs (0.06%) (Table S22). However, many of these colocalization differences between liver and blood may be due to limited caQTL sample sizes. Larger studies using identical caQTL pipelines are needed to robustly identify cell type-specific caQTL-GWAS colocalizations.

To identify plausible regulatory mechanisms at GWAS loci, we integrated our GWAS-colocalized caQTLs with TF motif-disrupting variants and predicted caPeak target genes. Of the 111 caPeaks at potentially colocalized caQTL-GWAS signals for liver function or cardiometabolic traits, 85 harbored a TF motif-disrupting variant, 56 had a predicted target gene, and 45 of these overlapped with both types of data. The gene with a TSS closest to the GWAS lead variant was predicted to be a target gene for 25 of 56 caPeaks (45%).

We identified seven liver function or cardiometabolic GWAS-caQTL colocalized signals with strong evidence of regulatory mechanisms. At these GWAS loci, the caPeak had a target gene identified by two approaches and harbored TF motif-disrupting variants (Table 2). We identified colocalized caQTL, eQTL, and GWAS signals and a correlated caPeak-promoter peak pair (Tables 2 and S20; Figures 4C–4F) at the *SORT1* locus associated with LDL cholesterol for which the alternate allele (rs12740374-T) has been shown to create a CEBP binding site and increase hepatic *SORT1* expression.[77] At a less well-characterized locus, the caQTL signal with lead variant rs13395911 associated with caPeak119621 is colocalized with GWAS signals for plasma liver enzyme levels in European[48] and Asian[78] individuals and an eQTL for *EFHD1*[11] (MIM: 611617; Figures 5A–5C and S7). Increased accessibility corresponds to higher *EFHD1* expression level and higher liver enzyme levels. caPeak119621 physically interacts with the promoter of *EFHD1* in liver tissue promoter capture Hi-C data[2] (Figure 5D), further suggesting that caPeak119621 may affect *EFHD1* expression. CaPeak119621 does not overlap an ATAC peak in macrophages[4] (Table S12). The peak overlaps ChIP-seq peaks for 12 TFs in liver (Figure 5E), and rs13395911 disrupts motifs for eight TFs expressed in liver (Tables S16 and S20). The motif with the largest difference between rs13395911 alleles is for *FOXA2*, and the allele with higher chromatin accessibility matches the motif better (Figure 5F). These and other

Table 1. Colocalized GWAS-caQTL signals by trait

Trait	Number of GWAS signals[a]	Number of colocalized caQTL-GWAS signals[b]	Percent of colocalized caQTL-GWAS signals[c]
Liver enzymes	77	14	18.2
Total cholesterol	292	18	6.2
Glucose	54	3	5.6
Insulin	18	1	5.6
LDL cholesterol	240	13	5.4
Bilirubin	20	1	5.0
HDL cholesterol	314	12	3.8
C-reactive protein	81	3	3.7
Triglycerides	279	10	3.6
Cardiovascular disease	454	14	3.1
Body mass index	986	29	2.9
Blood pressure	1,540	38	2.5
Type 2 diabetes	268	5	1.9
HbA1c	66	1	1.5
WHRadjBMI	209	3	1.4
Glycated albumin	2	0	0.0
Liver injury	17	0	0.0
NAFLD	9	0	0.0
Serum albumin	15	0	0.0

LDL, low-density lipoprotein; HDL, high-density lipoprotein; WHRadjBMI, waist-hip ratio adjusted for BMI; NAFLD, non-alcoholic fatty liver disease.
[a] Counted as lead GWAS variants not in high LD ($r^2 < 0.8$) with another.
[b] Colocalized if the caQTL lead variant was in strong LD ($r^2 > 0.8$) with the GWAS lead.
[c] Percent of all GWAS signals that are colocalized with a caQTL.

connections provide potential regulatory mechanisms linking variants to regulatory element, transcription factors and genes that may influence the GWAS traits.

Identification of a putative functional variant at the *LITAF* locus

Near *LITAF* (MIM: 603795), which encodes lipopolysaccharide (LPS)-induced TNF factor, we identified a caQTL signal for caPeak75869 and tested variants for allelic differences in transcriptional activity and protein binding. This caQTL signal is potentially colocalized with a GWAS signal for LDL cholesterol[79] and an eQTL signal for *LITAF*[11] (Figures 6A, 6B, and S8). caPeak75869 loops to the promoter of *LITAF* in liver tissue promoter capture Hi-C^2 (Figure 6C). caPeak75869 contains the lead caQTL variant rs57792815 (caQTL $p < 5.0 \times 10^{-17}$) and two additional variants in strong LD with the caQTL lead, rs3784924 ($r^2 = 0.95$) and rs11644920 ($r^2 = 0.98$). The haplotype associated with higher accessibility consists of the rs57792815-T, rs3784924-A, and rs11644920-A alleles. We tested a 666-bp DNA construct spanning the three variants for haplotype differences in transcriptional activity using luciferase reporter assays, testing the construct in two orientations relative to a minimal promoter. Given that *LITAF* is involved in lipopolysaccharide (LPS)-stimulated immune response,[80] we tested transcriptional activity in four cell types: HepG2 hepatocytes, THP-1 monocytes, THP-1 differentiated macrophages, and LPS-stimulated THP-1 macrophages. In all four cell types, the forward orientation construct containing the alleles associated with higher accessibility showed significantly higher transcriptional activity than the construct containing the other alleles, with the strongest differences observed in hepatocytes (fold change = 2.49, $p = 2 \times 10^{-4}$) and LPS-stimulated macrophages (fold change = 1.39, $p = 7 \times 10^{-4}$; Figure 6D). The same haplotype showed significantly higher transcriptional activity in the reverse orientation for hepatocytes ($p = 1 \times 10^{-4}$) and unstimulated macrophages ($p = 0.02$) and a trend toward higher transcriptional activity in the other cell types (Figure S8G). Although allelic differences were observed in all four cell types, caPeak75869 does not overlap an ATAC peak in macrophages[4] (Table S12). We next tested each of the three haplotype variants for allelic differences in protein binding using nuclear extract from HepG2 cells. Only rs11644920 showed allele-specific binding, with the T allele showing increased binding (Figures 6E and S8H). caPeak75869 contained liver ChIP-seq binding sites for numerous TFs and all three variants within the peak disrupted motifs (Figure 6F; Tables S16 and S20). We focused on the motif disrupted by rs11644920 because it

Table 2. Selected caQTLs at GWAS loci

caQTL variant	caPeak	GWAS variant	GWAS trait	LD r^2 [a]	Gene	Methods[b]	caQTL, eQTL directions[c]
rs12740374	peak9372	rs12740374	LDL cholesterol	1.00	*SORT1*	eQTL, Corr	D, D
rs17276527	peak13768	rs4077194	HDL cholesterol	1.00	*RALGPS2*	eQTL, HiC	D, D
rs13395911	peak119621	rs13395911	ALT	1.00	*EFHD1*	eQTL, HiC	I, I
rs2232015	peak9185	rs1730859	LDL cholesterol	0.97	*PRMT6*	TSS, eQTL	D, D
rs2037517	peak71475	rs832890	Pulse pressure	0.90	*PLEKHO2*	eQTL, HiC	D, D
rs12677006	peak205272	rs1906672	Sys. blood pressure	0.89	*DDHD2*	eQTL, HiC	I, I
rs57792815	peak75869	rs34318965	LDL cholesterol	0.81	*LITAF*	eQTL, HiC	I, I

Loci are shown for shared caQTL-GWAS signals if the caPeak was linked to a target gene by two methods and if the caPeak harbored motif-disrupting variants. ALT, alanine aminotransferase levels; Sys, systolic.

[a] LD r^2 between the caQTL and GWAS lead variants.

[b] Methods that linked the caPeak to a gene. Corr, correlation between caPeak and promoter peak accessibility.

[c] Direction of chromatin accessibility and gene expression relative to the allele associated with an increase in the GWAS trait, where "I" indicates increased and "D" indicates decreased accessibility or expression. Additional traits and loci are listed in Table S20.

was the only variant that showed allelic differences in binding in the EMSA results. Variant rs11644920 disrupted a motif for *ATF2*, and the A allele matched the motif better (Figure S8I), which is also the allele associated with higher chromatin accessibility. This result contrasts the EMSA results, which showed greater binding for the T allele. Together, these results suggest that altered transcription factor binding at rs11644920 and increased chromatin accessibility of the regulatory element marked by caPeak75869 may lead to increased transcriptional activity and higher *LITAF* expression.

Discussion

We profiled chromatin accessibility in 20 individuals and identified caQTLs in human liver tissue. caQTL variants frequently disrupt TF binding motifs, and alleles that better match a motif often have higher chromatin accessibility, consistent with TFs stabilizing chromatin in an accessible state. We identified 1,461 putative caPeak-gene links using four approaches, suggesting that caPeaks frequently regulate gene expression. We identified 110 caQTLs at GWAS signals, including 56 with a predicted caPeak target gene, identifying regulatory mechanisms that may be responsible for trait variation. Among variants at a colocalized caQTL, eQTL, and LDL cholesterol GWAS signal near *LITAF*, one variant showed allelic differences in transcriptional activity and *in vitro* TF binding. This study contributes to the epigenomic characterization of human liver tissue and will aid in functional characterization of GWAS loci that act in liver.

Combining caQTLs, caPeak-gene links, and disrupted TF motifs helps identify mechanisms at GWAS loci. At the well-characterized *SORT1* GWAS locus for lipid and cardiovascular traits,[77] we showed that the previously described functional variant rs12740374 is associated with chromatin accessibility and that the caPeak containing this variant is correlated with a peak at the *SORT1* promoter. We also identified plausible regulatory mechanisms at less well-characterized loci. At a GWAS signal for BMI[81] and LDL cholesterol,[79] we identified a caQTL potentially colocalized with a *PRMT6* (MIM: 608274) eQTL signal and observed that the caPeak overlapped the *PRMT6* TSS. *PRMT6* has been shown to regulate hepatic glucose metabolism in mice.[82] Our data suggest that a variant at this locus may increase chromatin accessibility and alter TF binding at the *PRMT6* TSS, leading to higher *PRMT6* expression and decreased LDL cholesterol (Table S20). At a GWAS locus for plasma liver enzyme levels,[48,78] we predicted *EFHD1* as a target gene based on both caQTL-eQTL colocalization and a promoter capture Hi-C link. While *EFHD1* is expressed in liver tissue, the GTEx portal shows that expression is much higher in other tissues,[1] and the gene's roles in liver have not been characterized.[83] The caPeak at this locus does not overlap an ATAC peak in macrophages[4] (Table S12), but additional experiments, such as single nucleus ATAC-seq, are needed to determine the relevant cell type within liver tissue. Our data suggest that *EFHD1* may be a target gene at this locus and act through one or more of the cell types in liver tissue. These and other results highlight the utility of caQTLs to identify mechanisms at GWAS loci.

At the *LITAF* locus, we provided direct evidence that variant rs11644920 can alter transcriptional regulation. Here, the caQTL, liver eQTL, and LDL cholesterol GWAS signals are colocalized, and the variant, mechanism, and cell type responsible for these associations were unknown. *LITAF* encodes a transcription factor that can mediate effects on inflammation,[80] suggesting a potential role in hepatocytes and/or macrophages in an inflammatory environment. We showed that variants in the caPeak alter transcriptional reporter activity in hepatocytes, monocytes, macrophages, and lipopolysaccharide-stimulated macrophages. In all cell types, the caPeak showed a similar magnitude of enhancer activity and alleles showed differences in transcriptional activity, suggesting that the variant may act in any or all of these cell types. The caPeak at this locus does not overlap an ATAC peak in macrophages[4] (Table S12), but additional experiments, such as

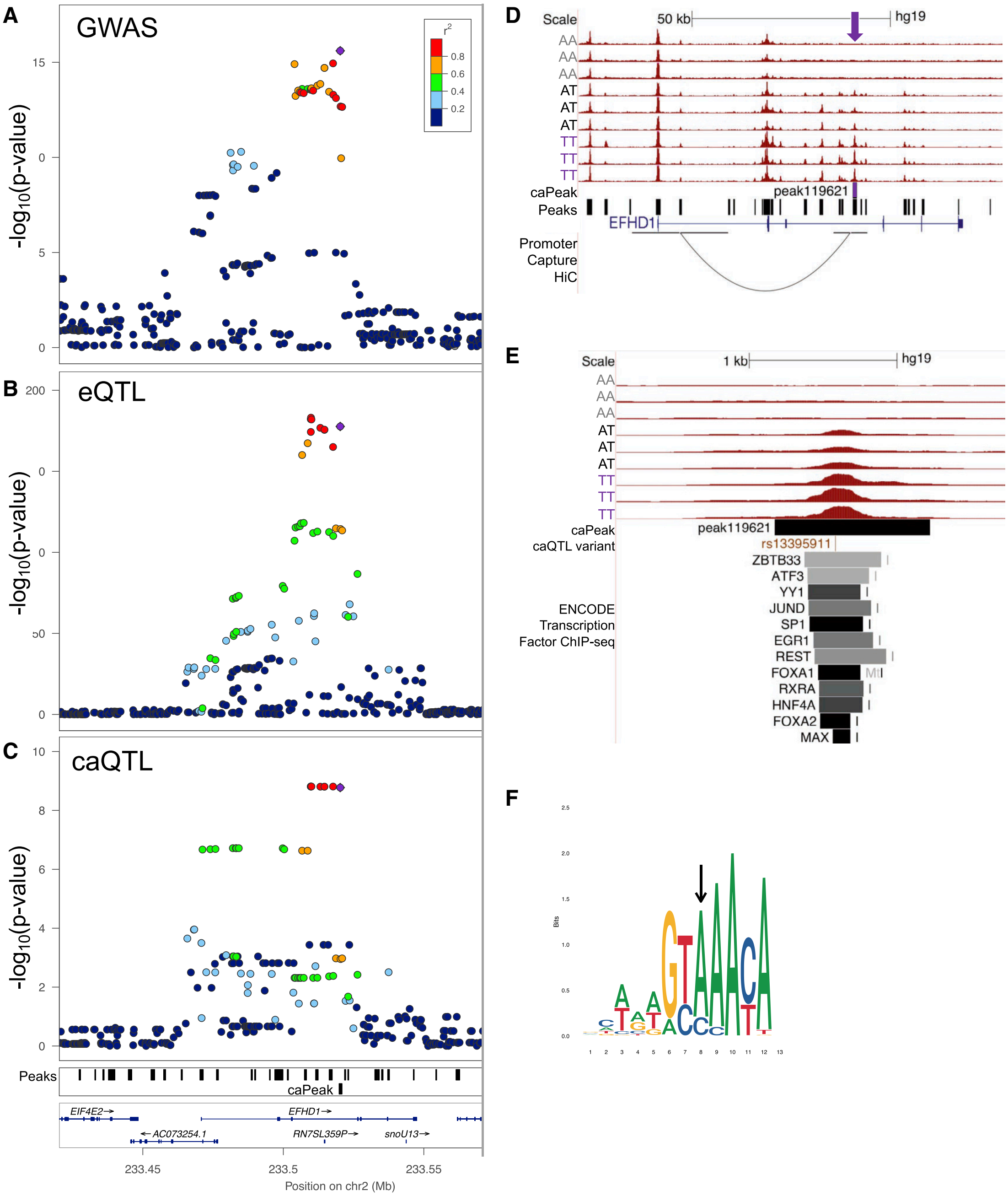


Figure 5. A plausible regulatory mechanism at the *EFHD1* locus for plasma liver enzyme levels

(A–C) GWAS association with plasma levels of the liver enzyme alanine transaminase in Japanese individuals (A), eQTL association for *EFHD1* (B), and caQTL associations for caPeak119621 (C). For all three plots, the caQTL lead variant within 1 kb of the peak center is indicated by a purple diamond and LD is based on 1000G phase 3 East Asians (A) or Europeans (B and C). Additional plots are shown in Figure S7.

(D) Hi-C chromatin contact shown as an arc between caPeak119621 and the *EFHD1* promoter. Selected ATAC-seq signal tracks are shown for each rs13395911 genotype. More accessible homozygotes, purple; heterozygotes, black; less accessible homozygote, gray.

(E) Transcription factor ChIP-seq peaks in liver tissue from ENCODE that overlap caPeak119621.

(F) Sequence logo plot for the *FOXA2* motif s disrupted by caQTL variant rs13395911 (arrow). The motif match is shown on the negative strand, and variant alleles in (D) and (E) are shown on the positive strand.

single nucleus ATAC-seq, are needed to determine the relevant cell type within liver tissue. We further provided evidence that rs11644920 alters protein binding, at least *in vitro*. Further study is needed to provide direct evidence that these variants alter transcription of *LITAF* and how altered levels of *LITAF* may affect cholesterol levels.

The maximum distance threshold between peaks and tested variants had a substantial impact on caQTL detection. Analyzing variants within a narrow region around a peak reduced the multiple testing burden for nearby variants, whereas testing variants in a broader region allowed identification of variants within one peak that may also

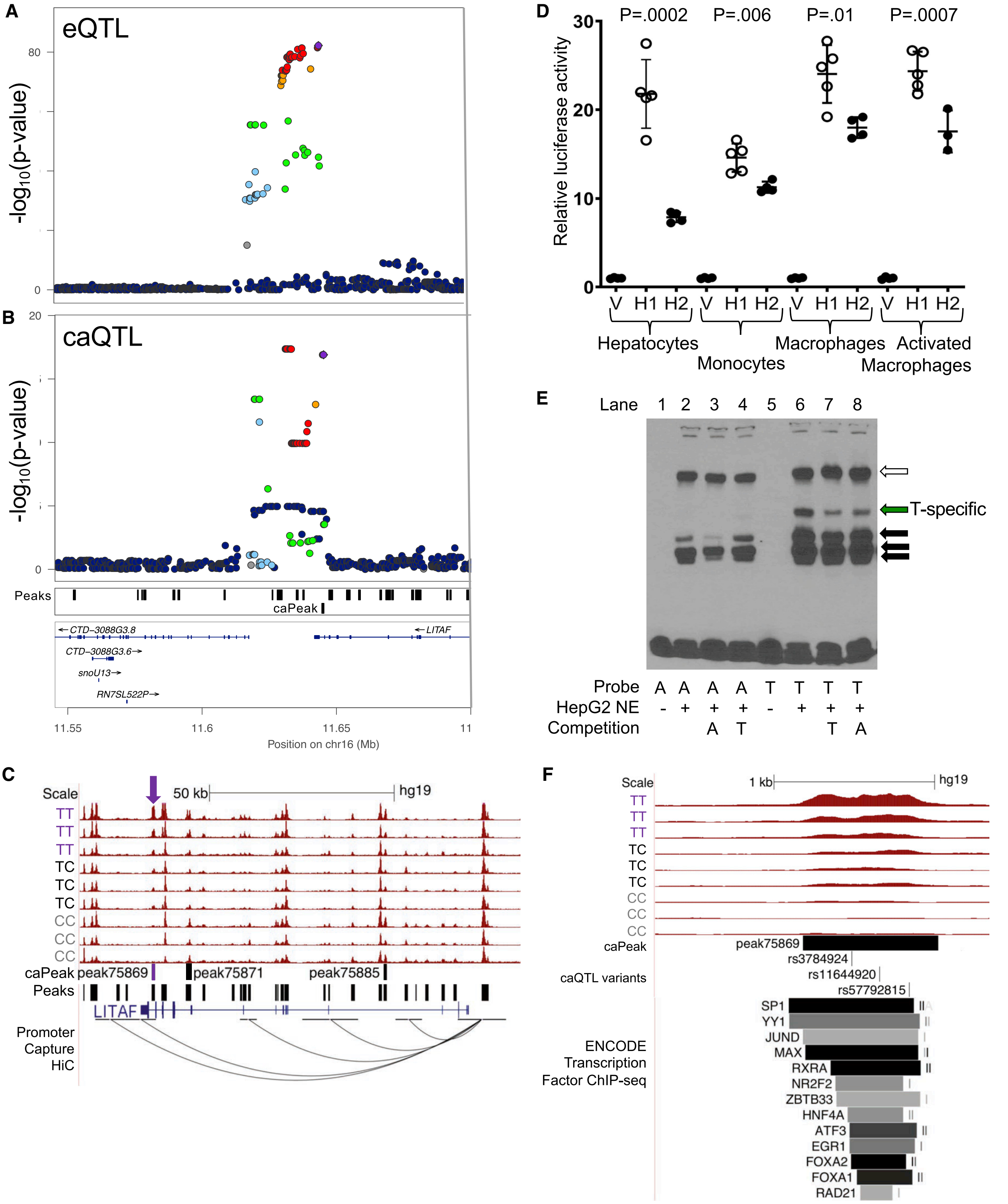


Figure 6. Identification of a putative functional variant at the *LITAF* locus for LDL cholesterol

(A and B) eQTL association for *LITAF* (A) and caQTL associations for caPeak75869 (B) at an LDL cholesterol GWAS signal. In both plots, the caQTL lead variant within 1 kb of the peak center is indicated by a purple diamond, and LD is based on 1000G phase 3 Europeans. Additional plots are shown in Figure S8.

(C) Hi-C chromatin contact between caPeak75869 and the *LITAF* promoter. Selected ATAC signal tracks are shown for each rs57792815 genotype. More accessible homozygotes, purple; heterozygotes, black; less accessible homozygotes, gray.

(D) Transcriptional activity of a 666-bp DNA element spanning caPeak75869 and containing rs3784924, rs11644920, and rs57792815 in HepG2 hepatocytes, THP-1 monocytes, THP-1 differentiated macrophages, and LPS-stimulated THP-1 macrophages. The DNA element

(legend continued on next page)

influence another peak. A wide range of distance thresholds have been applied to caQTL discovery, including variants within 1 kb and 20 kb of peak centers,[6] 50 kb from peak ends,[4] and 1 Mb from peak ends.[8] We found many more significant results when using variants within 1 kb of peak centers compared to variants within 100 kb of peak centers, potentially due to reduced multiple testing burden and low power to detect long-range caQTL effects due to small sample size. Future caQTL studies with larger sample sizes will be more powered to detect longer-range caQTLs.

Due to the modest sample size of this study, we only tested for caQTLs using common variants (MAF $\geq$ 0.1) and did not predict regulatory variants at low-frequency GWAS signals. Based on three large GWASs for height,[52] body mass index,[52] and blood lipids[79] (see web resources), 77%–91% of signals had lead variant MAF $\geq$ 0.1, suggesting that we could test the majority of GWAS signals for caQTLs. However, allele frequencies in small sample sizes may differ from population allele frequencies, and larger caQTL studies will have more power to detect caQTLs at low frequency variants.

We used four approaches to suggest genes that may be regulated by caPeaks. However, several factors limit how many caPeak-gene connections can be identified and how many are shared by two or more approaches. TSS proximity is useful to detect variation in promoter accessibility, although our results showed that only 4% of caPeaks are TSS proximal, and caQTL-eQTL colocalization is the only method we tested that can corroborate TSS proximity. Promoter capture Hi-C data[2] identifies distal regions that physically interact with promoters, although additional Hi-C loops may be identified in additional samples and with higher sequencing depth. Hi-C loops < 15 kb were removed,[2] indicating that the Hi-C data cannot corroborate caQTL-eQTL colocalizations or caPeak-promoter peak/gene expression correlations located < 15 kb from the promoter. The identification of caPeaks correlated with promoter peaks[84] or with gene expression is limited by sample size, and gene expression is affected by many other factors. The LD-based method we used to predict colocalized caQTL and eQTL signals helps identify peaks and genes with a shared genetic basis, although this method is influenced by low-resolution fine-mapping of the lead caQTL variant, use of an LD threshold, and choice of LD reference panel. Due to the modest sample size of this study, we were underpowered to detect colocalizations using coloc,[66] and we recommend that future caQTL studies consider larger sample sizes for more robust colocalizations. Identification of conditional liver eQTLs, which tend to be further from gene TSSs compared to primary eQTLs,[85,86] could lead to additional caQTL-eQTL colocalizations. While each of these approaches was useful to predict links between caPeaks and genes, additional experiments are needed to identify causal relationships.

The caQTLs presented here are a resource for studying liver regulatory elements and will help identify mechanisms at GWAS loci for multiple traits that act through liver. The 56 caQTLs at GWAS loci with predicted target genes are strong candidates for future functional studies. While caQTLs can pinpoint functional regulatory variants, the modest sample size and analyses restricted to common variants limit fine-mapping potential and highlight the importance of considering LD proxies. The promising regulatory mechanisms identified here motivate identification of liver caQTLs in larger sample sizes.

Data and code availability

Summary statistics for caQTL data are available at https://mohlke.web.unc.edu/data/.

Raw and processed genotype, ATAC-seq, and RNA-seq data are available in the Gene Expression Omnibus (GEO). The accession number for the data reported in this paper is GEO: GSE164942.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.05.001.

Acknowledgments

This study was supported by NIH grants R01DK072193, U01DK105561, R01DK093757, UM1DK126185, and 1-ZIA-HG000024. Individuals were supported by NIH grants F31HL146121 (K.W.C.), T32GM067553 (K.W.C.), T32HL069768 (H.J.P.), T32HL129982 (K.A.B.), and R25GM055336 (K.W.C). Liver samples were obtained from the Liver Tissue Cell Distribution System supported by NIH contract N01DK70004/HHSN267200700004C.

We thank Dan Liang, Jason Stein, Bryan Quach, and Terry Furey for helpful discussions on computational analyses and Gautam Pandey for literature reviews and biological interpretation of loci.

We also thank the Roadmap Epigenomics Project for liver tissue chromatin states, the ENCODE Consortium for liver transcription factor ChIP-seq data, Cis-BP for TF binding motifs, and GTEX for liver tissue gene expression data. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal.

was tested in the forward orientation relative to the genome (reverse orientation in Figure S8G). V, empty vector; H1, haplotype 1 of more accessible alleles rs3784924-A, rs11644920-A, and rs57792815-T; H2, haplotype 2 of less accessible alleles rs3784924-G, rs11644920-T, and rs57792815-C. Symbols represent 4–5 independent clones for each haplotype tested in duplicate wells; bars indicate mean ± standard deviation; p values from t tests of allelic differences.

(E) EMSA using HepG2 nuclear extract (NE) shows allelic differences in protein binding for rs11644920. rs3784924 and rs57792815 are shown in Figure S8H. Green arrow, band represents T-allele-specific binding; black arrows, T-allele-preferential binding; white arrow, non-specific binding. Competition probes were unlabeled and included in 10-fold excess.

(F) TF ChIP-seq peaks in liver tissue from ENCODE that overlap caPeak75869.

Declaration of Interests

J.P.D. is a current employee of DNAnexus.

Received: October 26, 2020
Accepted: May 4, 2021
Published: May 25, 2021

Web resources

Blood lipids GWAS summary statistics for stratified LD score regression, http://lipidgenetics.org/

Blood lipids significant GWAS lead variants, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6521726/bin/NIHMS1502930-supplement-3.xlsx

Body mass index GWAS summary statistics for stratified LD score regression, https://portals.broadinstitute.org/collaboration/giant/images/1/14/Bmi.giant-ukbb.meta-analysis.combined.23May2018.HapMap2_only.txt.gz

Body mass index significant GWAS lead variants, https://portals.broadinstitute.org/collaboration/giant/images/e/e2/Meta-analysis_Locke_et_al%2BUKBiobank_2018_top_941_from_COJO_analysis_UPDATED.txt.gz

BWA, https://github.com/lh3/bwa

Coronary artery disease GWAS summary statistics for stratified LD score regression, http://www.cardiogramplusc4d.org/media/cardiogramplusc4d-consortium/data-downloads/UKBB.GWAS1KG.EXOME.CAD.SOFT.META.PublicRelease.300517.txt.gz

CTA, https://github.com/ParkerLab/cta

dbSNP build 151 common variants, ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/00-common_all.vcf.gz

Gplots R package, https://rdrr.io/cran/gplots/

Height GWAS summary statistics for stratified LD score regression, https://portals.broadinstitute.org/collaboration/giant/images/6/63/Meta-analysis_Wood_et_al%2BUKBiobank_2018.txt.gz

Height significant GWAS lead variants, https://portals.broadinstitute.org/collaboration/giant/images/4/4b/Meta-analysis_Wood_et_al%2BUKBiobank_2018_top_3290_from_COJO_analysis.txt.gz

Liver enzymes GWAS summary statistics for stratified LD score regression, http://www.lolipopstudy.org/data-download

Macrophage ATAC peaks from Alasoo et al., https://zenodo.org/record/1188300/files/ATAC_peak_metadata.txt.gz

Macrophage caQTLs in 4 experimental conditions from Alasoo et al., https://zenodo.org/record/1133333#.X-T2NNhKg2w

Novoalign, http://www.novocraft.com/products/novoalign

OMIM, https://www.omim.org/

Picard, https://github.com/broadinstitute/picard

Promoter capture Hi-C data (liver code is LI11), http://kobic.kr/3div/download

Promoter capture Hi-C promoter baits, https://junglab.wixsite.com/home/db-link

Regions of unusually high linkage disequilibrium, https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)

Rheumatoid arthritis GWAS summary statistics for stratified LD score regression, http://plaza.umin.ac.jp/~yokada/datasource/files/GWASMetaResults/RA_GWASmeta_European_v2.txt.gz

swiss, https://github.com/statgen/swiss

T cell local caQTLs from Gate et al. (sheet 1): https://www.nature.com/articles/s41588-018-0156-2

Type 2 diabetes (T2D GWAS meta-analysis - Unadjusted for BMI) GWAS summary statistics for stratified LD score regression, https://diagram-consortium.org/downloads.html

WHRadjBMI GWAS summary statistics for stratified LD score regression, https://portals.broadinstitute.org/collaboration/giant/images/6/6e/Whradjbmi.giant-ukbb.meta-analysis.combined.23May2018.HapMap2_only.txt.gz

References

1. GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.
2. Jung, I., Schmitt, A., Diao, Y., Lee, A.J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet. *51*, 1442–1449.
3. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
4. Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., Hale, C., Dougan, G., Gaffney, D.J.; and HIPSCI Consortium (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet. *50*, 424–431.
5. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. *48*, 206–213.
6. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNaseI sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.
7. Bryois, J., Garrett, M.E., Song, L., Safi, A., Giusti-Rodriguez, P., Johnson, G.D., Shieh, A.W., Buil, A., Fullard, J.F., Roussos, P., et al. (2018). Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nat. Commun. *9*, 3121.
8. Gate, R.E., Cheng, C.S., Aiden, A.P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M.G., Subramaniam, M., Shamim, M., et al. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat. Genet. *50*, 1140–1150.
9. Khetan, S., Kursawe, R., Youn, A., Lawlor, N., Jillette, A., Marquez, E.J., Ucar, D., and Stitzel, M.L. (2018). Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. Diabetes *67*, 2466–2477.
10. Trefts, E., Gannon, M., and Wasserman, D.H. (2017). The liver. Curr. Biol. *27*, R1147–R1151.
11. Etheridge, A.S., Gallins, P.J., Jima, D., Broadaway, K.A., Ratain, M.J., Schuetz, E., Schadt, E., Schroder, A., Molony, C., Zhou, Y., et al. (2020). A New Liver Expression Quantitative Trait Locus Map From 1,183 Individuals Provides Evidence for Novel Expression Quantitative Trait Loci of Drug Response, Metabolic, and Sex-Biased Phenotypes. Clin. Pharmacol. Ther. *107*, 1383–1393.
12. Çalışkan, M., Manduchi, E., Rao, H.S., Segert, J.A., Beltrame, M.H., Trizzino, M., Park, Y., Baker, S.W., Chesi, A., Johnson, M.E., et al. (2019). Genetic and Epigenetic Fine Mapping of

Complex Trait Associated Loci in the Human Liver. Am. J. Hum. Genet. *105*, 89–107.

13. Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., and Weber, B.H.F. (2018). A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. Sci. Rep. *8*, 5865.
14. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
15. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
16. Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. *20*, 207–220.
17. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.
18. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
19. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
20. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
21. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.
22. Loh, P.-R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816.
23. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
24. Varshney, A., Scott, L.J., Welch, R.P., Erdos, M.R., Chines, P.S., Narisu, N., Albanus, R.D., Orchard, P., Wolford, B.N., Kursawe, R., et al.; NISC Comparative Sequencing Program (2017). Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc. Natl. Acad. Sci. USA *114*, 2301–2306.
25. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
26. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.
27. Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am. J. Hum. Genet. *91*, 839–848.
28. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47* (D1), D766–D773.
29. Hartley, S.W., and Mullikin, J.C. (2015). QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics *16*, 224.
30. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.
31. Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., et al. (2016). The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat. Commun. *7*, 11764.
32. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.
33. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal *17*, 10–12.
34. Rai, V., Quang, D.X., Erdos, M.R., Cusanovich, D.A., Daza, R.M., Narisu, N., Zou, L.S., Didion, J.P., Guan, Y., Shendure, J., et al. (2020). Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. Mol. Metab. *32*, 109–121.
35. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.; and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
36. Orchard, P., Kyono, Y., Hensley, J., Kitzman, J.O., and Parker, S.C.J. (2020). Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. Cell Syst. *10*, 298–306.e4.
37. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.
38. Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinformatics *47*, 1–34.
39. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.
40. R Core Team (2015). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).
41. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, D493–D496.
42. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.
43. Castro-Mondragon, J.A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2017). RSAT matrix-clustering:

dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. *45*, e119.

44. McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 165.
45. Ramaker, R.C., Savic, D., Hardigan, A.A., Newberry, K., Cooper, G.M., Myers, R.M., and Cooper, S.J. (2017). A genome-wide interactome of DNA-associated proteins in the human liver. Genome Res. *27*, 1950–1960.
46. Jou, J., Gabdank, I., Luo, Y., Lin, K., Sud, P., Myers, Z., Hilton, J.A., Kagda, M.S., Lam, B., O'Neill, E., et al. (2019). The ENCODE Portal as an Epigenomics Resource. Curr. Protoc. Bioinformatics *68*, e89.
47. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
48. Chambers, J.C., Zhang, W., Sehmi, J., Li, X., Wass, M.N., Van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S.E., et al.; Alcohol Genome-wide Association (AlcGen) Consortium; Diabetes Genetics Replication and Meta-analyses (DIAGRAM+) Study; Genetic Investigation of Anthropometric Traits (GIANT) Consortium; Global Lipids Genetics Consortium; Genetics of Liver Disease (GOLD) Consortium; International Consortium for Blood Pressure (ICBP-GWAS); and Meta-analyses of Glucose and Insulin-Related Traits Consortium (MAGIC) (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nat. Genet. *43*, 1131–1138.
49. Pulit, S.L., Stoneman, C., Morris, A.P., Wood, A.R., Glastonbury, C.A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., et al.; GIANT Consortium (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum. Mol. Genet. *28*, 166–174.
50. van der Harst, P., and Verweij, N. (2018). Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ. Res. *122*, 433–443.
51. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. *50*, 1505–1513.
52. Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., Visscher, P.M.; and GIANT Consortium (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum. Mol. Genet. *27*, 3641–3649.
53. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al.; RACI consortium; and GARNET consortium (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.
54. van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods *12*, 1061–1063.
55. Huang, Q.Q., Ritchie, S.C., Brozynska, M., and Inouye, M. (2018). Power, false discovery rate and Winner's Curse in eQTL studies. Nucleic Acids Res. *46*, e133.
56. Davis, J.R., Fresard, L., Knowles, D.A., Pala, M., Bustamante, C.D., Battle, A., and Montgomery, S.B. (2016). An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. Am. J. Hum. Genet. *98*, 216–224.
57. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.
58. Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biol. *16*, 195.
59. Yee, T.W. (2015). Vector generalized linear and additive models: with an implementation in R (Springer).
60. Iotchkova, V., Ritchie, G.R.S., Geihs, M., Morganella, S., Min, J.L., Walter, K., Timpson, N.J., Dunham, I., Birney, E., Soranzo, N.; and UK10K Consortium (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat. Genet. *51*, 343–353.
61. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598.
62. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.
63. Mitchelmore, J., Grinberg, N.F., Wallace, C., and Spivakov, M. (2020). Functional effects of variation in transcription factor binding highlight long-range gene regulation by epromoters. Nucleic Acids Res. *48*, 2866–2879.
64. de la Torre-Ubieta, L., Stein, J.L., Won, H., Opland, C.K., Liang, D., Lu, D., and Geschwind, D.H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell *172*, 289–304.e18.
65. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.
66. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.
67. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47* (D1), D1005–D1012.
68. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.
69. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336–2337.
70. Fogarty, M.P., Cannon, M.E., Vadlamudi, S., Gaulton, K.J., and Mohlke, K.L. (2014). Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. PLoS Genet. *10*, e1004633.

71. Hosoda, H., Tamura, H., and Nagaoka, I. (2015). Evaluation of the lipopolysaccharide-induced transcription of the human TREM-1 gene in vitamin D3-matured THP-1 macrophage-like cells. Int. J. Mol. Med. *36*, 1300–1310.

72. Nagaki, M., and Moriwaki, H. (2008). Transcription factor HNF and hepatocyte differentiation. Hepatol. Res. *38*, 961–969.

73. Kim, S., Yu, N.-K., and Kaang, B.-K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. Exp. Mol. Med. *47*, e166.

74. Oishi, Y., and Manabe, I. (2018). Krüppel-Like Factors in Metabolic Homeostasis and Cardiometabolic Disease. Front. Cardiovasc. Med. *5*, 69.

75. Mayran, A., and Drouin, J. (2018). Pioneer transcription factors shape the epigenetic landscape. J. Biol. Chem. *293*, 13795–13804.

76. Lau, E., and Ronai, Z.A. (2012). ATF2 - at the crossroad of nuclear and cytosolic functions. J. Cell Sci. *125*, 2815–2824.

77. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714–719.

78. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.

79. Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., et al.; Global Lipids Genetics Consortium; Myocardial Infarction Genetics (MIGen) Consortium; Geisinger-Regeneron DiscovEHR Collaboration; and VA Million Veteran Program (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet. *50*, 1514–1523.

80. Myokai, F., Takashiba, S., Lebo, R., and Amar, S. (1999). A novel lipopolysaccharide-induced transcription factor regulating tumor necrosis factor alpha gene expression: molecular cloning, sequencing, characterization, and chromosomal assignment. Proc. Natl. Acad. Sci. USA *96*, 4518–4523.

81. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am. J. Hum. Genet. *104*, 65–75.

82. Han, H.-S., Jung, C.-Y., Yoon, Y.-S., Choi, S., Choi, D., Kang, G., Park, K.-G., Kim, S.-T., and Koo, S.-H. (2014). Arginine methylation of CRTC2 is critical in the transcriptional control of hepatic glucose metabolism. Sci. Signal. *7*, ra19.

83. Dütting, S., Brachs, S., and Mielenz, D. (2011). Fraternal twins: Swiprosin-1/EFhd2 and Swiprosin-2/EFhd1, two homologous EF-hand containing calcium binding adaptor proteins with distinct functions. Cell Commun. Signal. *9*, 2.

84. Sheffield, N.C., Thurman, R.E., Song, L., Safi, A., Stamatoyannopoulos, J.A., Lenhard, B., Crawford, G.E., and Furey, T.S. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. *23*, 777–788.

85. Raulerson, C.K., Ko, A., Kidd, J.C., Currin, K.W., Brotman, S.M., Cannon, M.E., Wu, Y., Spracklen, C.N., Jackson, A.U., Stringham, H.M., et al. (2019). Adipose Tissue Gene Expression Associations Reveal Hundreds of Candidate Genes for Cardiometabolic Traits. Am. J. Hum. Genet. *105*, 773–787.

86. Dobbyn, A., Huckins, L.M., Boocock, J., Sloofman, L.G., Glicksberg, B.S., Giambartolomei, C., Hoffman, G.E., Perumal, T.M., Girdhar, K., Jiang, Y., et al.; CommonMind Consortium (2018). Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS. Am. J. Hum. Genet. *102*, 1169–1184.