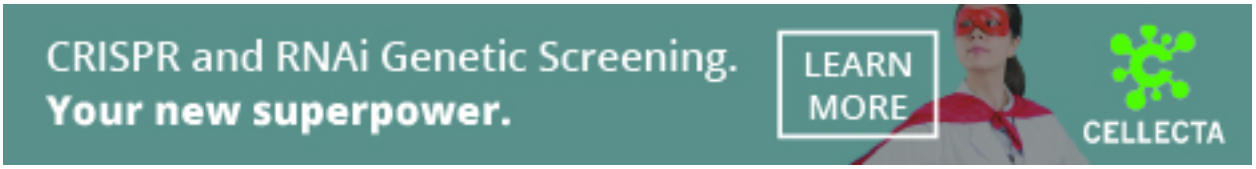
Functional noncoding SNPs in human endothelial cells fine-map vascular trait associations

Anu Toropainen, Lindsey K. Stolze, Tiit Örd, et al.

Genome Res. 2022 32: 409-424 originally published online February 22, 2022
Access the most recent version at doi:10.1101/gr.276064.121



Functional noncoding SNPs in human endothelial cells fine-map vascular trait associations

Anu Toropainen,[1,5] Lindsey K. Stolze,[2,3,5] Tiit Örd,[1,5] Michael B. Whalen,[2] Paula Martí Torrell,[1] Verena M. Link,[4] Minna U. Kaikkonen,[1,6] and Casey E. Romanoski[2,3,6]

[1]*A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio 70211, Finland;* [2]*The Department of Cellular and Molecular Medicine, The University of Arizona, Tucson, Arizona 85721, USA;* [3]*The Genetics Interdisciplinary Graduate Program, The University of Arizona, Tucson, Arizona 85721, USA;* [4]*Metaorganism Immunity Section, Laboratory of Host Immunity and Microbiome, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA*

Functional consequences of genetic variation in the noncoding human genome are difficult to ascertain despite demonstrated associations to common, complex disease traits. To elucidate properties of functional noncoding SNPs with effects in human endothelial cells (ECs), we utilized our previous molecular quantitative trait locus (molQTL) analysis for transcription factor binding, chromatin accessibility, and H3K27 acetylation to nominate a set of likely functional noncoding SNPs. Together with information from genome-wide association studies (GWASs) for vascular disease traits, we tested the ability of 34,344 variants to perturb enhancer function in ECs using the highly multiplexed STARR-seq assay. Of these, 571 variants validated, whose enriched attributes included: (1) mutations to TF binding motifs for ETS or AP-1 that are regulators of the EC state; (2) location in accessible and H3K27ac-marked EC chromatin; and (3) molQTL associations whereby alleles associate with differences in chromatin accessibility and TF binding across genetically diverse ECs. Next, using pro-inflammatory IL1B as an activator of cell state, we observed robust evidence (>50%) of context-specific SNP effects, underscoring the prevalence of noncoding gene-by-environment (GxE) effects. Lastly, using these cumulative data, we fine-mapped vascular disease loci and highlighted evidence suggesting mechanisms by which noncoding SNPs at two loci affect risk for pulse pressure/large artery stroke and abdominal aortic aneurysm through respective effects on transcriptional regulation of *POU4F1* and *LDAH*. Together, we highlight the attributes and context dependence of functional noncoding SNPs and provide new mechanisms underlying vascular disease risk.

[Supplemental material is available for this article.]

Genome-wide association studies (GWASs) have revealed thousands of associations between genetic variants and clinical phenotypes. A large majority of disease-associated loci, and thus functional variants that underpin trait differences, are not protein-coding (Hindorff et al. 2009). This suggests that noncoding disease-associated variants alter transcription through mechanisms such as recruitment of transcriptional activators and/or repressors at regulatory elements, or by changing chromatin conformation. Because regulatory elements, especially enhancers, are frequently cell type–specific (Roadmap Epigenomics Consortium et al. 2015) or restricted by cell state in their activities (Kaikkonen et al. 2013; Ostuni et al. 2013), identification of disease-predisposing cells and tissues within the body remains a barrier toward functional understanding of risk loci.

Despite significant advancements in the catalogs linking human sequence variants to clinical phenotypes and molecular 'omic profiles, a major bottleneck in functional genomics remains—namely, pinpointing functional regulatory variants and their mechanisms of action in relevant biological systems at scale. The identification of functional variants relative to proxy variants in high linkage disequilibrium (LD) presents an additional challenge. Recent studies have begun to map molecular quantitative trait loci (molQTLs) for gene regulatory traits, such as histone modification, transcription factor (TF) binding, and chromatin accessibility (Hogan et al. 2017; Alasoo et al. 2018). These studies have identified allelic differences that associate with quantitative epigenetic differences in *cis*. However, causative roles for regulatory variants require experimental validation by separation from linked variants. To this end, massively parallel reporter assays (MPRAs) represent a high-throughput solution for experimental validation, compared to luciferase or EMSA assays, because several thousand sequences can be tested for regulatory function simultaneously (Arnold et al. 2013; Kheradpour et al. 2013; Zhang et al. 2018). Various MPRA techniques have been developed and implemented to identify genomic sequences and compare allele-specific effects on enhancer activity (Vockley et al. 2015; Tewhey et al. 2016; Ulirsch et al. 2016; Liu et al. 2017; Zhang et al. 2018; van Arensbergen et al. 2019).

In a recent study, we mapped expression quantitative trait loci (eQTLs) and molQTLs for molecular traits in a set of genetically diverse human aortic endothelial cells (HAECs) under basal and pro-inflammatory conditions, mimicked by stimulation with cytokine interleukin 1 beta (IL1B) (Stolze et al. 2020). This thoroughly

[5]These authors are co-first authors and contributed equally to this work.
[6]These authors are co-last authors.
Corresponding authors: cromanoski@arizona.edu, minna.kaikkonen@uef.fi
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276064.121.

phenotyped genetic panel of human cells provides a unique opportunity to pinpoint functional regulatory variants as well as to experimentally define genomic features that best correspond to validation in MPRAs. To this end, we now extend the analysis of this resource by generating a matching MPRA data set using self-transcribing active regulatory region sequencing (STARR-seq). This allowed us to investigate the correlation between allele-specific activity in STARR-seq and molQTLs, identify attributes that best associate with variant effects, and fine-map noncoding variants at GWAS loci.

Results

STARR-seq validates ETS and AP-1 factor motifs enriched in HAEC enhancer elements

In STARR-seq, enhancer function was tested by placement of the 198-base pair oligo downstream from a core promoter (origin of replication, ORI), followed by a poly-adenylation (poly[A]) track, such that oligos that enhance transcription from the promoter can be identified by sequencing of the resulting noncoding RNAs (Fig. 1A). A total of 34,344 bi-allelic variants were selected

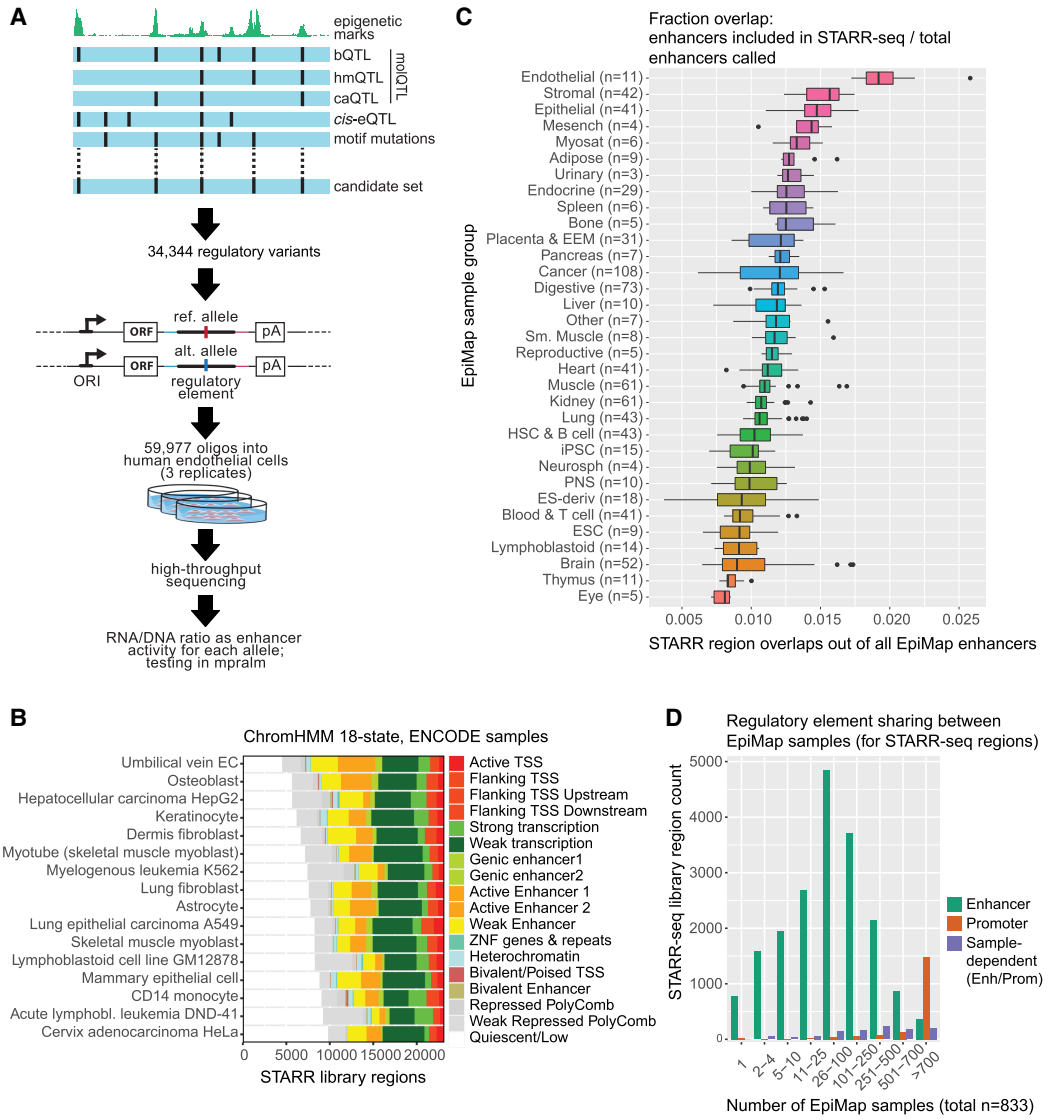


Figure 1. Construction and characterization of the STARR-seq library. (*A*) Schematic of the STARR-seq library design. A candidate set of 198-bp oligos was selected based on their overlap with binding (b)QTLs, histone mark (hm)QTLs, chromatin accessibility (ca)QTLs, and *cis*-eQTLs. Reference and alternative variants were cloned into a plasmid vector under origin of replication core-promoter. The reporter library was transfected into cultured teloHAECs, and deep sequencing was conducted. The enrichment of reporter RNA expression over the input DNA library directly and quantitatively reflects enhancer activity. (*B*) Chromatin state distribution of the STARR-seq library regions across the core set of ENCODE samples. STARR-seq library regions were intersected with previously published genome-wide chromatin partitioning into 18 chromatin states by ChromHMM. (*C*) STARR-seq library regions that overlap an enhancer, relative to the total number of enhancers called in the sample. Enhancer coordinates and sample grouping were obtained from EpiMap. (*D*) Sharing of STARR-seq library enhancers and promoters across a diverse set of epigenomes. STARR-seq library regions were intersected with the ChromHMM chromatin states (18-state model) for every sample in EpiMap, an epigenetic compendium spanning the human body. For each STARR-seq region, the number of samples with a promoter or enhancer overlap was counted. The different enhancer states (active, bivalent, genic, and weak) were combined, as were the promoter states (active TSS, bivalent TSS, flanking TSS, flanking TSS upstream, and flanking TSS downstream).

to test for allele-specific enhancer activity in our STARR-seq library. Alleles were either present as the sole variant nucleotide, when no other variants in the 198-bp window were present (30% of oligos), or in existing European haplotypes when additional polymorphic variants (5% minor allele frequency [MAF]) were located within the window (63% of oligos; up to the top five most frequent haplotypes represented, based on 1000 Genomes EUR reference population) (Supplemental Fig. S1C). SNP selection was based on one of two approaches. First, we utilized our recent QTL analyses (Stolze et al. 2020) to prioritize variants (58% SNPs in the library) using combinations of the following attributes: (1) significance as a SNP underlying molQTLs (chromatin accessibility [caQTL], histone H3 lysine 27 acetylation (H3K27ac; hmQTL), and DNA binding of the TFs ERG and RELA (a component of NF-kB; bQTL) and/or eQTLs (RNA-seq and microarray) in HAECs; (2) allele-specific mutation to TF binding motifs; and (3) GWAS for coronary artery disease (CAD) (Supplemental Fig. S1B; Methods). Second, GWAS variants associated with CAD, myocardial infarction, and type 2 diabetes and overlapping accessible chromatin elements in several disease-relevant cell types were selected (41% SNPs in the library) (Methods). Additionally, 265 control regions were also included, resulting in 59,976 unique oligos (Supplemental Fig. S1A).

To further categorize the regions included in the STARR-seq library, we annotated the library utilizing the 18-state chromatin annotation model generated using ChromHMM for the Encyclopedia of DNA Elements (ENCODE) core sample set (Boix et al. 2021). Approximately 67% of the STARR-seq library regions are categorized into one of the active chromatin states in human umbilical vein ECs (HUVECs), with ~36% as enhancers, 7% as transcription start sites (TSSs)/promoters, and 24% as other transcribed regions (Fig. 1B). Using enhancer and promoter regions defined by the EpiMap compendium (Boix et al. 2021), where chromatin states were surveyed across 833 samples spanning the human body, 82% of STARR-seq library regions overlap an enhancer in at least one cell or tissue type, 8% overlap a promoter, and 5% overlap regions that are variably annotated as promoter or enhancer depending on the sample (Fig. 1C,D). Enhancer annotations of the STARR-seq regions are less frequently shared between multiple cell types than annotations of promoters (Fig. 1D). The fraction of STARR-seq oligos overlapping active enhancers differs across different cell types and tissues, with the highest representation seen for endothelial enhancers (Fig. 1C), whereas promoter representation by tissue or cell type is much less variable (Supplemental Fig. S2). These data are consistent with the promoter chromatin state being less variant across cell types and tissues than enhancers and demonstrates that our STARR-seq library was enriched for endothelial cell regulatory elements relative to other cell types.

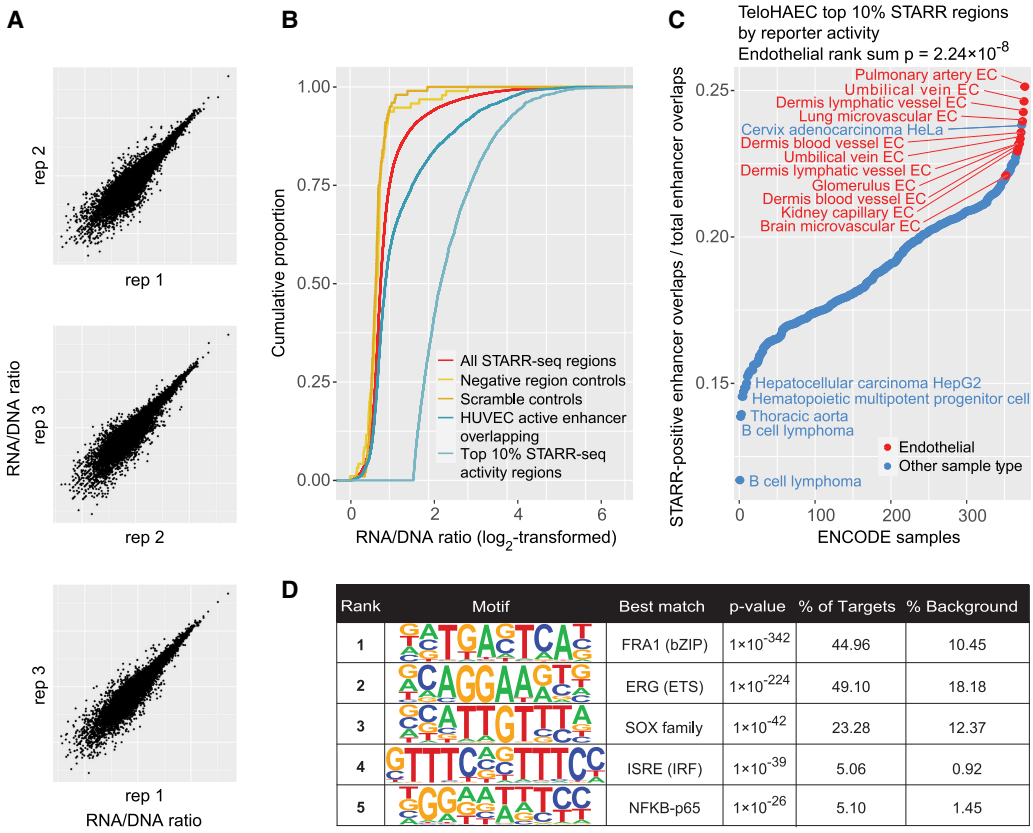
The STARR-seq library was transfected into teloHAECs (HAECs immortalized with hTERT) and enhancer activity was quantified as the ratio of transcripts from each oligo relative to the amount of DNA plasmid for that oligo (i.e., RNA/DNA ratio). As shown in Figure 2A and Supplemental Figure S3, reporter activity for individual oligo sequences was highly concordant across triplicate biological replicates. Oligos from genomic loci exhibiting active enhancer elements in HUVECs (from Fig. 1B) had greater enhancer activity by STARR-seq compared to both oligos derived from negative control regions (i.e., lacking any ENCODE [The ENCODE Project Consortium 2012] active chromatin marks) and scrambled sequences (Fig. 2B). This demonstrates that regions with enhancer activity in the STARR-seq assay were indicative of EC gene regulation.

To evaluate the cell type–specificity of STARR-seq signal strength, we intersected STARR-seq regions with the enhancer regions of more than 350 ENCODE samples individually and tabulated how many ranked within the top 10% of active STARR-seq regions as measured in teloHAECs. The results reveal that enhancers demarked in EC samples are more likely than other cell types to generate high STARR-seq activity in teloHAEC ($P = 2.24 \times 10^{-8}$ for ECs vs. others by Wilcoxon rank-sum test) (Fig. 2C). Further, de novo motif analysis of the upper 10th percentile of STARR-seq regions by enhancer reporter activity uncovered enrichment for the AP-1 (FRA1), ERG (ETS), and SOX family TF motifs when compared to all other regions in the STARR-seq library (Fig. 2D). This is consistent with our previous reports that these motifs are enriched in HAEC enhancers in the native chromatin context (Kaikkonen et al. 2014; Hogan et al. 2017). More specifically, the ETS motif was frequently bound by ERG in HAECs, which is an essential TF that regulates vascular development in mice (Lathen et al. 2014; Birdsey et al. 2015). Similarly, the AP-1 motif was also found as enriched at HAEC enhancers and bound by JUN (Hogan et al. 2017). These data provide evidence of concordance between enhancer activity measured by STARR-seq in teloHAECs, by ChIP-seq in the genomic context of HAECs, and by ECs from other tissues. Based on the analyses of enhancer activity, we expect that the STARR-seq system is a reliable means to quantify allele-specific effects of common genetic variation on the EC regulome.

Among sequence attributes, TF motif mutations are most enriched for allele-specific regulatory activity

To assess differential enhancer activity between alleles in the STARR-seq library, we employed the mpralm method (Myint et al. 2019) from the mpra R package that uses linear models to analyze MPRA data (Law et al. 2014). Of the 34,344 variants tested for allele-specific activity, 3829 variants (11.14% of variants quantified) were significant at 5% FDR in untreated teloHAECs (Fig. 3A). For simplicity, these variants are referred to as STARR-seq "validated" variants.

To test whether sequence-based genomic features differentiated validated from unvalidated variants, we evaluated the following metrics or test variables: (1) GC content in the STARR-seq oligos; (2) genomic distance of variants to the nearest TSS; (3) average sequence conservation across vertebrates in oligo sequences; and (4) SNPs that mutate TF binding motifs. Enrichment scores (ESs) were calculated in each analysis by dividing the observed number of the test variable (e.g., GC content between 0.8 and 1) in the validated STARR-seq SNP set by the number that we would expect by the number of the test variable included in the STARR-seq library. Enrichments were verified using statistical testing by a hypergeometric test (Methods). Of the genomic features tested, we observed that oligos with GC content between 0.368 and 0.478 tended to harbor SNPs with allele-specific enhancer activity; similarly, regions with greater conservation scores were more likely to validate, and variants that were farther from TSSs were also more likely to validate than the other groups (Supplemental Fig. S4A). Notably, regions with 60%–70% GC content were underrepresented in the library with a concurrent increase in data variability, whereas GC > 80% was rarely detected, suggesting that high GC content limits the detection accuracy (Supplemental Fig. S4B,C). Still, only 2% of the input library had GC > 70%, which is why we expect this to minimally affect downstream analysis. Additionally, TF motif-mutating SNPs were significantly associated, as a set, with allele-specific regulatory activity (Fig. 3G); we



Rank	Motif	Best match	p-value	% of Targets	% Background
1		FRA1 (bZIP)	1×10^{-342}	44.96	10.45
2		ERG (ETS)	1×10^{-224}	49.10	18.18
3		SOX family	1×10^{-42}	23.28	12.37
4		ISRE (IRF)	1×10^{-39}	5.06	0.92
5		NFKB-p65	1×10^{-26}	5.10	1.45

Figure 2. Enhancer activity profile of the STARR-seq reporter regions in teloHAEC cells. (*A*) STARR-seq replicate correlations for the measured reporter activity of library inserts normalized to DNA amounts in the input plasmid pool (the RNA/DNA ratio). (*B*) Cumulative distribution of enhancer activity (RNA/DNA ratio) for oligos overlapping active enhancers in HUVECs, compared to 100 scrambled regions,100 negative region oligos (that did not overlap any active chromatin marks in the ENCODE study). The activity distributions of the all STARR-seq regions and the top 10% most active regions are also shown. When selecting the top 10% most active STARR-seq regions, only the allele with the highest activity was considered for each region. (*C*) For different epigenomes, the fraction of enhancer-overlapping STARR-seq regions that are among the top 10% most active STARR-seq regions, as measured in teloHAECs. The active enhancers of each ENCODE sample were intersected with STARR-seq library regions to obtain the epigenome-specific set of enhancer overlapping oligos, which was then intersected with overall strongest reporter signals from teloHAEC STARR-seq. The top 10% regions were selected as in panel *B*. (*D*) De novo motif analysis of the STARR-seq regions showing enhancer activity within the upper 10th percentile in teloHAECs. The top 10% regions were selected as in panel *B*, and motif enrichment was calculated against all other regions in the library.

reasoned this relationship was driven by a particular set of motifs corresponding to EC-lineage determining TFs. Among the TF motifs with strongest associations were members of the AP-1 family (AP-1, BATF, ATF3/7), as well as members of the ETS family (EWS, ETV2, ERG, ETS1) (Fig. 3B). Mutations to both AP-1/ATF3 and ETS/ERG motifs were enriched in validated SNPs over expectations (ES scores). In addition to qualitative enrichment, we also observed that SNPs whose alleles produced a greater deviation from an established position weight matrix (PWM) achieved more significant STARR-seq allelic *P*-values (Fig. 3C,D). These data are consistent with AP-1 and ETS interactions at these loci being important for HAEC regulatory function.

It is reported that support vector machine (SVM) models applied to in vitro TF-DNA binding measurements from SNP evaluation by systematic evolution of ligands by exponential enrichment (SNP-SELEX) experiments more accurately explain effects of allelic mutations to TF motifs than position weight matrices for some TFs (Yan et al. 2021). Therefore, we tested if SVM models were better indicators of allelic validation in our STARR-seq data set than PWM mutation results. All 94 TF binding models that were considered to be high-confidence by the original authors were used with reported thresholds for affecting binding. Among the 30,792 SNPs queried, TFs were predicted to bind 28,885, and among those, 5615 SNPs demonstrated predicted allele-dependent gain or loss of binding. Besides ETS family factors, SVM-predictions were most successful for AP-1- and CREB-like motifs (JDP2, ATF3, CREB1, etc.) (Supplemental Fig. S5). We observed a significant correlation between SVM-based and PWM-based predicted allelic effects for many motifs, including ERG and ATF3 (Fig. 3E,F). Though predicted effects of both approaches were enriched in validated SNPs, SVM effects were slightly more associated with validation in STARR-seq. These data underscore the utility for both PWM and SVM methods for identification of functional SNPs in enhancers.

Among epigenetic attributes, transcription factor binding most significantly associates with allelic effects in STARR-seq

Epigenetic marks such as chromatin accessibility and histone modifications that are measured in native chromatin contexts are frequently used to signify genomic loci with enhancer function and to prioritize functional regulatory SNPs. By relating validated

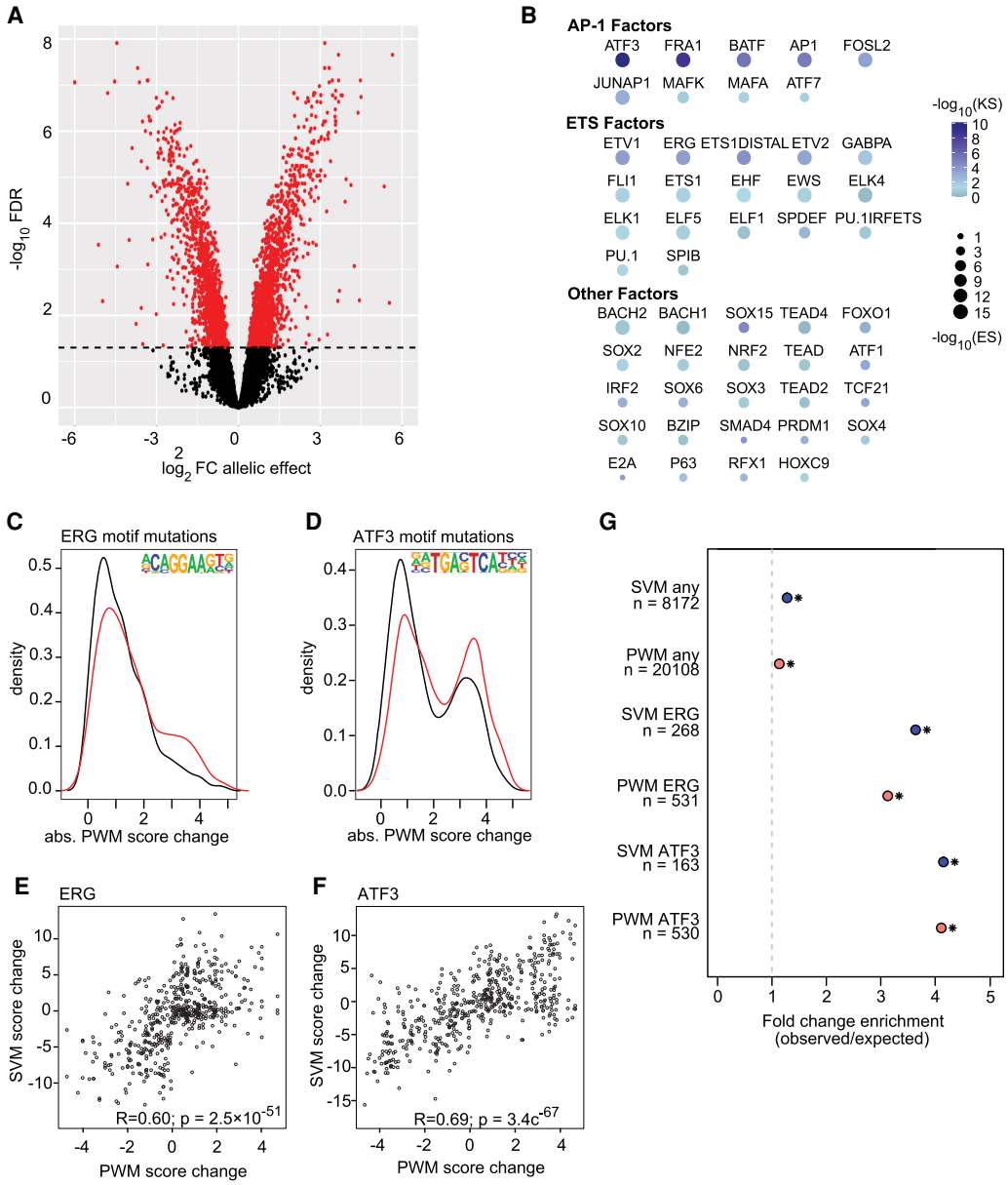


Figure 3. Allele-specific enhancer function is associated with motif mutations. (*A*) Volcano plot of allele-specific effects in STARR-seq with the $\log_2$ fold change on the *x* axis and $-\log_{10}$ false discovery rate on the *y*-axis. All oligos, representing both reference and alternative alleles for each variable position, from each replicate were analyzed. (*B*) Dot-plot presenting the hypergeometric enrichment –log(*P*-value) and KS testing –log(*P*-value) of the top 50 most enriched motifs mutated in the validated set from STARR-seq. Motifs that are directly for ETS family or AP-1 family transcription factors are indicated. (*C*) PWM motif mutation score density for ERG motif mutations in validated (red) and nonvalidated (gray). (*D*) PWM motif mutation score density for ATF3 motif mutations in validated (red) and nonvalidated (gray). (*E*) PWM motif mutation score (*x*-axis) versus the SVM motif mutation score for ERG motif mutation SNPs. (*F*) PWM motif mutation score (*x*-axis) versus the SVM motif mutation score for ATF3 motif mutation SNPs. (*G*) Enrichment of SVM (blue) and PWM motif mutations scores (orange) in the STARR-seq-validated set.

STARR-seq SNPs with genomic data measured in ECs, we sought to better understand what genomic features are most associated with validation. To achieve this, we utilized our previously generated chromatin accessibility data by ATAC-seq among 44 genetically distinct HAEC donors, H3K27ac measured by ChIP-seq across 42 HAEC donors, and the binding of the EC-relevant TF, ERG by ChIP-seq across 22 HAEC donors (Stolze et al. 2020). Of SNPs that were included in our STARR-seq library, 18,457 were in a chromatin accessibility peak, H3K27ac peak, or ERG binding peak within HAECs, and 2615 of these validated with allele-specific enhancer effects by STARR-seq (Fig. 4A). Upon calculating enrichment scores, we initially found an enrichment of STARR-seq-validated SNPs in HAEC-accessible regions, H3K27ac marked regions, and ERG binding regions; however, SNPs at H3K27ac peaks were less enriched for functional effects (Fig. 4B). This led us to consider one major distinction between these data types; namely, the size of the peaks. ATAC-seq and ERG binding by ChIP-seq generates focal peaks (median 90 bp and 280 bp, respectively), whereas H3K27ac ChIP-seq produces distributed peaks (median 1492 bp) (Fig. 4C). We hypothesized that functional

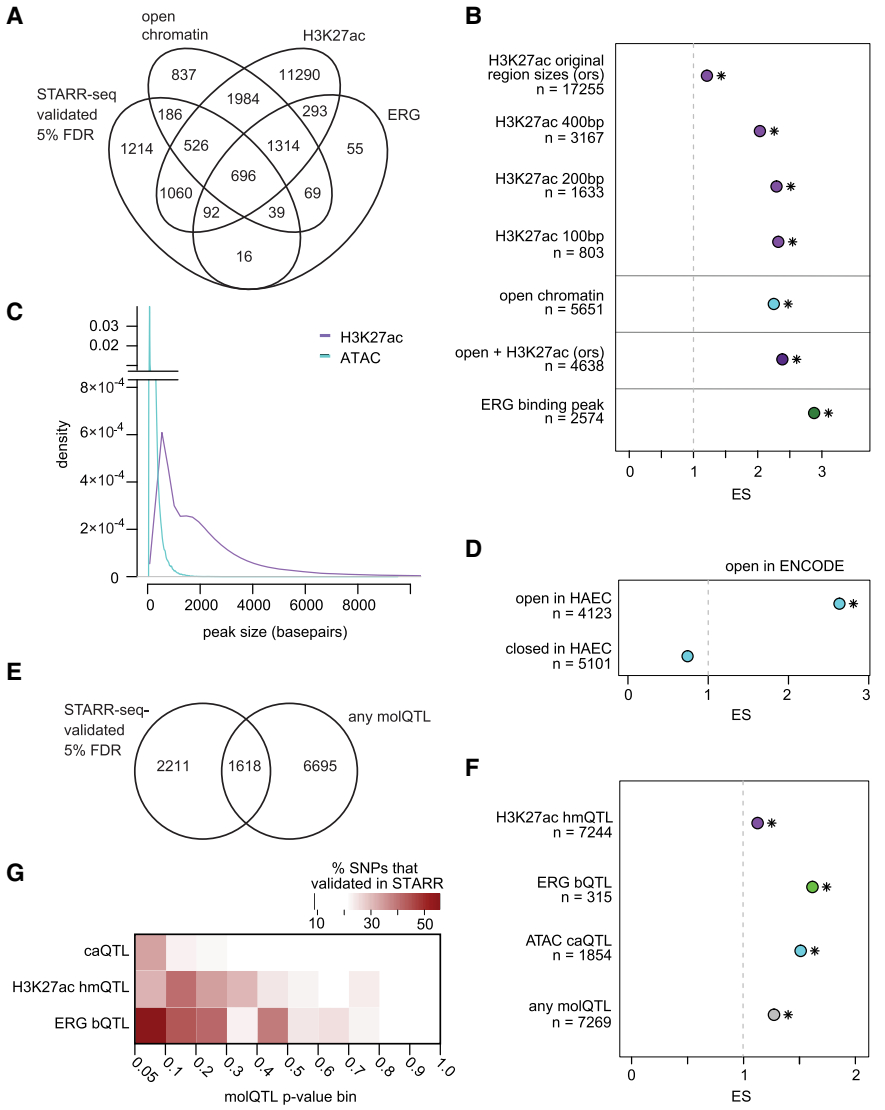


Figure 4. Epigenetic traits are linked to validation in STARR-seq. (*A*) Venn diagram comparing the number of SNPs in an open chromatin region in HAECs (*bottom*), in an H3K27ac peak (*right*), or validated in STARR-seq (*left*). (*B*) Enrichment of SNPs in epigenetic trait peaks in the STARR-seq-validated SNP set. H3K27ac (*top* four) restricted to various sizes around the center of the peak. (*C*) Peak size density of ATAC-seq peaks and H3K27ac ChIP-seq peaks. (*D*) Enrichment of SNPs in ENCODE open regions that are shared and not shared with HAECs. (*E*) Venn Diagram comparing the number of SNPs that are molQTLs (*right*) and validated in STARR-seq (*left*). (*F*) Enrichment of molQTL SNPs in STARR-seq-validated SNPs. (*G*) Heat map of number of STARR-seq-validated SNPs in subsignificant molQTL bins.

SNPs are more likely to occur in H3K27ac regions when they are located near nucleosome-free regions (NFRs), where TFs bind DNA. Restricting H3K27ac regions around their centers (100 bp, 200 bp, and 400 bp) improved enrichment scores for STARR-seq-validated variants, demonstrating that SNPs closer to the center of the H3K27ac regions are more likely to perturb enhancer function than those further away (Fig. 4B). This is consistent with ERG binding having the highest enrichment among epigenetic attributes, which would bind in the NFR.

We next asked whether SNPs in accessible regions in ECs were more likely to validate than SNPs that were open in another cell type but closed in ECs. Using DNase I hypersensitivity peaks from 95 cell types in ENCODE, we found that open regions shared with HAECs are far more likely to validate than SNPs in open regions that are closed in HAECs (Fig. 4D). Extending this analysis to 350 ENCODE samples from EpiMap (Boix et al. 2021), variants residing in EC enhancers showed a higher proportion of significant allelic effects than other samples ($P = 1.29 \times 10^{-5}$ for EC samples compared to others using the Wilcoxon rank-sum test) (Supplemental Fig. S6A). Along these same lines, we observed diminished enrichment for HAEC-accessible SNPs when the same STARR-seq library was transfected into HepG2 (Supplemental Fig. S6B; Selvarajan et al. 2021). These data are consistent with previous reports demonstrating that cell type is an important determinant for which regulatory elements, motifs, and allele-specific effects are detectable (Tewhey et al. 2016).

SNPs underlying molecular quantitative trait loci frequently validate in STARR-seq

MolQTL mapping is one approach to identify functional noncoding variants. In molQTL analysis of epigenetic data, alleles/genotypes are tested for association with differences in quantitative epigenetic traits measured at a given locus. Significant molQTLs provide evidence for functional regulatory SNPs and often suggest a mechanism of action (e.g., an allele that perturbs TF binding). In prior work, we identified molQTLs for chromatin accessibility, H3K27ac histone modification (hmQTLs), and ERG binding (bQTLs) in untreated HAECs (Stolze et al. 2020). Of the SNPs input into our STARR-seq library, 8313 were significant molQTLs (at 5% FDR), with 1618 of these validating by STARR-seq (Fig. 4E). We found that each set of molQTLs was significantly enriched for validation by STARR-seq, with modest but significant correlation between STARR-seq validation statistics and molQTL effect sizes (Fig. 4F; Supplemental Fig. S7A,B,D,E,G,H). It is important to note that the enrichments we observe for molQTLs are in addition to the enrichments noted for epigenetic marks alone (Fig. 4B). This is because only SNPs within the epigenetic peaks (H3K27ac ChIP-seq peaks; ATAC-seq peaks; TF peaks) were tested in molQTL analysis. For example, of SNPs polymorphic in the HAEC population at 5% MAF, 0.57% reside in ERG binding peaks in the human genome; however, 7.65% of SNPs tested in our STARR-seq are in ERG genomic peaks (13.4-fold enrichment over genomic). In contrast, we observed that 22% of STARR-seq-validated SNPs are in ERG genomic peaks (2.9-fold over expectation). Further, only SNPs in ERG peaks were included in ERG bQTL analysis, and only 1.5% of SNPs in ERG peaks are significant ERG bQTLs.

Table 1. Percentages of SNPs meeting epigenetic, molQTL, and STARR-seq library criteria

	ERG		ATAC		H3K27ac	
	In peaks	bQTLs (in peaks)	In peaks	caQTLs (in peaks)	In peaks	hmQTLs (in peaks)
Of all genotyped SNPs in HAECs (at MAF > 5% in given data set)	0.5724% (36,478/6,372,621)	0.00874% (557/6,372,614)	1.6593% (110,509/6,659,935)	0.05863% (3905/6,659,935)	6.1126% (404,620/6,619,412)	0.3871% (25,621/6,619,412)
Of mutually tested SNPs in the STARR-seq library	7.6475% (2574/33,658)	14.025% (315/2246)	16.7895% (5651/33,658)	39.0316% (1854/4750)	51.2657% (17,255/33,658)	49.5587% (7244/14,617)
Of allele-specific SNPs by STARR-seq	22.0162% (843/3829)	22.72% (177/779)	37.7906% (1447/3829)	59.1177% (804/1360)	62.0005% (2374/3829)	55.9981% (1195/2134)
Of SNPs in given data set's peaks	N/A	1.5269% (557/364,78)	N/A	3.5336% (3905/110,509)	N/A	6.3321% (25,621/404,620)

Among SNPs mutually tested in ERG bQTL analysis and our STARR-seq library, 14% are significant ERG bQTLs. In contrast, we observed that 23% of STARR-seq-validated SNPs are ERG bQTLs (1.64-fold over expectation). These values are provided for each epigenetic mark analyzed in Table 1.

Among molQTLs, ERG bQTLs were most enriched, supporting ERG's importance in EC gene regulation (Fig. 4F). This result was replicated for bQTLs and caQTLs using Kolmogorov–Smirnov (KS) testing ($P < 0.001$) (Supplemental Fig. S7C,F,I). The relatively lower enrichment of hmQTLs could be explained by the broader size of peaks compared to bQTLs and caQTLs reducing the power to detect enrichment or histone modifications being less sequence-dependent than TF binding (Heinz et al. 2013; Huang and Ovcharenko 2015). Based on these analyses, we conclude that molQTLs effectively prioritize functional variants with focal epigenetic traits as most significantly associated. Still, we were interested that there is a subset of validated SNPs that are not significant molQTLs. We suspect this could be due to low statistical power, given the modest sample size of individuals submitted to molQTL analysis (n~20–50). Consistent with this hypothesis, we observed an increased concentration of validated STARR-seq SNPs in P-value bins just above the molQTL significance threshold (Fig. 4G), leading us to conclude that suboptimal molQTLs power in part explains the incomplete overlap between molQTLs and STARR-seq validated SNPs.

Allele-specific enhancer activity reveals interactions between genetic variation and environment

Gene-by-environment (GxE) interactions are one mechanism by which alleles can protect or predispose individuals to develop complex traits, including disease. In this study, we modeled the effect of an inflammatory environment on endothelial cells in culture by pro-inflammatory cytokine IL1B stimulation, which is a known hallmark and driver of disease progression, such as in atherosclerosis (Ridker et al. 2017). Using the same STARR-seq library, we measured enhancer activity in teloHAECs exposed to IL1B for 6 and 24 h, which loosely mimics early and late transcriptional responses to inflammation (Fig. 5A). In total, 3297 regions (5.6% of the 59,058 regions tested) had differential enhancer activity between treatments (2695 up-regulated by IL1B treatment, 602 down-regulated by IL1B treatment) (Supplemental Fig. S8A,B). As expected, the NF-kB motif was enriched in the regions with increased enhancer activity after IL1B treatment (Supplemental Fig. S8C; Hogan et al. 2017). Likewise, the ERG motif was enriched in regions with decreased enhancer activity after IL1B treatment (Supplemental Fig. S8C). This could be explained in part by the down-regulation of ERG protein upon pro-inflammatory stimulation (Yuan et al. 2009).

We identified 5711 SNPs with allele-specific enhancer activity in at least one IL1B treatment condition (0-h, 6-h, or 24-h) with

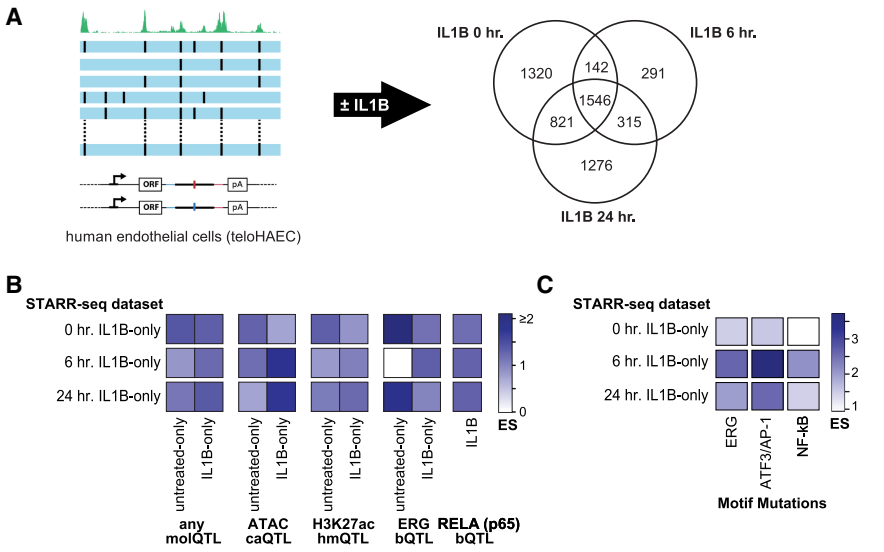


Figure 5. IL1B treatment trends with molQTLs and motif mutations. (*A*) Diagram of treatment design and Venn diagram of SNPs that validated in untreated (*left*), IL1B-treated for 6 h (*right*), and IL1B-treated for 24 h (*bottom*). (*B*) Heat map of enrichment scores comparing SNPs uniquely in specific treatment conditions between STARR-seq and the molQTLs (i.e., SNPs only significant in STARR-seq 0-h and not in 6-h or 24-h IL1B-treated). (*C*) Heat map of enrichment scores comparing SNPs uniquely in specific treatment conditions between STARR-seq and the motif mutations.

27% of these having shared effects across all treatment conditions (Fig. 5A). The 6-h IL1B treatment had the least number of significant results ($n = 2294$ SNPs), with the majority being similarly significant in either the 0-h or 24-h treatments. Thus, the 0-h and 24-h treatments uncover most SNPs that perturb enhancer activity. This could be indicative of the slightly lower similarity across 6-h replicates, or perhaps reflect a limitation of our bulk assay when in fact individual cells are in various transition activation states (Supplemental Fig. S3).

We next evaluated relationships between allelic effects by STARR-seq (IL1B at 6 and 24 h) and previously published molQTL results that were measured in HAECs after 4 h IL1B treatment (Stolze et al. 2020). Though there was widespread enrichment across the STARR-seq data sets for several molQTL data sets, we did observe differences with intriguing implications. As expected, IL1B- restricted caQTLs validated with a higher ES in IL1B-restricted STARR-seq data sets than in the untreated STARR-seq data set. However, HAEC ERG bQTLs restricted to basal conditions were enriched by STARR-seq in 0-h and 24-h IL1B treatment. In contrast, HAEC ERG bQTLs that were uniquely detected after IL1B treatment were most enriched in the STARR-seq-validated SNPs at the 6-h IL1B treatment (Fig. 5B). We also observed different rates of TF motif mutations (Fig. 5C), where both ERG and AP-1 motif mutations were most pronounced for SNPs with distinct allelic effects in the 6-h STARR-seq time point, whereas NF-kB motif mutations were preferentially evident in both IL1B STARR-seq time points compared to 0 h. Taken together, these data demonstrate the prevalence of GxE on enhancer activity and underscore that SNPs affecting chromatin accessibility and binding of TFs can exert their effects only in particular cellular activation states.

Validated functional SNPs fine-map GWAS signals for vascular diseases including abdominal aortic aneurysm and large artery stroke

To identify functional noncoding SNPs at GWAS loci, we cross-referenced our data generated in ECs with SNPs underlying GWAS signals for traits with appreciated vascular etiology. After restricting SNPs to those that validated by STARR-seq with at least one significant molQTL and an association (lead or proxy) with a vascular GWAS trait, we report 89 high-confidence functional regulatory SNPs that are associated with a variety of complex vascular diseases (Supplemental Fig. S9A; Supplemental Table S1), 14 of which are associated with coronary artery disease (Table 2; Supplemental Table S2). One of these SNPs, rs17114036, has already been shown to alter enhancer function for the shear stress-induced transcript phospholipid phosphatase 3 (*PLPP3* [previously known as *PPAP2B*]), thus serving as a positive control in our approach (Table 2; Krause et al. 2018). We also see evidence for functional noncoding SNPs at the 9p21 CAD locus and the rs17293632 SNP in the SMAD family member 3 (*SMAD3*) locus that has been extensively characterized in smooth muscle cells of the vascular wall (Miller et al. 2016; Turner et al. 2016). We find evidence that the *SMAD3* intronic SNP rs17293632 affects enhancer activity in ECs, and we further confirm that CRISPR-mediated deletion of the overlapping enhancer leads to significant reduction in the *SMAD3* expression in teloHAECs (Supplemental Figs. S9, S10). However, lack of an eQTL in HAECs at rs17293632 and conflicting eQTL directions in GTEx obscures mechanistic interpretations.

Two loci exhibited exceptional evidence for functional regulatory SNPs at enhancers that regulate target genes and modulate risk for (1) abdominal aortic aneurysm, and (2) large artery stroke and pulse pressure. Here, we present evidence for functional noncoding SNPs rs13385499 and rs13382862 at the lipid droplet associated hydrolase (*LDAH*) gene locus, and rs4304924 at the POU class 4 homeobox 1 (*POU4F1*) locus.

At a GWAS locus for abdominal aortic aneurysm (AAA) (Jones et al. 2017), two SNPs, rs13385499 and rs13382862, validated in STARR-seq at all time points and underlie RELA bQTLs in HAECs (Table 2; Supplemental Fig. S11A–C). SNPs rs13385499 and rs13382862 are in LD with each other (EUR $R^2 = 0.79$; HAEC population $R^2 = 0.89$) (Machiela and Chanock 2015). Both SNPs cause multiple motif mutations, including for RUNX (rs13385499) and ERG (rs13382862) motifs, which directionally coincide with the RELA bQTL (Supplemental Fig. S11D). These SNPs are at the 3′ end of the *LDAH* gene and demonstrate association with transcript expression levels of *LDAH* in HAECs (Supplemental Fig. S11A,E). In addition, both SNPs are eQTLs for *LDAH* in the same direction in multiple GTEx tissues including for artery aorta (Supplemental Fig. S11F). *LDAH* has been associated with cholesterol deposition in atherosclerotic plaques (Goo et al. 2014); however, the role of *LDAH* in ECs has not been well described. *LDAH* is highly expressed in teloHAECs, enabling us to test whether this enhancer region specifically regulates *LDAH* expression. Targeted deletion of the 456-bp enhancer carrying both SNPs in teloHAECs resulted in decreased *LDAH* expression with no statistically significant effect on other nearby genes (Supplemental Table S3; Supplemental Figs. S10A, S11G). Unfortunately, GWAS summary statistics for the AAA GWAS are not publicly available, precluding evaluation of whether or not the eQTL and GWAS signals are likely to be caused by the same genetic signal.

Another interesting locus is located on Chromosome 13 at a pleiotropic GWAS signal for large artery stroke and pulse pressure (Dichgans et al. 2014; Evangelou et al. 2018). We identified that alleles of rs4304924 co-associate with *POU4F1* transcript abundance in HAECs and in arterial tissues of GTEx (Fig. 6A,B). Importantly, rs4304924 is the lead SNP at this locus for both pulse pressure, large artery stroke, and *POU4F1* expression in HAECs and GTEx, strongly supporting rs4304924 as causal (Fig. 6A; Supplemental Fig. S12). We observe marks of a regulatory element, demarked by chromatin accessibility, ERG binding, and H3K27ac at the genomic locale of rs4304924, which is located ~61 kb upstream (Fig. 6A, beneath locusZoom plots). Furthermore, rs4304924 is an hmQTL ($q = 3.4 \times 10^{-8}$) (Fig. 6C) in HAECs with the reference G allele predicted to mutate a CRX/GSC homeobox TF motif (Fig. 6D). Lastly, the alternate A allele demonstrated significantly greater enhancer activity by STARR-seq in all time points tested (Fig. 6E). *POU4F1* is a gene that encodes for a homeobox transcription factor with described roles in neuronal differentiation (Fedtsova and Turner 1995) and yet unknown roles in vasculature. Taken together, our data prioritize further inquiry into what role *POU4F1* plays in the arterial vasculature.

Discussion

In this study, paired MPRA and epigenetic data from genetically diverse HAECs enabled a series of discoveries. Importantly, it allowed us to specify which genomic features most often describe SNPs with functional effects on enhancer activity. Our findings suggest that functional noncoding SNPs likely have one or more of the following attributes: (1) an allelic mutation to a TF binding motif for a factor with an important role in the cell type and/or cell state of interest; (2) genomic location in a region marked by

Table 2. Select list of credible functional noncoding SNPs

rsID	Ref	Alt	Chr	position (hg19)	eQTL genes (FDR<0.05) Both HAEC-Only GTEx-Only	Type molQTL [.n = notx, .i = il1b] (q-value, higher tag allele)	Motif Mutated PWM	Predicted transcription factor binding distruption (SVM)	GWAS	STARR-seq pvals [<0.05 (Higher expression allele)] 0h 6h 24h
rs6475604	T	C	Chr 9	22052734	 *CDKN2A | CDKN2B*	ATAC.n (0.02, **T**)		YY2	Glaucoma|CAD	4.19×10^{-2} **(T)** 0.227 0.234
rs10757267	G	C	Chr 9	22052810	 *CDKN2B*	RELA.i (0.02, **G**)			Glaucoma|CAD	9.78×10^{-5} **(G)** 3.45×10^{-3} **(G)** 5.09×10^{-4} **(G)**
rs17293632	C	T	Chr 15	67442596	 *SMAD3 | AAGAB | PIAS1*	ATAC.n (2.5×10^{-3}, **C**)| RELA.i (4.2×10^{-3}, **C**)	NKX6.1|LHX2|ELF5|EHF|LHX3|LHX1| FOSL2|JUNAP1|FRA1|BATF|AP1|ATF3	ATF3|JDP2|NFE2	CAD|Allergic rhinitis|Asthma|Atopic asthma|Chronic inflammatory diseases ankylosing spondylitis Crohn's disease psoriasis primary sclerosing cholangitis ulcerative colitis pleiotropy|Crohn's disease|Eosinophil counts|Hay fever and or eczema|Inflammatory bowel disease|Ulcerative colitis	1.10×10^{-4} **(C)** 9.66×10^{-3} **(C)** 1.01×10^{-3} **(C)**
rs13385499	T	C	Chr 2	20882413	*LDAH* 	RELA.i (0.024, **T**)	CRX|ZNF264|TLX|OTX2|GSC|RUNX1		Abdominal aortic aneurysm	4.13×10^{-5} **(T)** 8.57×10^{-4} **(T)** 1.15×10^{-4} **(T)**
rs13382862	A	G	Chr 2	20882449	*LDAH* 	RELA.i (0.029, **A**)	AP1|BATF|ATF3|FRA1|FOSL2|NKX3.1| JUNAP1|SMAD3|TBX5|ERG	ERG|FLI1	Abdominal aortic aneurysm	4.13×10^{-5} **(A)** 8.57×10^{-4} **(A)** 1.15×10^{-4} **(A)**
rs4304924	G	A	Chr 13	79238925	*POU4F1 | OBI1* *OBI-AS1* 	H3K27ac.n (3.4×10^{-8}, **A**)|H3K27ac.i (0.011, **A**)	CRX|GSC|GATA3|GATA4		Large artery stroke|Pulse pressure	1.54×10^{-3} **(A)** 0.12516 2.91×10^{-2} **(A)**
rs17114036	A	G	Chr 1	56962821	*PLPP3* 	RELA.i (0.03, **G**)			Coronary artery disease|Coronary artery disease or ischemic stroke|Coronary artery disease or large artery stroke|Coronary heart disease	1.69×10^{-2} **(G)** 3.68×10^{-3} **(G)** 3.49×10^{-3} **(G)**

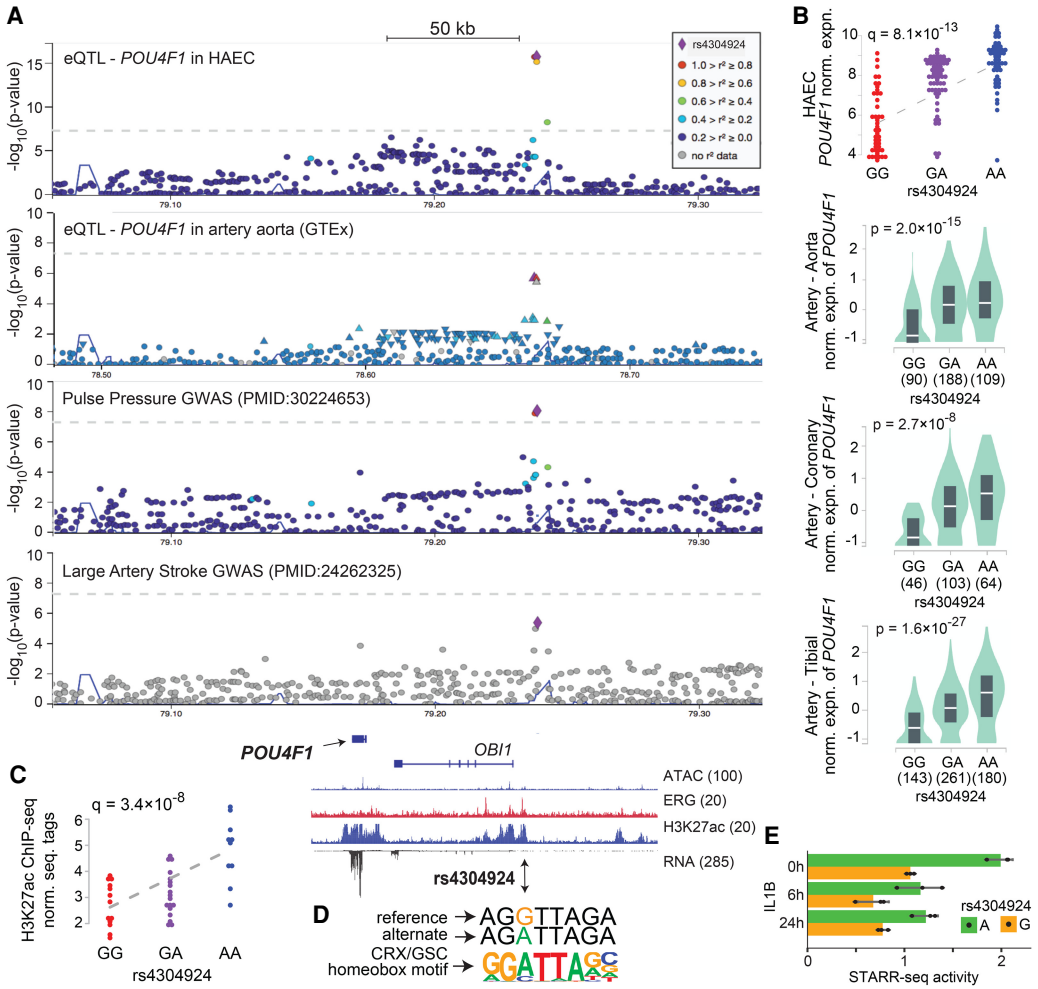


Figure 6. Large artery stroke- and pulse pressure–associated locus at gene *POU4F1*. (*A*) Locus zoom plots depicting the SNPs present in the region surrounding *POU4F1* locus and the –log(*P*-values) in the following data sets: eQTL within HAECs (Stolze et al. 2020); eQTLs in artery aorta by GTEx; GWAS statistics for pulse pressure (Evangelou et al. 2018); and GWAS statistics for large artery stroke (Dichgans et al. 2014). Additional information about genomic region is at the *bottom*, containing the putative enhancer where rs4304924 resides and the surrounding genes. Tracks display H3K27ac, ATAC-seq, ERG binding, RELA binding, and RNA-seq expression from HAECs. (*B*) *POU4F1* gene expression in HAECs by microarray (Probe set ID: 206940_s_at) and three arterial tissues from GTEx compared between genotypes at rs4304924. (*C*) RPKM-normalized H3K27ac tags in HAECs surrounding the rs4304924 SNP by genotype. (*D*) PWM for the CRX/GSC motif and the motif sequences created by the two alleles at rs4304924. (*E*) STARR-seq allele-specific expression (RNA/DNA ratio) for rs4304924 across the different treatment points.

accessible chromatin and H3K27ac; and (3) association to an epigenetic measurement in *cis*- (molQTL), with effects on chromatin accessibility and TF binding being most significant. Using IL1B as a modulator of cell state and environmental simulant, we observe robust evidence for context-specific regulatory SNP effects, thereby underscoring the prevalence of GxE effects on enhancer function.

We find that AP-1 and ETS motif mutations are the most enriched at functional regulatory loci detected in our STARR-seq experiment (Fig. 3). Previous findings indicate that perturbed ERG and JUN bindings are likely affected by these motif mutations, although other members of the ETS and AP-1 families could also be affected (Hogan et al. 2017). The enrichment of functional mutations in ETS and AP-1 motifs is consistent with other reports demonstrating that functional mutations in regulatory elements often reveal motifs of TFs that are selectively active in the respective cell types (van Arensbergen et al. 2019). Additionally, both AP-1 and ETS factors are recognized as pioneering factors, able to establish accessible chromatin, that serve as binding sites for collaborating TFs. We and others have previously shown that genetic variants that perturb the binding of AP-1 and ETS factors consequently affect binding of signal-dependent TFs (Heinz et al. 2013; Hogan et al. 2017; Alasoo et al. 2018). Irrespective of cell type, AP-1 motifs have been shown to produce high enhancer activities (Nguyen et al. 2016), whereas ETS factors seem to show high episomal activity (Klein et al. 2020), suggesting that these motifs might be more evident in validated elements by MPRAs. Still, based on similar enrichment values of STARR-seq-validated SNPs, we show that SVM-based ERG motif mutations have only slightly lower enrichment among validated SNPs than the combination of being within an ERG peak and an ERG bQTL, supporting the high value of motif mutation prediction in identifying functional SNPs. Importantly, many validated SNPs did not fall into any motif mutations. This could be due to poor prediction power of the PWM for weak TF binding and the lack of SVM-based binding models for all TFs and highlights the need for improved algorithms to identify functional variants in the future (Yan et al. 2021).

Many SNPs that fit the epigenetic and/or mutation criteria outlined above did not validate in STARR-seq. This could be explained by the possibility that (1) causal alleles of weak effect may fall below the limit of detection in STARR-seq, or (2) their detection requires longer sequence context that is not captured in the 198-bp oligo, (3) SNP effects might require the endogenous genomic position, chromatinization not captured by episomal assays, and (4) perturbations of silencing effects by alleles are unlikely to be detected due to the low basal activity of the minimal promoter. These are limitations that should be addressed in future experimental iterations. Further insight into such discrepancies between genomic features and MPRA results will be necessary to accelerate discovery into the functional noncoding genome.

A major motivation for this study was to fine-map functional noncoding variants at disease loci for complex human vascular diseases, as summarized in Table 2. These include the *SMAD3* locus where rs17293632 introduces an AP-1 motif mutation and represents a candidate causal variant in smooth muscle cells (Miller et al. 2016; Turner et al. 2016; Zhao et al. 2019; Örd et al. 2021). Despite the absence of *cis*-eQTL for *SMAD3* in HAECs, we demonstrate that deletion of the rs17293632-carrying enhancer does abrogate *SMAD3* expression in ECs (Supplemental Fig. S9F), suggesting potential context dependence of this enhancer variant on gene expression. The current study is the first study to implicate this SNP in ECs. Still, further investigation will be required to dissect mechanisms of action at this interesting locus.

One of the most interesting findings from our study was the in-depth characterization of a locus associated with large artery stroke and pulse pressure. We present strong evidence for the causal role of rs4304924 in an enhancer regulating expression of *POU4F1* (Fig. 6). The G allele of this SNP mutates a homeobox TF motif, exhibits diminished H3K27ac in the adjacent chromatin, and has reduced *POU4F1* expression compared to the A allele. This eQTL is replicated in all three arterial GTEx tissues. To our knowledge, there are no publications linking this transcript with function in the arterial vasculature. The other most interesting finding is at a locus associated with AAA where we find two SNPs, rs13385499 and rs13382862, whose alternate alleles correspond to diminished RELA binding and expression of the nearby gene *LDAH* (Supplemental Fig. S11). The alternate alleles mutate TF motifs, and rs13382862 is likewise an eQTL for *LDAH* in GTEx's artery aorta data set. A limitation of this finding, however, is the lack of GWAS summary statistics to evaluate shared signal strength between eQTL and GWAS results. Still, our findings support a role for these SNPs in enhancer function in HAECs that warrants deeper investigations. *LDAH* itself has been mostly described in macrophages where up-regulation of *LDAH* is linked with a reduction in intracellular cholesterol (Goo et al. 2014; Robichaud et al. 2021). *LDAH* has also been shown to be highly expressed in macrophage-laden atherosclerotic plaques (Goo et al. 2014).

Taken together, the implication of our findings is that a comprehensive understanding of noncoding functional SNPs will require experimental observations from comprehensive sets of cell types and cell states. Such context specificity exemplifies the vast complexity of gene regulatory networks.

Methods

Quantitative trait locus analysis

All QTL analysis was done in a prior study (Stolze et al. 2020). Briefly, eQTL analysis was performed using the program matrixeqtl (Shabalin 2012) for SNPs within 1 Mb of the test gene on three separate data sets: RNA-seq collected for 53 HAEC donors in untreated conditions; RNA-seq collected for 53 HAEC donors under 4-h IL1B treatment; and microarray expression collected for 157 HAEC donors in untreated conditions. eQTL analysis was followed by a *P*-value calculation on the gene level (SNP *P*-values were corrected for all the SNPs tested for a specific gene's expression). eQTLs were called using an adjusted *P*-value threshold of 0.05. Molecular QTL analysis was performed using the program RASQUAL (Kumasaka et al. 2016) on seven separate data sets: ChIP-seq for transcription factor ERG binding in 22 HAEC donors in untreated conditions; ChIP-seq for ERG binding in 20 HAEC donors under 4-h IL1B treatment; ChIP-seq for H3K27ac in 42 HAEC donors in untreated conditions; ChIP-seq for H3K27ac in 42 HAEC donors under 4-h IL1B-treated conditions; ChIP-seq for transcription factor RELA (NF-kB) binding in 36 HAEC donors under 4-h IL1B treatment; ATAC-seq for 45 HAEC donors in untreated conditions; and ATAC-seq for 44 HAEC donors under 4-h IL1B treatment. The SNPs tested against these traits were restricted to within the peaks (Heinz et al. 2010). MolQTLs were called using a 0.05 FDR threshold. The QTL data used in this study is available from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE30169 and GSE139377, the NCBI database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) accession phs002057.v1.p1, and summary statistics are available on figshare (https://doi.org/10.6084/m9.figshare.17303804.v1).

Generating haplotypes for STARR-seq library

The selection of genomic loci and SNPs for the STARR-seq library involved a multifaceted approach. As shown in Supplemental Figure S1, three major strategies were used: (1) QTL overlap (QTL_set: 36,024 oligos; 19,581 SNPs); (2) GWAS overlap (GWAS_and_accessibilty, 16,507 oligos; SNPs = 10,640); and (3) custom selection (custom_other: 7245 oligos; SNPs = 4104 [remaining 200 oligos and 19 SNPs were contained in control regions]) (Supplemental Fig. S1A). For the QTL overlap set, we utilized results from our previous study that identified molQTLs in genetically diverse HAEC cultures (Shabalin 2012; Kumasaka et al. 2016; Stolze et al. 2020). In the current study, we selected STARR-seq regions by overlapping all of the molQTL-associated SNPs to select SNPs that had one or more significant associations across the multiple traits (e.g., accessibility; ERG binding, etc.). As shown in Supplemental Figure S1B, this resulted in library regions that contained SNPs spanning most combinatorial categories of QTLs. In conjunction, we utilized results from motif mutation analysis by SNPs (described in the Motif Mutation Analysis section below). Additional prioritization for QTL SNPs was given for those that were genome-wide-significant in the van der Harst CAD GWAS (van der Harst and Verweij 2018).

For the GWAS overlap, SNPs associated with CAD, coronary heart disease, myocardial infarction, and type 2 diabetes were extracted from the GWAS database (Buniello et al. 2019) in May 2018. As majority of the GWAS lead SNPs came from studies that were based on subjects of European ancestry; the co-inherited, proximal SNPs (dbSNP version 146) in tight LD ($r2 > 0.80$) with the GWAS lead SNPs were determined using the 1000 Genomes (The 1000 Genomes Project Consortium 2015) (phase 3, version 5a) European samples using PLINK (Purcell et al. 2007) version 1.90b5.3 with the following settings: '--extract < dbSNP v146 rsIDs>, --keep < EUR sample IDs>, --maf 0.01, --r2, --ld-snp-list < GWAS lead SNP rsIDs>, --ld-window 100000, --ld-window-kb 1000, --ld-window-r2 0.8'.21. The set of candidate SNPs that overlapped the peaks extracted from the following data sources were selected for the STARR-seq library: ATAC-seq/DNaseHS and TF

peaks from HCASMCs (GEO; GSM1876021-28); HAECs (GEO; GSM2394391-8); macrophages (GEO; GSM2592781-4); AoSMCs (GEO; GSM816638); monocytes (GEO; GSM1008582); HUVECs (GEO; GSM816646); HepG2 (GEO; GSM816662, GSE32465, GSE98983); SGBS (GEO; GSE64233, GSE41629) (The ENCODE Project Consortium 2012); liver (GEO; GSM2400294); adipose tissue (ENCSR350CYJ), and strong common DNase hypersensitive peaks assayed in a large collection of cell types by the ENCODE project that can be found under "DNase I Hypersensitivity in 95 cell types" (hotspots) and "wgEncodeRegDnaseClusteredV3.bed.gz" (score ≥1000). The overlap analysis was performed using the HOMER v4 (Heinz et al. 2010) command "mergePeaks" -cobound option. For the custom set, we selected HUVEC enhancers from topologically associated domains (TADs) enriched for hypoxia-responsive genes and TADs that harbor VEGF family members (GEO; GSE94872, GSE52642). In addition, top enhancers responsive to hypoxia, oxPAPC, VEGF, and IL1B in HUVECs were selected based on published studies (GEO; GSE136813, GSE94872, GSE52642). As controls, 65 IFN-responsive enhancers (Muerdter et al. 2018), 100 scrambled regions, and 100 random negative regions that did not overlap any active chromatin marks in the ENCODE database were selected.

The following computational pipeline was used to generate 198-bp sequences representing up to five haplotypes at each locus of interest for inclusion in the STARR-seq library. First, genomic loci of interest were identified using multiple criteria including overlap of HAEC epigenetic features and sequence analysis and formatted as HOMER peak files (Heinz et al. 2010). Second, phased alleles along haplotypes, in VCF format, were utilized from our previous study of 53 genotyped, imputed, and phased individuals that generated the HAEC epigenetic data (Stolze et al. 2020). Note that alleles were only included if they had minor allele frequencies greater than 5% in this population. Third, the HOMER subprogram *annotatePeaks.pl* was used by inputting the peak file from step 1 along with the options "-vcf phased.vcf.file -size given" which outputs another HOMER formatted peak file with columns noting the base pair positions within each peak and alleles of each haplotype. Fourth, the sequence of the reference hg19 human genome was retrieved within each peak's boundaries using the R (R Core Team 2020) package *seqinr()* (Charif and Lobry 2007). We do not expect that remapping all data to the GRCh38 would significantly affect the conclusions because the sequences selected and tested into the STARR-seq assay would have remained the same. A custom R script was then used to iterate through each peak to paste custom sequences together for each haplotype (Supplemental Code S1). Specifically, strings of nonpolymorphic sequence were separated from polymorphic alleles using coordinates in the previous peak file, and then these were pasted together again for each haplotype. Because 63% of oligos had two or more SNPs, represented as common haplotypes, all combinatorial alleles were not inserted; instead, only up to the five most common haplotypes were inserted. For example, a region with three SNPs that would typically be in LD is very unlikely to be represented by eight oligos for two reasons. First, we only included up to five common haplotypes per oligo. Secondly, less than eight haplotypes are likely to exist. This is how the SNP set N is represented by fewer than 2N oligos.

Massively parallel reporter assay

The hSTARR-seq_ORI plasmid (Addgene 99296) (Muerdter et al. 2018) was used as a backbone for the plasmid constructs. DNA inserts, 230 bp long and containing 198 bp of enhancer variant sequence, were synthesized by Agilent. The oligos were designed to have a 2-bp barcode in the 5′ end of the enhancer sequence and 15 bp matching to Illumina NGS sequencing primers in both ends. The first round of emulsion PCR using a Micellula DNA Emulsion & Purification kit (Roboklon) was performed to complete the sequencing primers and to double-strand the oligos. The second round was used to amplify the material. The plasmid was linearized using AgeI and SalI restriction enzymes, and inserts were cloned into the linearized plasmid in 17 reactions using the standard InFusion cloning (Clontech) protocol. The cloned DNA library was transformed to XL-10 gold ultracompetent bacteria (Agilent) in 15 reactions. Plasmid was purified using an EndoFree Maxiprep kit (Qiagen).

Immortalized human aortic endothelial cells (teloHAECs) were purchased from ATCC and cultured in Vascular Cell Basal Medium (ATCC PCS-100-030), supplemented with Vascular Endothelial Cell Growth kit-VEGF (ATCC PCS-100-041), 100 U/mL penicillin, and 100 µg/mL streptomycin. Cells were incubated at 37°C in 5% CO_2 and passaged every 3 d until the number of cells needed for the experiment was achieved.

The plasmid library was transfected in triplicate to 90 million cells/replicate using the Neon transfection system (Invitrogen). The cells were detached with trypsin, centrifuged, washed with PBS, and resuspended in resuspension buffer R at 5×10^7 cells/mL. The cell suspension was mixed with 15 µg of the STARR-seq library/reaction and subjected to electroporation using the 100-µL tip and three 1350V pulses of 10 msec width. The cells were allowed to recover from electroporation for 24 h, after which they were treated with inflammatory stimulus (IL1B, 10 ng/mL). The cells were divided into three treatment groups, each with 30 million cells: 6-h stimulus, 24-h stimulus, and nonstimulated control. Cells were harvested 48 h posttransfection and RNA was extracted using an RNeasy midi kit (Qiagen). Messenger RNA was purified from bulk RNA using Dynabeads Oligo(dT)25 beads (Invitrogen) with a 2:1 beads to total RNA volume ratio. The purified mRNA was treated with Turbo DNase I (Ambion) and purified using an RNeasy MinElute clean up kit (Qiagen). Reverse transcription was performed using UMI-primers. Unique molecular identifiers (UMIs) were added during cDNA synthesis to tag identifiable replicates of the constructs, which improves the data analysis by accounting for PCR duplicates (Kalita et al. 2018). The samples were pooled and RNase A–treated, and the cDNA was purified with AMPure XP beads using a 1.8:1 beads to cDNA ratio. The libraries were amplified using junction PCR. The junction PCR for the RNA library was performed with junction_RNA_fwd and junction_RNA_rev primers (Muerdter et al. 2018), which allow amplification of correctly inserted enhancer sequence cDNA. The jPCR products were purified using AMPure XPbeads with a beads to sample ratio of 0.8. A second PCR step was run to add the index primers for sequencing (NEBNext Multiplex Oligos for Illumina Dual Index Primers Set 1 and 2). PCR products were purified using SPRIselect beads (Beckman) (bead to sample ratio 0.8). High-throughput sequencing was performed on an Illumina NextSeq 500 platform as paired-end 75-cycle dual index runs using the following parameters: Read1: 37 bp (insert sequence), Index1: 10 bp (contains the UMI instead of an i7 index), Index2: 8 bp (i5 index for demultiplexing), Read2: 37 bp (insert sequence). Raw data from STARR-seq experiments is available at NCBI GEO (accession GSE180846).

Sequencing read demultiplexing was carried out using the i5 index (Index2) only, and the Index1 read (containing the UMI) was extracted as regular sequence read. UMI-tools version 1.0.1 (Smith et al. 2017) was used to remove any reads where the UMI did not match the expected sequence pattern RDHBVDHBVD (Kalita et al. 2018). The remaining reads with valid UMIs were aligned to a custom reference genome consisting of all the oligonucleotide sequences included in the library. Alignment was

performed with the STAR aligner (Dobin et al. 2013) by running the nf-core RNA-seq pipeline version 1.4.2 (Ewels et al. 2020). UMI deduplication was performed on the resulting BAM files using the UMI-tools default method. Reads mapping uniquely and strand-specifically to a library oligonucleotide were summed using featureCounts (Subread Python package version 1.6.5) (Liao et al. 2014). To identify enhancers displaying allele-specific expression, STARR-seq read counts for each genetic variant were summed by variant allele (reference or alternative), and the mpralm R package version 1.4.1 (Myint et al. 2019) with the "mean" aggregation method was used. Only regions having at least one count in every replicate for both DNA and RNA samples were considered. The three biological replicates were inputted into mpralm as independent samples. To account for multiple testing, *P*-values were adjusted using the FDR (Benjamini–Hochberg) method and $p\text{-adj} < 0.05$ was considered significant.

De novo motif analysis

De novo motif analysis was performed using HOMER (Heinz et al. 2010) to identify short nucleotide sequences that are overrepresented statistically within a subset of STARR-seq oligos. The command "findMotifs.pl" was used with FASTA files containing STARR-seq oligonucleotide sequences as input and default settings for motif finding parameters. For selecting the top 10% most active STARR-seq regions, only the most active haplotype of each STARR-seq region was considered. Regions in the upper 10th percentile of enhancer activity in teloHAECs (untreated) were compared to all other regions of the STARR-seq library.

PWM-based and SVM-based TF motif mutation analysis

To test for evidence of functional SNPs by identifying alleles distinguishing DNA motifs, we used the MMARGE software package (Link et al. 2018). The human reference hg19 build (International Human Genome Sequencing Consortium 2001) was input with genotyped and imputed SNPs of our HAEC population (dbGaP: phs002057.v1.p1) to create reference and alternate builds using MMARGE's "prepare_files" function. A MMARGE-formatted peak file was created using reference genomic chromosome and position boundaries for all oligos input into the STARR-seq library. These were input into MMARGE's mutation_analysis function, using the "-keep" flag to save temporary output files, along with all TF motifs available in the HOMER (Heinz et al. 2010) database. The temporary files returned were queried to create a unique list of genomic positions, motifs detected, and motif scores for the reference and alternative genome sequence builds. The positions for the alternate genome were shifted to reference coordinates to remove positional differences resulting from indels between reference and alternate builds. This was achieved with tabix (Li 2011) by adding back the difference in allele base pair lengths between reference and alternate alleles to the alternate positions. Motif scores were compared for corresponding motif positions and the difference is reported as the delta PWM in this study. When a motif was only detected in either the reference or alternate, the minimum PWM detection score (from HOMER) was used for the absent genome. As an alternative approach for the prediction of TF binding disruption due to genetic variants, recently published deltaSVM models (Yan et al. 2021) based on in vitro protein-DNA binding data were run for all single-nucleotide variants in the current study (because the models were trained by the original authors using single-nucleotide variants). Code and models to run deltaSVM were obtained from GitHub (https://github.com/ren-lab/deltaSVM [accessed 03/12/2021]). All 94 high-confidence TF binding models published by the original authors were included and run with the authors' recommended thresholds for sequence binding and allelic disruption.

Enrichment testing

To calculate enrichment of any test data set in the STARR-seq significant data sets, we restricted summary statistic lists of both data sets to what was mutually tested in both (e.g., only include variants that have a *P*-value for both STARR-seq allele specificity and the test data set). We counted how many SNPs were significant in the test data set that were included in the STARR-seq library (white), how many SNPs were not significant in the test data set that were included in the STARR-seq library (black), how many SNPs were significant in the test data set and significant in the STARR-seq allele-specific testing (white_drawn), and the number of SNPs that were significant in the allele-specific testing of STARR-seq (n_pick). These values were used in a hypergeometric test using the following command: (1-phyper(white_drawn-1, white, black, n_pick)). This command provides the probability of picking the number of white drawn or higher.

To calculate the enrichment score, the same values from above were put into a simple formula as follows: (white_drawn/((white/(white + black)) × n_pick)) (Supplemental Code S2).

Genomic characteristics enrichment testing

To test for enrichment of bins for conservation, the 100 vertebrate conservation BED file for hg19 was downloaded from the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/gbdb/hg19/multiz100way/phastCons100way.wib). The conservation assigned to one SNP was determined by averaging the conservation score of all base pairs in the region including the SNP that was input into STARR-seq. These were subsequently ranked by conservation, then binned so that each bin had the same number of SNPs. Presence or nonpresence in each bin was what was considered a success or a failure, respectively.

For the GC content enrichment, the command annotatePeaks.pl from HOMER was used. A peak file of all of the regions input into the STARR-seq library (all haplotypes) was put into annotatePeaks.pl with the option –CpG which provides two columns, one of which is a GC content for each peak. GC content was assigned to a SNP by averaging the GC content across all of the haplotypes where the SNP (in some form) is included. These were then ranked and binned so that each bin had the same number of SNPs. Enrichment was done by presence or absence of a SNP in the bin corresponding to success or failure, respectively.

ENCODE and EpiMap comparisons

To assess the native chromatin state of the regions included in the STARR-seq library, previously published epigenetic region sets were downloaded from the EpiMap Repository website (http://compbio.mit.edu/epimap; accessed 09/20/2021) (Boix et al. 2021). Briefly, the EpiMap project aggregated and uniformly processed human epigenetic data sets across multiple data generation projects, including ENCODE and Roadmap Epigenomics, spanning a total of 833 epigenomes from 33 tissue groups (categories) (Boix et al. 2021). For high-detail chromatin state annotations, each STARR-seq library region was intersected with the 18-state chromatin partitioning ChromHMM results for the deeply profiled ENCODE 2012 subset of EpiMap. For each cell type, the most highly annotated epigenome sample (having the least quiescent chromatin) was retained. An overlap of ≥100 bp was required between a STARR-seq oligo (198 bp) and a ChromHMM chromatin state region; thus, each STARR-seq oligo was assigned to only one ChromHMM state in each epigenome. For allelic variants, the

minimum overlap was 1 bp. To classify STARR-seq regions broadly into enhancers and promoters across all EpiMap samples, all enhancer states (active, bivalent, genic, and weak) and all promoter states (active TSS, bivalent TSS, flanking TSS, flanking TSS upstream, and flanking TSS downstream) from the 18-state ChromHMM tracks were included (Boix et al. 2021). Similarly to Boix et al., regions with ≥75% of annotations of one type (enhancer or promoter) were classified as such, whereas the remaining regions (neither specifically enhancer nor promoter) were classified as a separate category ("sample-dependent"). For analyses intersecting STARR-seq library regions with active regulatory elements, we used the EpiMap "active enhancer" or "active promoter" region sets that were generated by the original authors by intersecting DNase hypersensitivity regions with H3K27ac signal regions in each sample (Boix et al. 2021). As these regions are size-constrained, a ≥50-bp overlap was required for STARR-seq regions. For calculations of overlap fractions (such as allelic effect-significant enhancer variants relative to total enhancer-overlapping variants in the same epigenome), epigenomes with ≤50,000 annotated enhancers were excluded. This removed 13 out of 390 ENCODE epigenomes (none endothelial; two adrenal, one pancreas, one cardiac, two neuronal, one thyroid, one colon, two stem cell, two cancer cell line, one immortalized cell line).

To compare enrichment of allelic effect variants between HAEC open chromatin regions versus regions that are not open in HAECs but are open in other cell types, the ENCODE track was downloaded in BED file format (encode.bed) from the UCSC Genome Browser (Kent et al. 2002). This file was processed using HOMER (Heinz et al. 2010) to restrict peak sizes to 200 bp using command (annotatePeaks.pl encode.bed hg19 -size 200). This peak file (encode.peaks) was then separated into those that are shared with HAECs based on our ATAC-seq data across 44 individuals (atac.peaks), also using HOMER as follows (mergePeaks -cobound 1 encode.peaks atac.peaks).

Kolmogorov–Smirnov testing

As a validation of the enrichment testing, we verified the associations by using Kolmogorov–Smirnov testing between the QTL analysis FDRs of the STARR-seq-validated sets of SNPs vs the SNPs that did not validate. The FDRs for one data set (e.g., ERG binding QTL FDRs) were pulled for all of the SNPs that validated (valid) and all SNPs that did not validate (noValid). The R code to perform testing is as follows using command from package "stats": ks.test(valid, noValid, alternative = "less"). Plotting of the densities for graphical representation of the distributions tested was done by using the "density" command from package "stats". Cumulative distributions were created with command "Freq" from package "DescTools".

CRISPR validation experiments

CRISPR-Cas9-mediated deletion of target regions was performed using the Alt-R CRISPR-Cas9 System (Integrated DNA Technologies). Briefly, CRISPR-Cas9 single gRNAs (Supplemental Table S3) flanking the target enhancers were designed using an online CRISPR design tool (https://eu.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE) and ordered from IDT as crRNAs. These crRNAs were annealed to a tracrRNA (IDT 1072532) and complexed with Cas9 endonuclease (S.p. HiFi Cas9 Nuclease V3; IDT 1081060) to form the ribonucleoprotein complex (RNP). The RNP complexes were then delivered into 150,000 teloHAEC (ATCC) cells per replicate by electroporation using the Neon transfection system (Invitrogen) with a 1350 V pulse of 30 msec width, following the IDT CRISPR genome editing protocol for RNP electroporation, Neon transfection system. Two days later, transfected cells were collected for RNA and gDNA. In order to analyze the efficiency of enhancer deletion, genomic DNA was extracted using a NucleoSpin tissue kit (Macherey-Nagel) and amplified by PCR using specific primers (Supplemental Table S2) flanking the deletion sites and DreamTaq DNA Polymerase (Thermo Fisher Scientific EP0701). PCR products were then analyzed by electrophoresis in a 1% agarose gel. To analyze the effects of enhancer deletion on *cis* target gene expression, the RNA was extracted using an RNeasy-Plus Micro kit (Qiagen 74034). RNA library preparation was carried out using a QuantSeq 3′ mRNA-seq Library Prep kit FWD for Illumina (Lexogen) according to the manufacturer's instructions. The resultant library was quantified using a Qubit dsDNA HS Assay kit (Thermo Fisher Scientific Q32854) and its quality was checked with a Bioanalyzer using High Sensitivity DNA kit (Agilent Technologies 5067-4626). Individual libraries were pooled in equimolar ratio (4 nM total) and sequenced with the NextSeq 550 platform (Illumina) in a single-end 75-cycle high-output run. The sequencing reads were processed using the nf-core RNA-seq pipeline (version 3.1) (Ewels et al. 2020). Genes with very low expression were filtered out with the filterByExpr function of the edgeR package (version 3.24.3) (Robinson et al. 2010) using minimum count five and minimum total count 15. The effect of CRISPR deletion on gene expression was studied for all the *cis* candidate genes within 1 Mb of the deleted region. The positive and negative guide RNA transfected cells were used as controls ($n = 4$ each). DESeq2 version 1.22.2 (Love et al. 2014) was used for the statistical analysis comparing the enhancer deleted samples ($n = 4$) to the controls ($n = 8$) and FDR < 5% was considered significant. The data is represented as $\log_2$ fold change and standard error for the estimated coefficients on the $\log_2$ scale.

Data access

All raw data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE180846 and GSE191253.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This research was supported by the National Institutes of Health (NIH) (R01HL147187 to C.E.R. and T32 HL007249-42 to L.K.S.), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant No. 802825 to M.U.K.), the American Heart Association (20PRE35200195 to L.K.S.), the Academy of Finland (Grant Nos. 287478 and 319324 to M.U.K.), the Finnish Foundation for Cardiovascular Research, the Sigrid Juselius Foundation, and the Doctoral Program of Molecular Medicine at University of Eastern Finland.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, HIPSCI Consortium, Hale C, Dougan G, Gaffney DJ. 2018. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat Genet* **50:** 424–431. doi:10.1038/s41588-018-0046-7

Arnold CD, Gerlach D, Stelzer C, Boryń LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339:** 1074–1077. doi:10.1126/science.1232542

Birdsey GM, Shah AV, Dufton N, Reynolds LE, Osuna Almagro L, Yang Y, Aspalter IM, Khan ST, Mason JC, Dejana E, et al. 2015. The endothelial transcription factor ERG promotes vascular stability and growth through Wnt/β-catenin signaling. *Dev Cell* **32:** 82–96. doi:10.1016/j.devcel.2014.11.016

Boix CA, James BT, Park YP, Meuleman W, Kellis M. 2021. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590:** 300–307. doi:10.1038/s41586-020-03145-z

Buniello A, MacArthur JAL, Cerezo M, Harris LW, 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:** D1005–D1012. doi:10.1093/nar/gky1120

Charif D, Lobry JR. 2007. Seqinr 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural approaches to sequence evolution: molecules, networks, populations* (ed. Bastolla U, et al.), pp. 207–232. Springer Verlag, New York.

Dichgans M, Malik R, König IR, Rosand J, Clarke R, Gretarsdottir S, Thorleifsson G, Mitchell BD, Assimes TL, Levi C, et al. 2014. Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke* **45:** 24–36. doi:10.1161/STROKEAHA.113.002707

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, Ntritsos G, Dimou N, Cabrera CP, Karaman I, et al. 2018. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat Genet* **50:** 1412–1425. doi:10.1038/s41588-018-0205-x

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* **38:** 276–278. doi:10.1038/s41587-020-0439-x

Fedtsova NG, Turner EE. 1995. Brn-3.0 expression identifies early post-mitotic CNS neurons and sensory neural precursors. *Mech Dev* **53:** 291–304. doi:10.1016/0925-4773(95)00435-1

Goo YH, Son SH, Kreienberg PB, Paul A. 2014. Novel lipid droplet–associated serine hydrolase regulates macrophage cholesterol mobilization. *Arterioscler Thromb Vasc Biol* **34:** 386–396. doi:10.1161/ATVBAHA.113.302448

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589. doi:10.1016/j.molcel.2010.05.004

Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK. 2013. Effect of natural genetic variation on enhancer selection and function. *Nature* **503:** 487–492. doi:10.1038/nature12615

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* **106:** 9362–9367. doi:10.1073/pnas.0903103106

Hogan NT, Whalen MB, Stolze LK, Hadeli NK, Lam MT, Springstead JR, Glass CK, Romanoski CE. 2017. Transcriptional networks specifying homeostatic and inflammatory programs of gene expression in human aortic endothelial cells. *eLife* **6:** e22536. doi:10.7554/eLife.22536

Huang D, Ovcharenko I. 2015. Identifying causal regulatory SNPs in ChIP-seq enhancers. *Nucleic Acids Res* **43:** 225–236. doi:10.1093/nar/gku1318

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921. doi:10.1038/35057062

Jones GT, Tromp G, Kuivaniemi H, Gretarsdottir S, Baas AF, Giusti B, Strauss E, Van't Hof FN, Webb TR, Erdman R, et al. 2017. Meta-analysis of genome-wide association studies for abdominal aortic aneurysm identifies four new disease-specific risk loci. *Circ Res* **120:** 341–353. doi:10.1161/CIRCRESAHA.116.308765

Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha RK, Benner C, et al. 2013. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* **51:** 310–325. doi:10.1016/j.molcel.2013.07.010

Kaikkonen MU, Niskanen H, Romanoski CE, Kansanen E, Kivelä AM, Laitalainen J, Heinz S, Benner C, Glass CK, Ylä-Herttuala S. 2014. Control of VEGF-A transcriptional programs by pausing and genomic compartmentalization. *Nucleic Acids Res* **42:** 12570–12584. doi:10.1093/nar/gku1036

Kalita CA, Brown CD, Freiman A, Isherwood J, Wen X, Pique-Regi R, Luca F. 2018. High-throughput characterization of genetic effects on DNA–protein binding and gene transcription. *Genome Res* **28:** 1701–1708. doi:10.1101/gr.237354.118

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996–1006. doi:10.1101/gr.229102

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23:** 800–811. doi:10.1101/gr.144899.112

Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, Ahituv N, Shendure J. 2020. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* **17:** 1083–1091. doi:10.1038/s41592-020-0965-y

Krause MD, Huang RT, Wu D, Shentu TP, Harrison DL, Whalen MB, Stolze LK, Di Rienzo A, Moskowitz IP, Civelek M, et al. 2018. Genetic variant at coronary artery disease and ischemic stroke locus 1p32.2 regulates endothelial responses to hemodynamics. *Proc Natl Acad Sci* **115:** E11349–E11358. doi:10.1073/pnas.1810568115

Kumasaka N, Knights AJ, Gaffney DJ. 2016. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet* **48:** 206–213. doi:10.1038/ng.3467

Lathen C, Zhang Y, Chow J, Singh M, Lin G, Nigam V, Ashraf YA, Yuan JX, Robbins IM, Thistlethwaite PA. 2014. ERG-APLNR axis controls pulmonary venule endothelial proliferation in pulmonary veno-occlusive disease. *Circulation* **130:** 1179–1191. doi:10.1161/CIRCULATIONAHA.113.007822

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15:** R29. doi:10.1186/gb-2014-15-2-r29

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27:** 718–719. doi:10.1093/bioinformatics/btq671

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656

Link VM, Romanoski CE, Metzler D, Glass CK. 2018. MMARGE: Motif Mutation Analysis for Regulatory Genomic Elements. *Nucleic Acids Res* **46:** 7006–7021. doi:10.1093/nar/gky491

Liu S, Liu Y, Zhang Q, Wu J, Liang J, Yu S, Wei GH, White KP, Wang X. 2017. Systematic identification of regulatory variants associated with cancer risk. *Genome Biol* **18:** 194. doi:10.1186/s13059-017-1322-z

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8

Machiela MJ, Chanock SJ. 2015. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31:** 3555–3557. doi:10.1093/bioinformatics/btv402

Miller CL, Pjanic M, Wang T, Nguyen T, Cohain A, Lee JD, Perisic L, Hedin U, Kundu RK, Majmudar D, et al. 2016. Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *Nat Commun* **7:** 12092. doi:10.1038/ncomms12092

Muerdter F, Boryń LM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, Pagani M, Haberle V, Kazmar T, Catarino RR, et al. 2018. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15:** 141–149. doi:10.1038/nmeth.4534

Myint L, Avramopoulos DG, Goff LA, Hansen KD. 2019. Linear models enable powerful differential activity analysis in massively parallel reporter assays. *BMC Genomics* **20:** 209. doi:10.1186/s12864-019-5556-x

Nguyen TA, Jones RD, Snavely AR, Pfenning AR, Kirchner R, Hemberg M, Gray JM. 2016. High-throughput functional comparison of promoter and enhancer activities. *Genome Res* **26:** 1023–1033. doi:10.1101/gr.204834.116

Örd T, Õunap K, Stolze LK, Aherrahrou R, Nurminen V, Toropainen A, Selvarajan I, Lönnberg T, Aavik E, Ylä-Herttuala S, et al. 2021. Single-cell epigenomics and functional fine-mapping of atherosclerosis GWAS loci. *Circ Res* **129:** 240–258. doi:10.1161/CIRCRESAHA.121.318971

Ostuni R, Piccolo V, Barozzi I, Polletti S, Termanini A, Bonifacio S, Curina A, Prosperini E, Ghisletti S, Natoli G. 2013. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152:** 157–171. doi:10.1016/j.cell.2012.12.018

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575. doi:10.1086/519795

R Core Team. 2020. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Ridker PM, Everett BM, Thuren T, MacFadyen JG, Chang WH, Ballantyne C, Fonseca F, Nicolau J, Koenig W, Anker SD, et al. 2017. Antiinflammatory therapy with canakinumab for atherosclerotic disease. *N Engl J Med* **377:** 1119–1131. doi:10.1056/NEJMoa1707914

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518:** 317–330. doi:10.1038/nature14248

Robichaud S, Fairman G, Vijithakumar V, Mak E, Cook DP, Pelletier AR, Huard S, Vanderhyden BC, Figeys D, Lavallée-Adam M, et al. 2021. Identification of novel lipid droplet factors that regulate lipophagy and cholesterol efflux in macrophage foam cells. *Autophagy* 3671–3689. doi:10.1080/15548627.2021.1886839

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

Selvarajan I, Toropainen A, Garske KM, López Rodríguez M, Ko A, Miao Z, Kaminska D, Õunap K, Örd T, Ravindran A, et al. 2021. Integrative analysis of liver-specific non-coding regulatory SNPs associated with the risk of coronary artery disease. *Am J Hum Genet* **108:** 411–430. doi:10.1016/j.ajhg.2021.02.006

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28:** 1353–1358. doi:10.1093/bioinformatics/bts163

Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27:** 491–499. doi:10.1101/gr.209601.116

Stolze LK, Conklin AC, Whalen MB, López Rodríguez M, Õunap K, Selvarajan I, Toropainen A, Örd T, Li J, Eshghi A, et al. 2020. Systems genetics in human endothelial cells identifies non-coding variants modifying enhancers, expression, and complex disease traits. *Am J Hum Genet* **106:** 748–763. doi:10.1016/j.ajhg.2020.04.008

Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, et al. 2016. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165:** 1519–1529. doi:10.1016/j.cell.2016.04.027

Turner AW, Martinuk A, Silva A, Lau P, Nikpay M, Eriksson P, Folkersen L, Perisic L, Hedin U, Soubeyrand S, et al. 2016. Functional analysis of a novel genome-wide association study signal in *SMAD3* that confers protection from coronary artery disease. *Arterioscler Thromb Vasc Biol* **36:** 972–983. doi:10.1161/ATVBAHA.116.307294

Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS, et al. 2016. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165:** 1530–1545. doi:10.1016/j.cell.2016.04.048

van Arensbergen J, Pagie L, FitzPatrick VD, de Haas M, Baltissen MP, Comoglio F, van der Weide RH, Teunissen H, Võsa U, Franke L, et al. 2019. High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet* **51:** 1160–1169. doi:10.1038/s41588-019-0455-2

van der Harst P, 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. doi:10.1161/CIRCRESAHA.117.312086

Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG, Lowe WL Jr., Reddy TE. 2015. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res* **25:** 1206–1214. doi:10.1101/gr.190090.115

Yan J, Qiu Y, Ribeiro Dos Santos AM, Yin Y, Li YE, Vinckier N, Nariai N, Benaglio P, Raman A, Li X, et al. 2021. Systematic analysis of binding of transcription factors to noncoding variants. *Nature* **591:** 147–151. doi:10.1038/s41586-021-03211-0

Yuan L, Nikolova-Krstevski V, Zhan Y, Kondo M, Bhasin M, Varghese L, Yano K, Carman CV, Aird WC, Oettgen P. 2009. Antiinflammatory effects of the ETS factor ERG in endothelial cells are mediated through transcriptional repression of the interleukin-8 gene. *Circ Res* **104:** 1049–1057. doi:10.1161/CIRCRESAHA.108.190751

Zhang P, Xia JH, Zhu J, Gao P, Tian YJ, Du M, Guo YC, Suleman S, Zhang Q, Kohli M, et al. 2018. High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nat Commun* **9:** 2022. doi:10.1038/s41467-018-04451-x

Zhao Q, Wirka R, Nguyen T, Nagao M, Cheng P, Miller CL, Kim JB, Pjanic M, Quertermous T. 2019. TCF21 and AP-1 interact through epigenetic modifications to regulate coronary artery disease gene expression. *Genome Med* **11:** 23. doi:10.1186/s13073-019-0635-9

Received July 30, 2021; accepted in revised form January 6, 2022.