The American Journal of Human Genetics, Volume *103*

Supplemental Data

Genetic Regulatory Mechanisms of Smooth Muscle Cells Map to Coronary Artery Disease Risk Loci

Boxiang Liu, Milos Pjanic, Ting Wang, Trieu Nguyen, Michael Gloudemans, Abhiram Rao, Victor G. Castano, Sylvia Nurnberg, Daniel J. Rader, Susannah Elwyn, Erik Ingelsson, Stephen B. Montgomery, Clint L. Miller, and Thomas Quertermous

Supplementary Figures

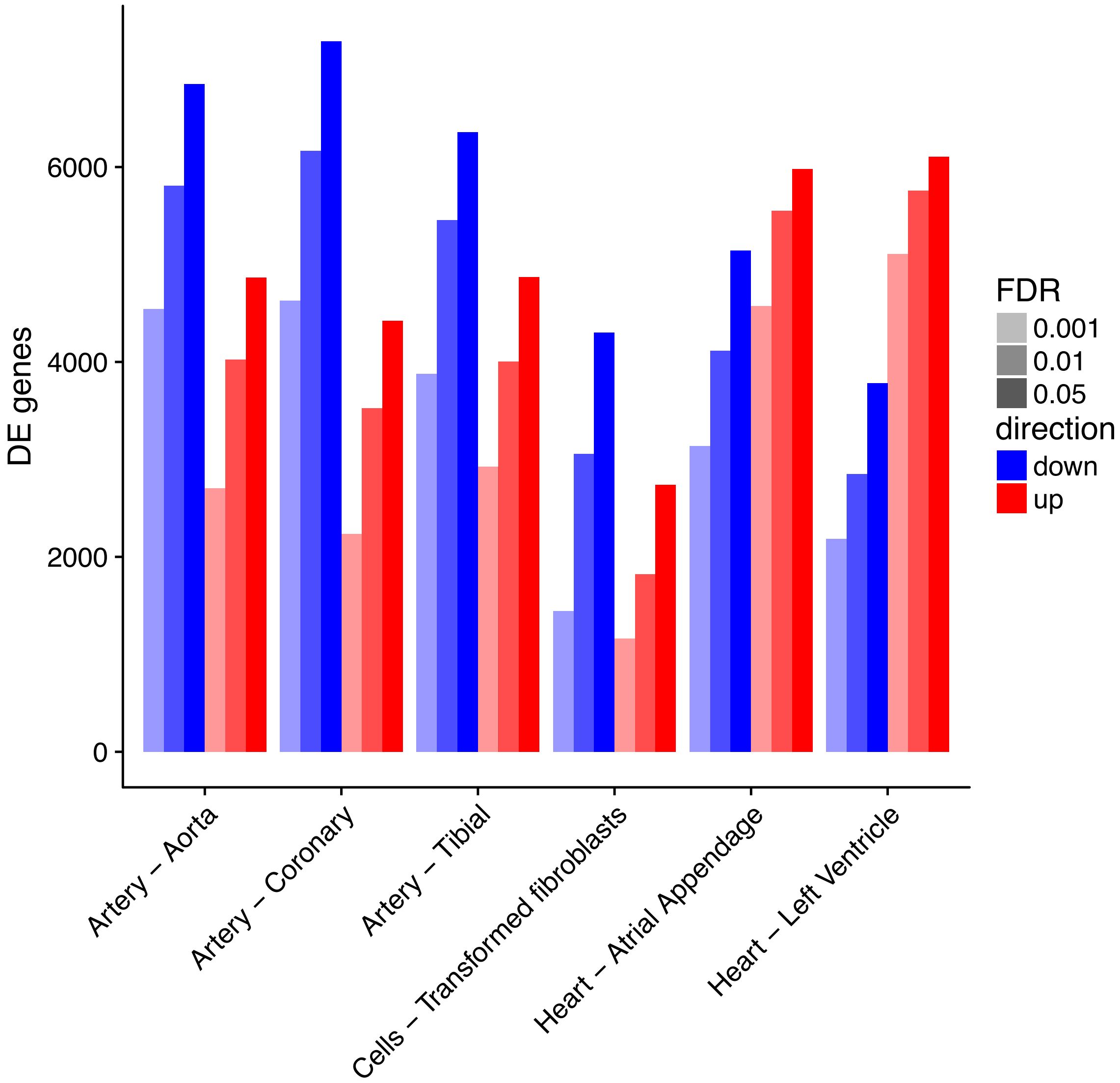
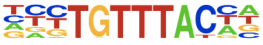
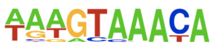
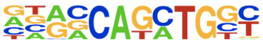
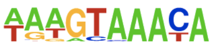
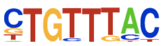


Figure S1. Differential expression between HCASMC and related GTEx tissues. We performed differential expression analysis between HCASMC and six related tissues: three artery tissues (tissue of origin), two heart tissues (tissue of origin), and fibroblast (closest neighbor). The number of differentially expressed genes at FDR < 0.001 are as follows: Artery – Aorta: 2703 (up) and 4542 (down); Artery – Coronary: 2234 (up) and 4630 (down); Artery – Tibial: 2923 (up) and 3877 (down); Fibroblast: 1164 (up) and 1446 (down); Heart – Atrial Appendage: 4572 (up) and 3139 (down); Heart – Left Ventricle: 5107 (up) and 2184 (down).

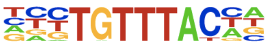
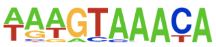
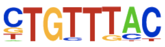
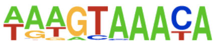
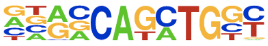
A

50bp

Motif Logo	Name	P-value	% of Target Sequences with Motif	% of Background Sequences with Motif
	FOXP1(Forkhead)/H9-FOXP1-ChIP-Seq(GSE31006)/Homer	1e-31	4.75%	1.53%
	FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-29	10.08%	5.07%
	Atoh1(bHLH)/Cerebellum-Atoh1-ChIP-Seq(GSE22111)/Homer	1e-28	5.23%	1.91%
	FOXA1(Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1e-26	8.61%	4.19%
	Foxo1(Forkhead)/RAW-Foxo1-ChIP-Seq(Fan et al.)/Homer	1e-24	10.66%	5.85%

B

200 bp

Motif Logo	Name	P-value	% of Target Sequences with Motif	% of Background Sequences with Motif
	FOXP1(Forkhead)/H9-FOXP1-ChIP-Seq(GSE31006)/Homer	1e-126	17.49%	5.69%
	FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-105	33.54%	17.76%
	Foxo1(Forkhead)/RAW-Foxo1-ChIP-Seq(Fan et al.)/Homer	1e-102	36.24%	20.09%
	FOXA1(Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1e-97	29.29%	14.96%
	Atoh1(bHLH)/Cerebellum-Atoh1-ChIP-Seq(GSE22111)/Homer	1e-85	17.64%	7.32%

C

1000 bp

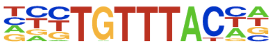
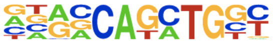
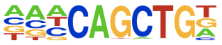
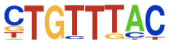
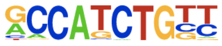
Motif Logo	Name	P-value	% of Target Sequences with Motif	% of Background Sequences with Motif
	FOXP1(Forkhead)/H9-FOXP1-ChIP-Seq(GSE31006)/Homer	1e-124	37.67%	20.62%
	Atoh1(bHLH)/Cerebellum-Atoh1-ChIP-Seq(GSE22111)/Homer	1e-93	39.42%	24.15%
	Ap4(bHLH)/AML-Tfap4-ChIP-Seq(GSE45738)/Homer	1e-88	40.04%	25.07%
	Foxo1(Forkhead)/RAW-Foxo1-ChIP-Seq(Fan et al.)/Homer	1e-77	71.16%	56.25%
	NeuroD1(bHLH)/Islet-NeuroD1-ChIP-Seq(GSE30298)/Homer	1e-76	31.04%	18.40%

Figure S2. Motif enriched within HCASMC-specific peaks. Motif enrichment analysis (see **Methods**) reveal that Forkhead box (FOX) motifs (FOXP1, FOXA1 and FOXO1) are enriched in HCASMC-specific chromatin accessibility regions. FOXA1 and other family members function as pioneer transcription factors involved in cell growth, proliferation and differentiation. The enrichment is robust to selection of window sizes around HCASMC-specific peaks (50bp, 200bp, and 1000bp tested), suggesting that the enrichment is robust to the selection of window size.

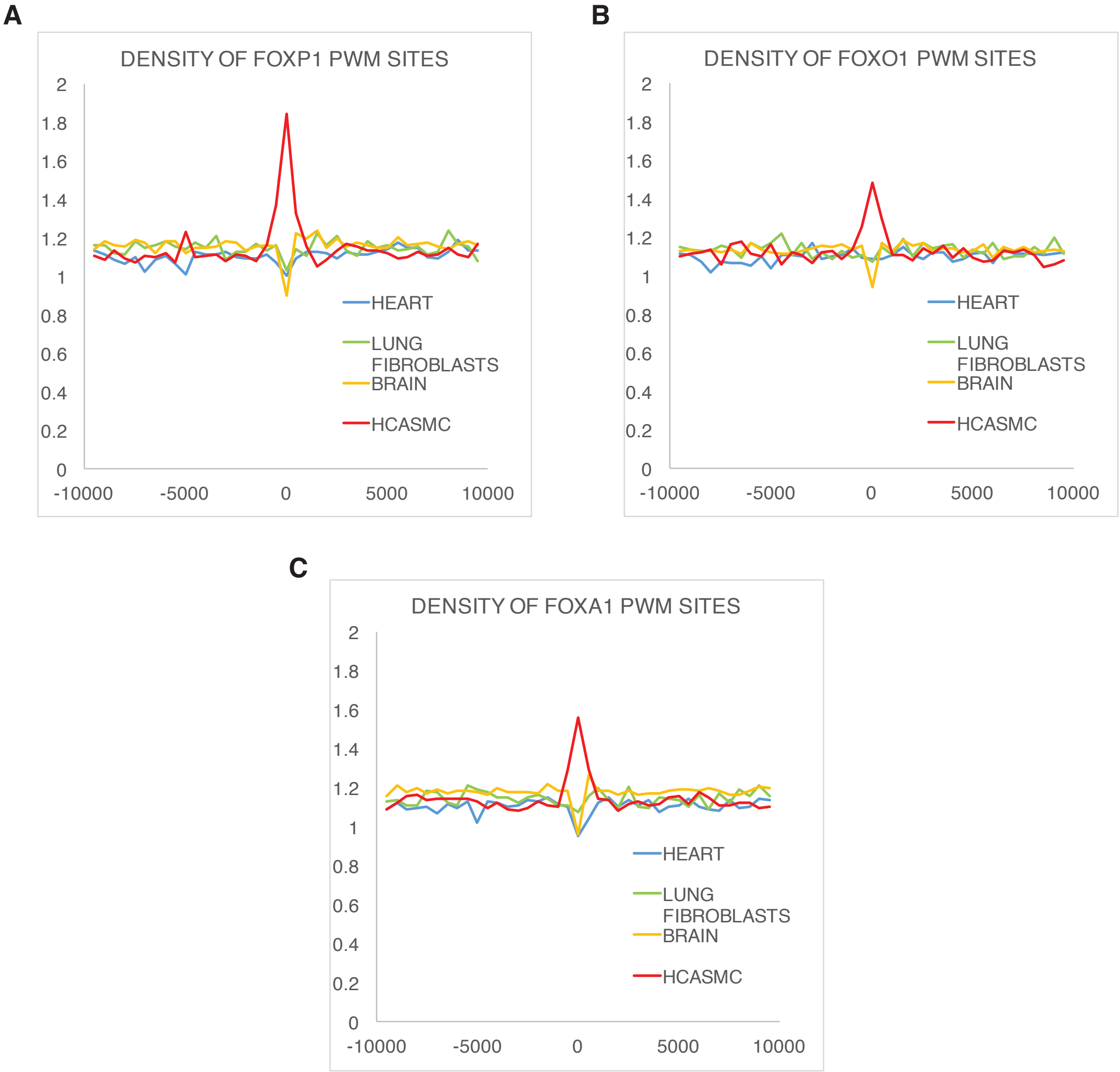


Figure S3. FOX family motif enrichment is specific to HCASMC. To validate that FOX motif enrichment is specific to HCASMC, we tested whether FOX motifs are enriched in lung fibroblast-, brain-, and heart-specific chromatin accessibility regions (see **Methods**). The results suggest that FOX motifs are not enriched, and to some extent depleted, in regions specific to tissues other than HCASMC.

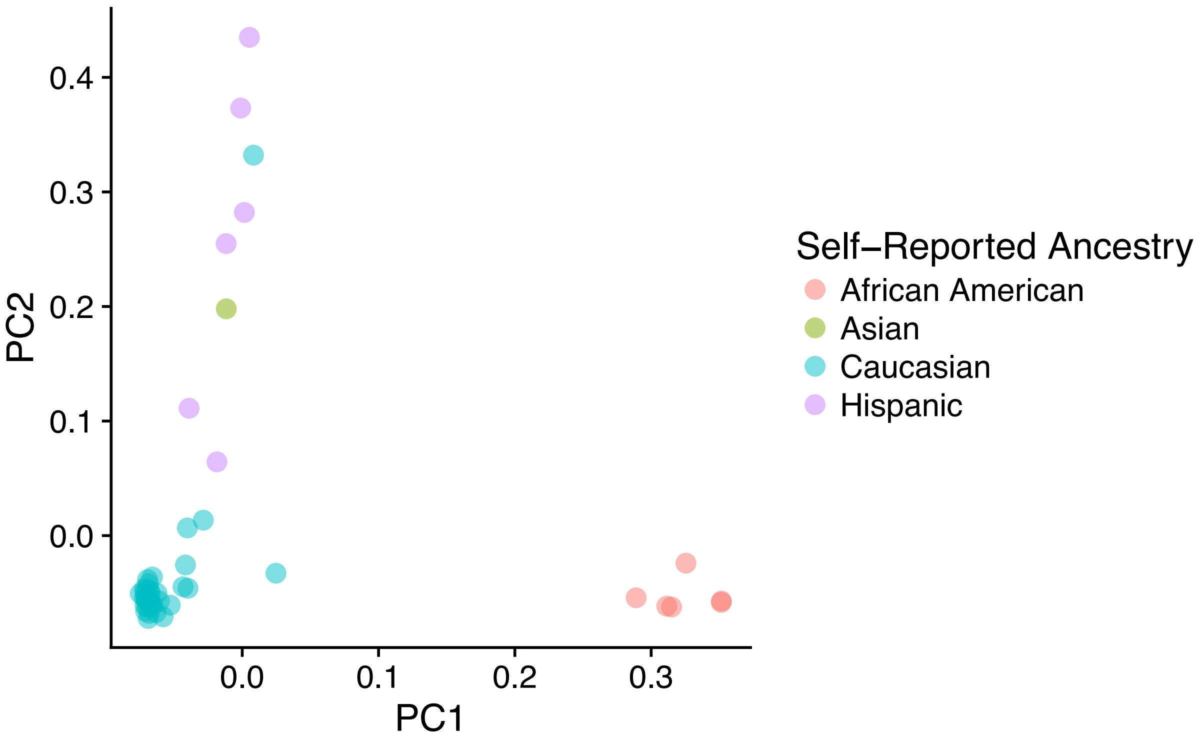


Figure S4. Correspondence between self-reported ancestry and genotype principal components. We estimated ancestry principal components using the R package SNPRelate (see **Methods**), and colored the data points according to self-reported ancestry. A self-reported Caucasian individual is likely Hispanic, and a self-reported Asian individual is likely admixed.

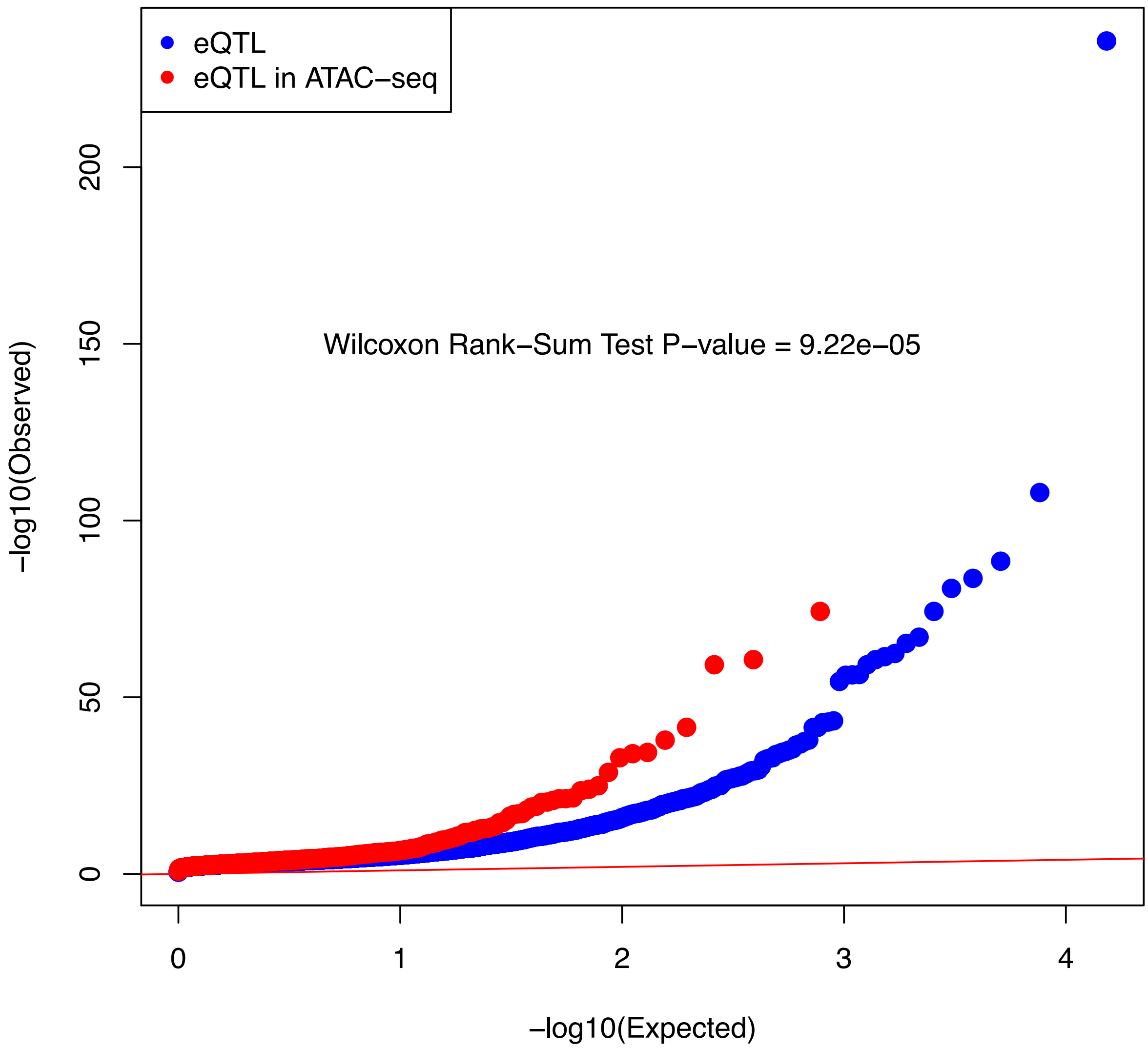


Figure S5. HCASMC eQTLs are enriched in HCASMC ATACseq regions. We tested whether eQTLs are enriched in HCASMC open chromatin regions by plotting all top eQTLs and top eQTLs overlapping ATAC-seq peaks. The latter has a stronger upward trend, suggesting a significant enrichment of eQTLs in open chromatin regions (two-sided Wilcoxon rank-sum test p-value = 9.22×10^{-5}).

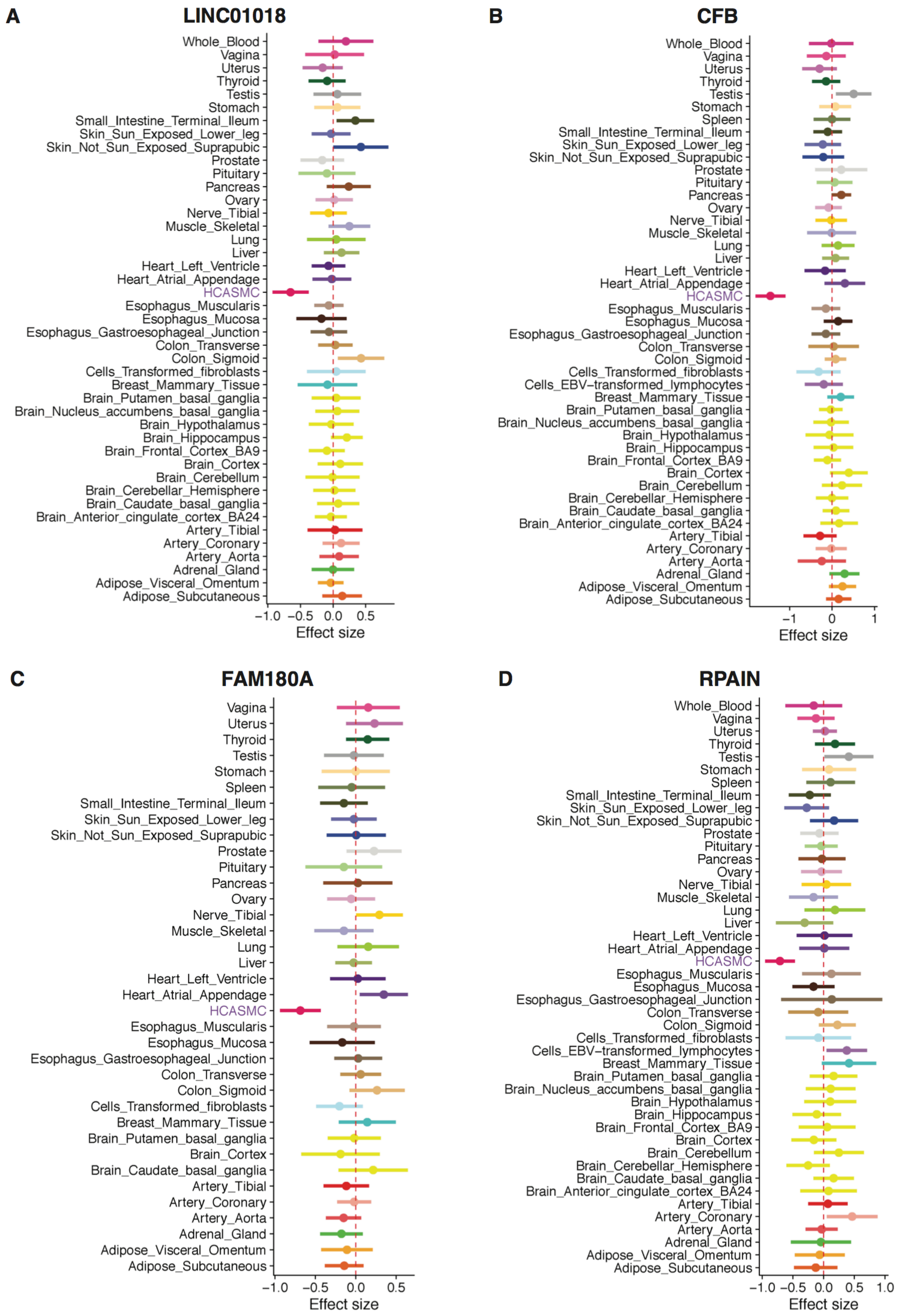


Figure S6 HCASMC-specific eQTLs *RPAIN*, *FAM180A*, *CFB*, and *LINC01018*. We performed HCASMC-specific eQTL calling with METASOFT (see **Supplementary Methods** Section 12). Under the most stringent criteria (m-value > 0.9 for HCASMC and m-value < 0.1 for GTEx tissues), we found 4 HCASMC-specific eQTLs (*RPAIN*, *FAM180A*, *CFB*, and *LINC01018*). This plot compares the effect size of HCASMC-specific eQTLs across all tissues. Error bar indicates 95% confidence intervals.

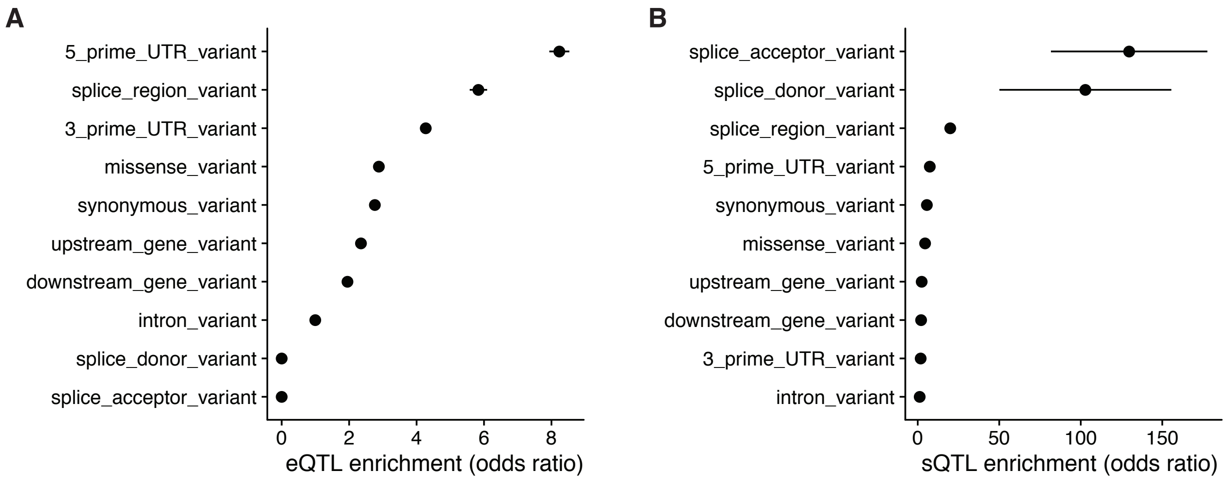


Figure S7. Enrichment of eQTL and sQTL SNPs in various genomic features. We tested enrichment of (A) eQTL and (B) sQTL variants within 11 genomic regions: downstream-gene variant, exonic variant, intronic variant, missense variant, splice-acceptor variant, splice-donor variant, splice-region variant, synonymous variant, upstream-gene variant, 3' UTR variant, and 5' UTR variant. As we expected, eQTL SNPs were most enriched around the promoter region or the 5' UTR regions. Splicing QTL SNPs were most enriched in splice donor and acceptor sites and the splice region. Error bars indicate standard error of the odds ratios.

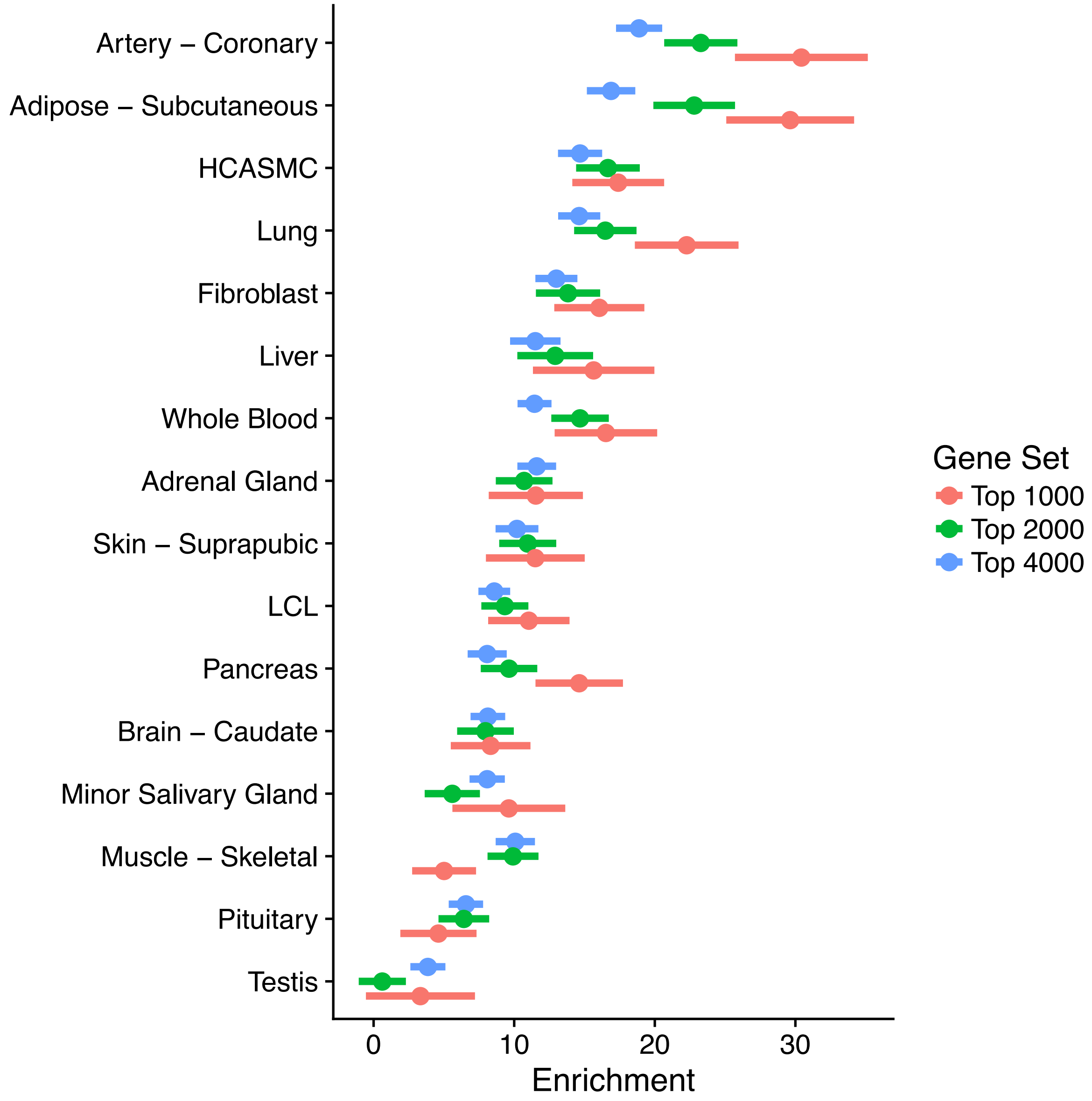


Figure S8. Enrichment of heritability estimate for tissue-specific genes. We ranked tissues according to the heritability enrichment estimated by stratified LD score regression. To determine whether the ranking is robust to the number of tissue-specific genes selected, we defined the top 1000, 2000, and 4000 genes with the highest z-scores as tissue-specific. With top 2000 and 4000 genes, HCASMC ranks number three, after coronary artery and adipose. With top 1000 genes, HCASMC ranks number four, after coronary artery, adipose and lung. These results suggest that enrichment estimates were robust to the selection of tissue-specific genes. Error bars indicate standard error of the enrichments.

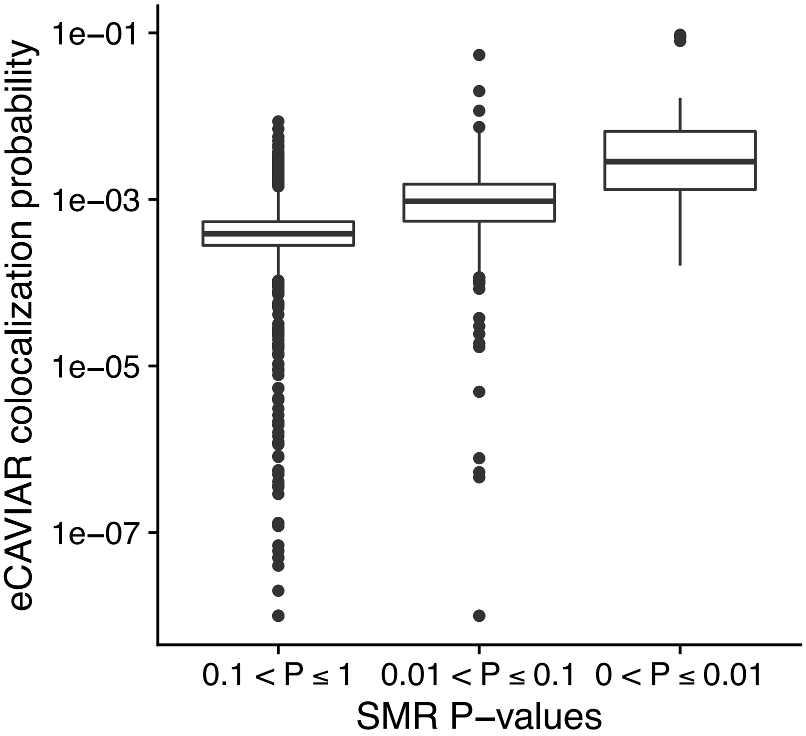


Figure S9 Comparing eCAVIAR and SMR results. We plotted the binned SMR p-values against eCAVIAR colocalization posterior probabilities (CLPPs). We observed that as SMR p-values becomed more significant, eCAVIAR CLPPs becomed more significant as well.

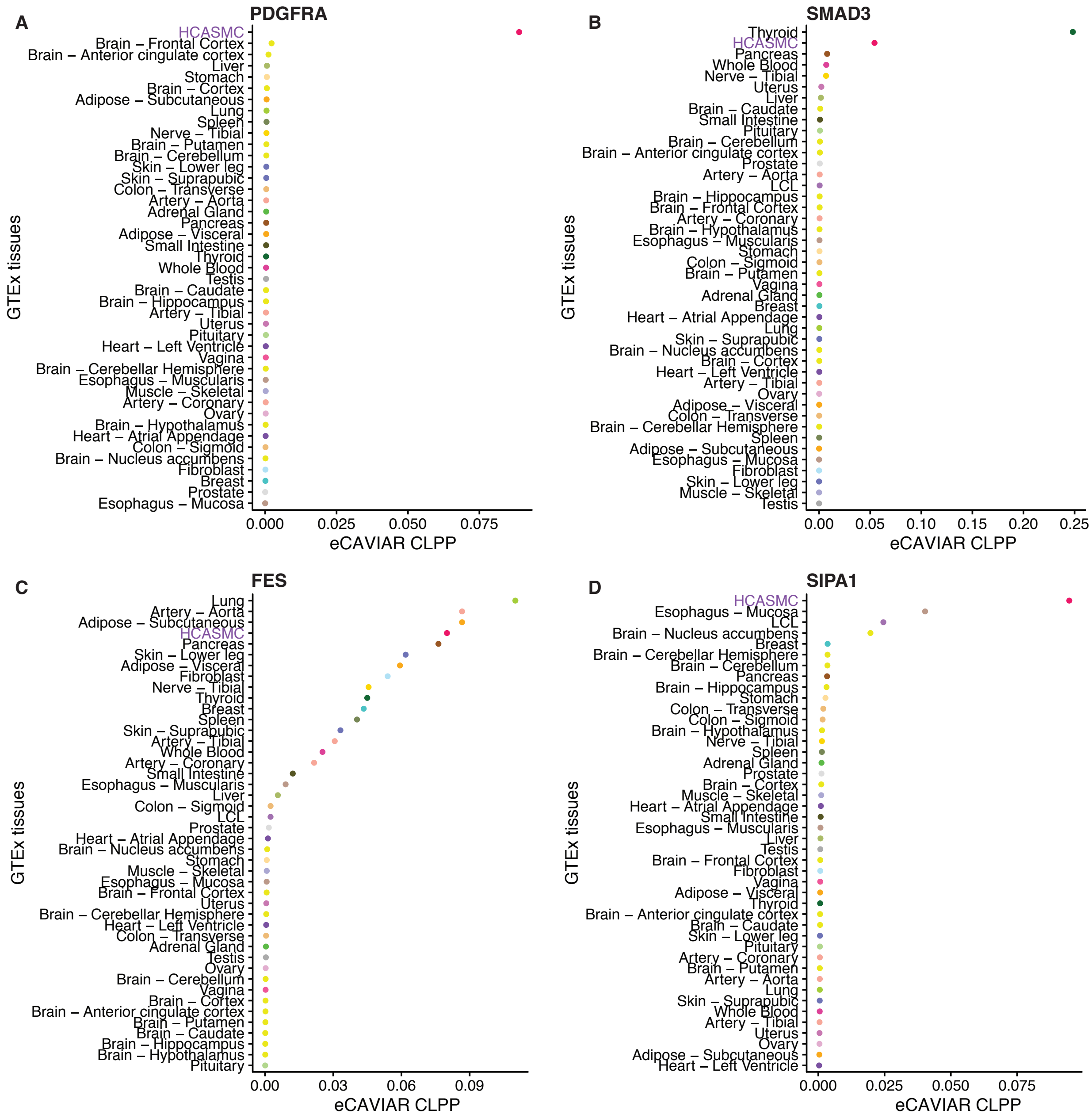


Figure S10. eCAVIAR Posterior colocalization probability in GTEx tissues. We estimated the colocalization posterior probability in GTEx tissues for four causal genes discovered by eCAVIAR. We observed that colocalization of *PDGFRA* and *SIPA1* are highly specific to HCASMC. *SMAD3* colocalization signal is shared in thyroid and HCASMC, and *FES* colocalization signal is shared across multiple tissues.

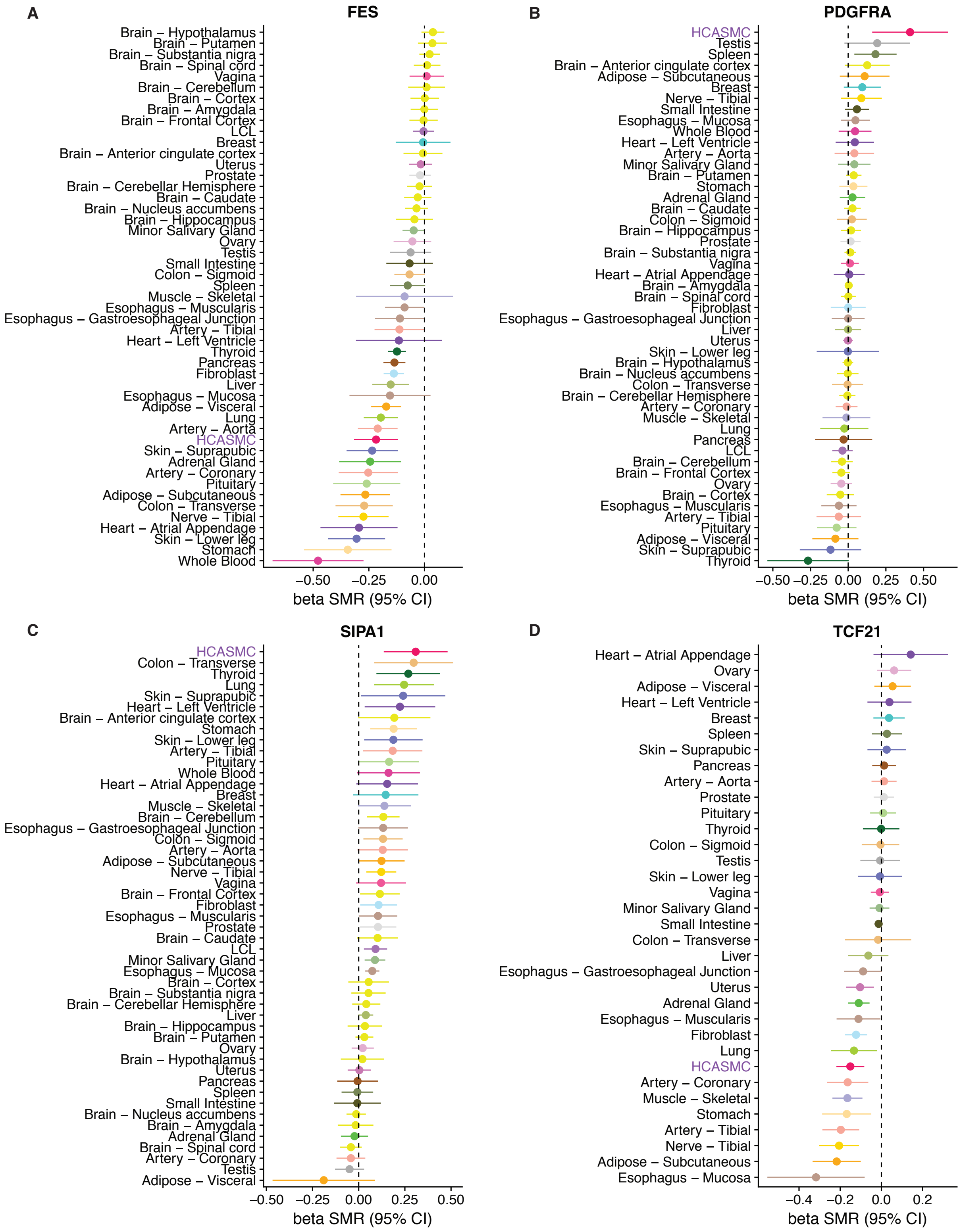


Figure S11. SMR p-values in GTEx tissues. We plotted SMR effect sizes in GTEx tissues for four candidate genes with nominal significance (*FES*, *PDGFRA*, *SIPA1*, and *TCF21)*. *FES* and *TCF21* colocalization are shared across multiple tissues. *PDGFRA* and *SIPA1* colocalization are strongest in HCASMC, in agreement with eCAVIAR results.

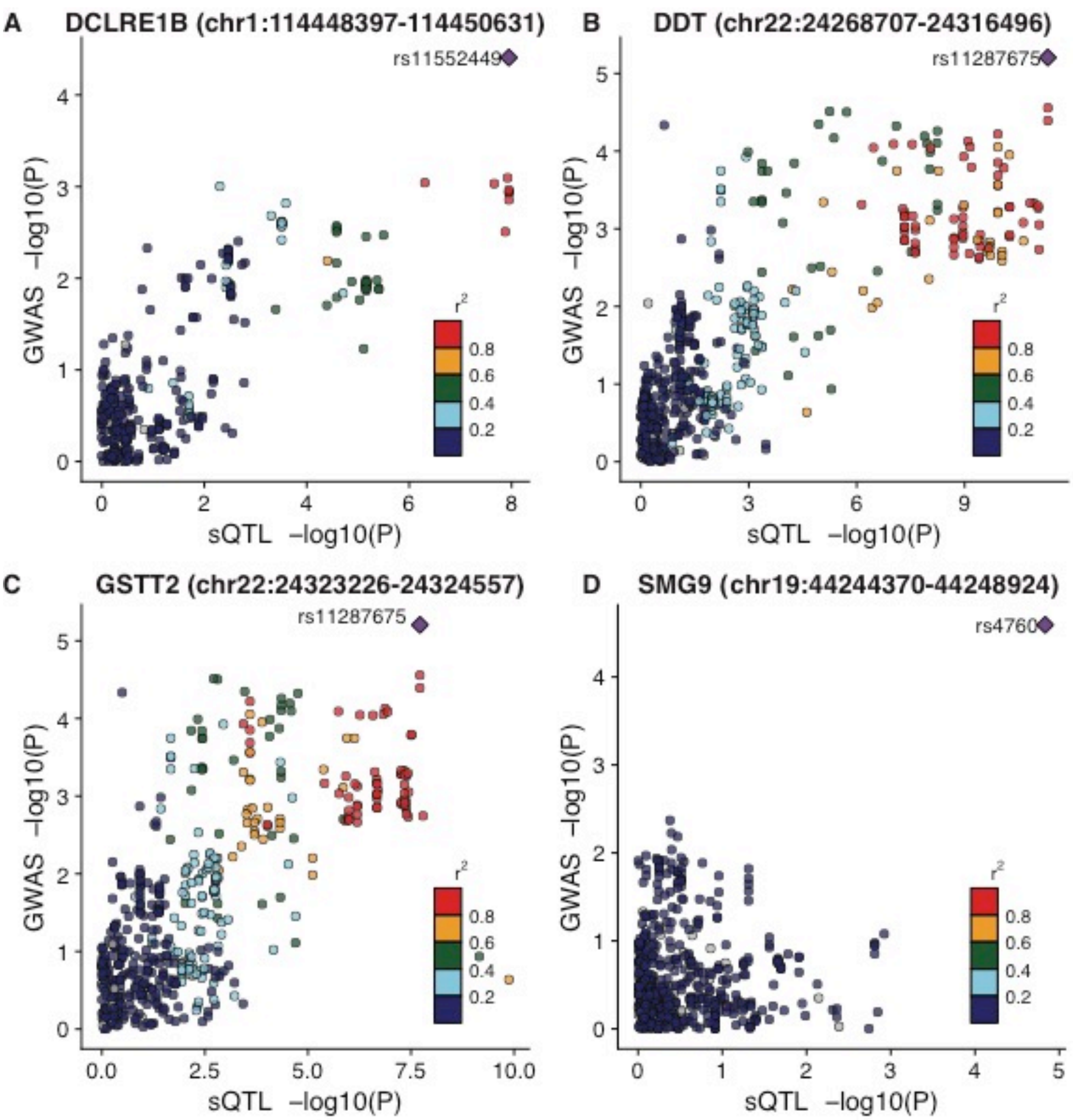
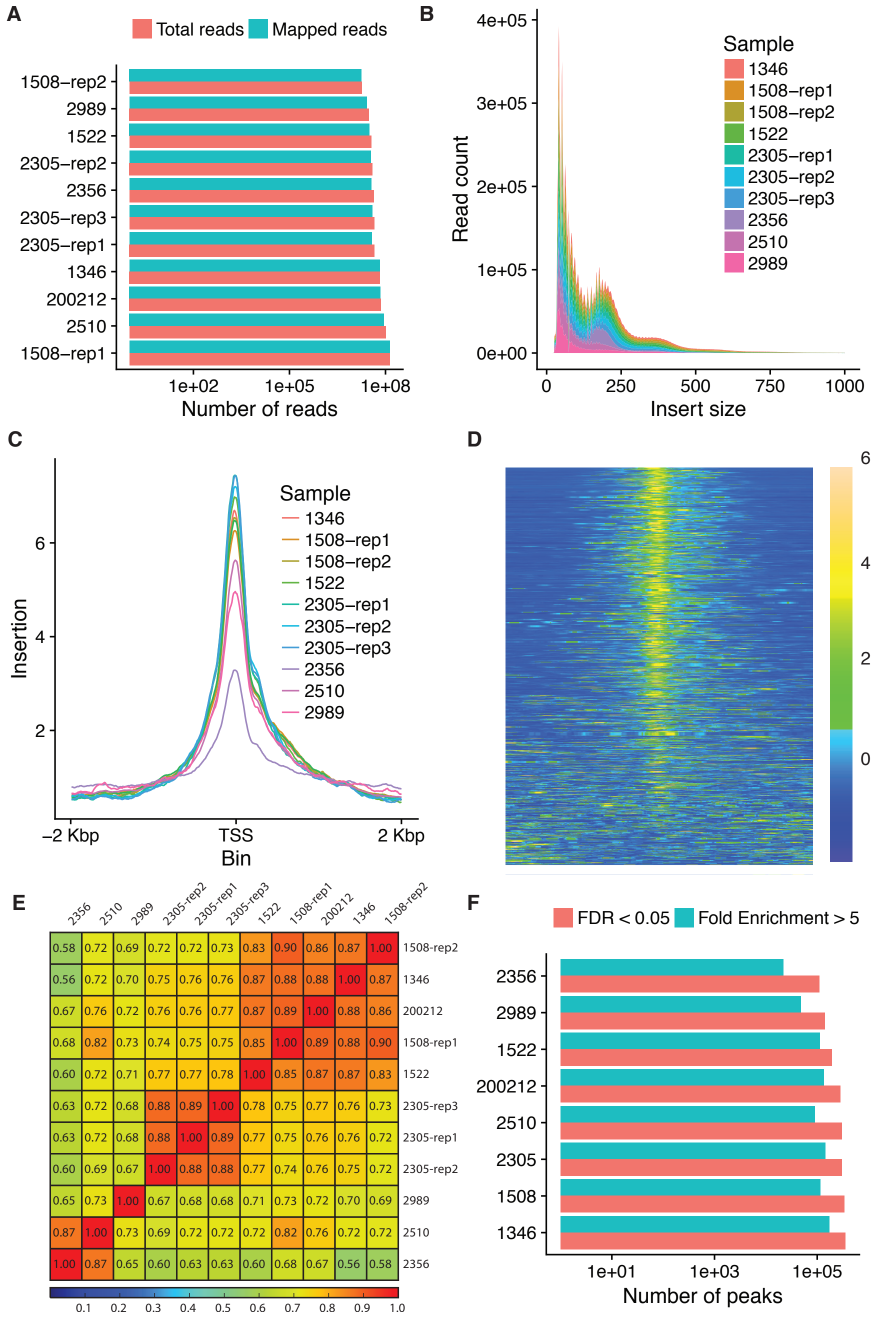


Figure S12. Splicing QTL colocalization. We performed sQTL colocalization analysis using both eCAVIAR and SMR. We found four potential causal genes at *DCLRE1B, DDT, GSTT2,* and *SMG9.* However, none of the GWAS p-values of the four loci reached genome-wide significance ($5x10^{-8}$).



	2356	2510	2989	2305-rep2	2305-rep1	2305-rep3	1522	1508-rep1	200212	1346	1508-rep2
1508-rep2	0.58	0.72	0.69	0.72	0.72	0.73	0.83	0.90	0.86	0.87	1.00
1346	0.56	0.72	0.70	0.75	0.76	0.76	0.87	0.88	0.88	1.00	0.87
200212	0.67	0.76	0.72	0.76	0.76	0.77	0.87	0.89	1.00	0.88	0.86
1508-rep1	0.68	0.82	0.73	0.74	0.75	0.75	0.85	1.00	0.89	0.88	0.90
1522	0.60	0.72	0.71	0.77	0.77	0.78	1.00	0.85	0.87	0.87	0.83
2305-rep3	0.63	0.72	0.68	0.88	0.89	1.00	0.78	0.75	0.77	0.76	0.73
2305-rep1	0.63	0.72	0.68	0.88	1.00	0.89	0.77	0.75	0.76	0.76	0.72
2305-rep2	0.60	0.69	0.67	1.00	0.88	0.88	0.77	0.74	0.76	0.75	0.72
2989	0.65	0.73	1.00	0.67	0.68	0.68	0.71	0.73	0.72	0.70	0.69
2510	0.87	1.00	0.73	0.69	0.72	0.72	0.72	0.82	0.76	0.72	0.72
2356	1.00	0.87	0.65	0.60	0.63	0.63	0.60	0.68	0.67	0.56	0.58

Figure S13. ATAC-seq quality control. (A) As a quality control, we plotted the sequencing depth for each library. The libraries were sequenced to a median of 44.6 million reads (interquartile range: 37.7 – 69.8 millions reads), among which a median of 37.1 millions reads (84%) were mapped (interquartile range: 33.3 – 67.7 millions reads). (B) We plotted the insert size distribution of ATAC-seq reads. The plot shows distinct peaks at length of the linker sequence (10-80 bp) as well as peaks spaced one nucleosome apart (~150 bp). (C) Further, we plotted the distribution of distance to TSS and observed that the ATAC-seq datasets are centered symmetrically around TSS. (D) An example V-plot (1346-rep1) is shown. (E) We plotted the spearman correlation in coverage across all pairs of samples. The correlations range between 0.58 (2356 and 1508-rep2) to 0.9 (1508-rep1 and 1508-rep2). Note that the correlation between replicates are high (>0.88 for 2305, and 0.9 for 1508). (F) The median number of peaks with FDR < 0.05 is 2.8x10^5 (interquartile range: 1.7x10^5 – 3.0x10^5).

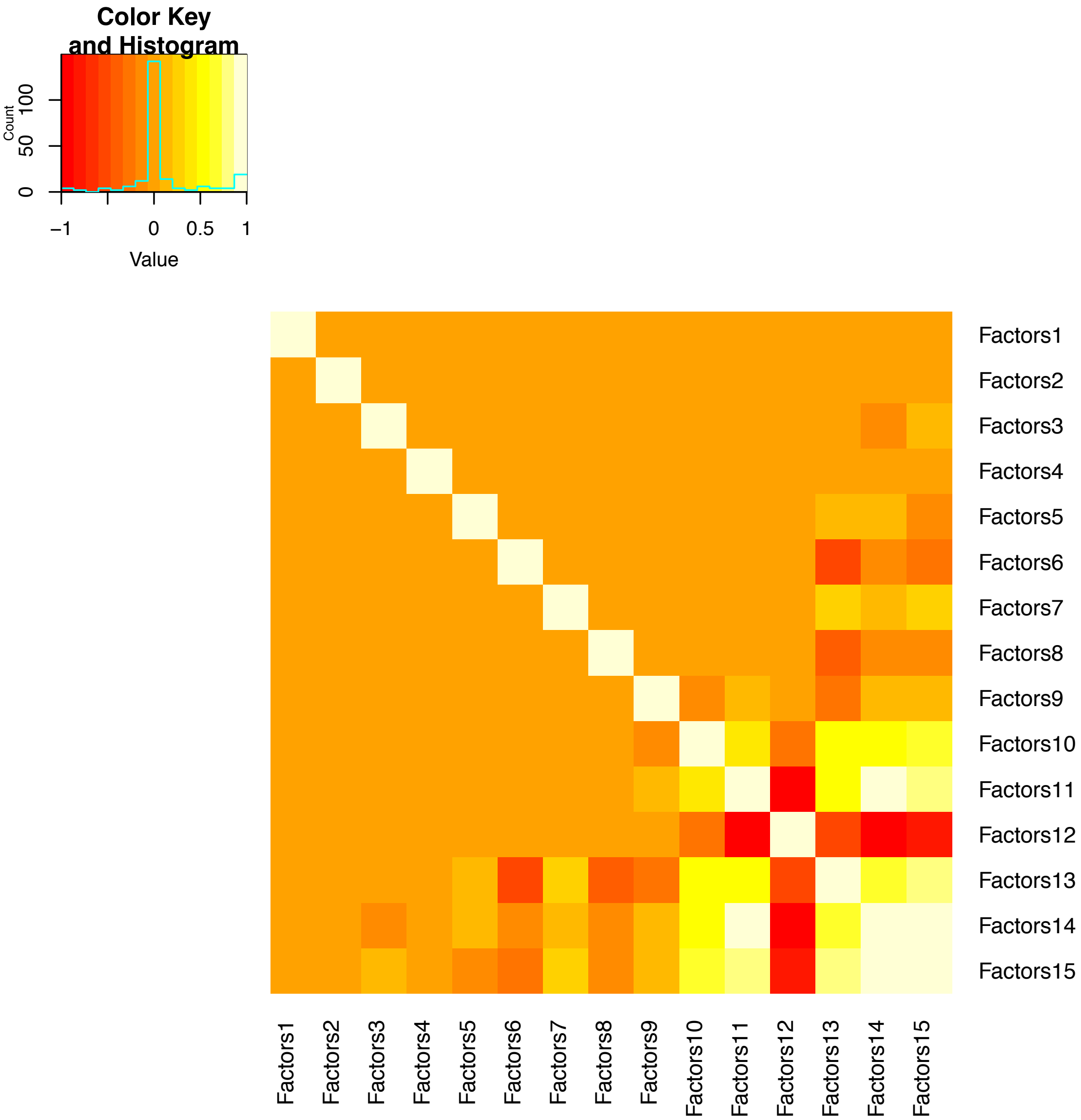


Figure S14. Pairwise PEER factor correlation. We observed factors 1 to 9 showed pairwise independences but factors 10 to 15 appeared correlated. In single-tissue eQTL mapping, we found that 8 peer factors achieve the maximum power for eQTL discovery. We did not use factors 9-15, and regression results were not affected by the correlation in these factors.

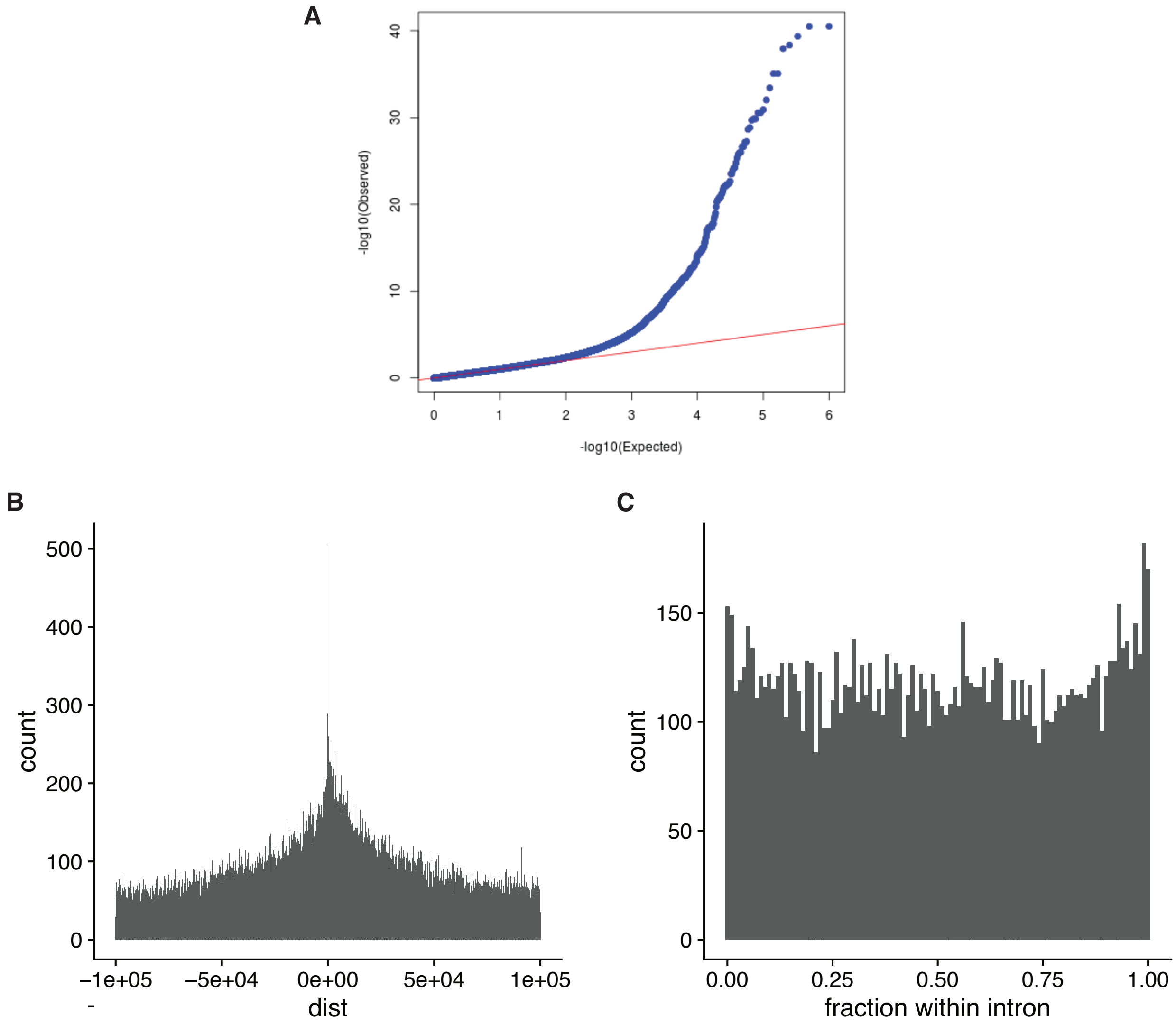


Fig. S15. Splicing QTL quality control. We used a quantile-quantile plot to visualize the p-value distribution (A), distance from sQTL SNP to splice donor site (B), and normalized distance from intronic sQTL SNP within the intron boundary (C). We observed that p-values are enriched towards 0, and that sQTL SNPs are enriched around splice donor sites and acceptor sites.

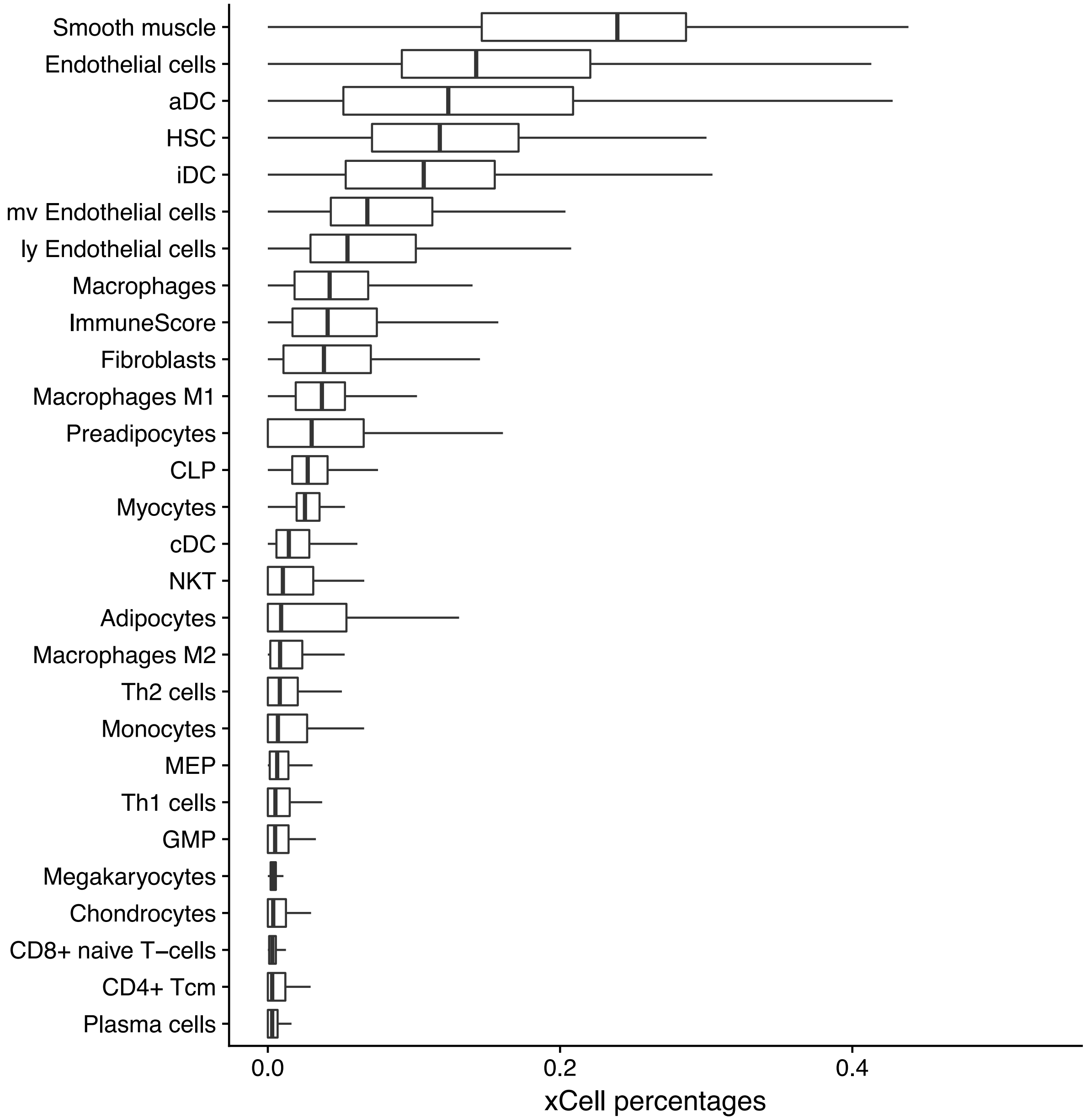


Fig. S16 Cell type deconvolution of coronary artery. We performed cell type deconvolution using xCell on GTEx coronary artery and found that the most enriched cell type is smooth muscle (20-30%), and it is followed by endothelial cells (10-20%).

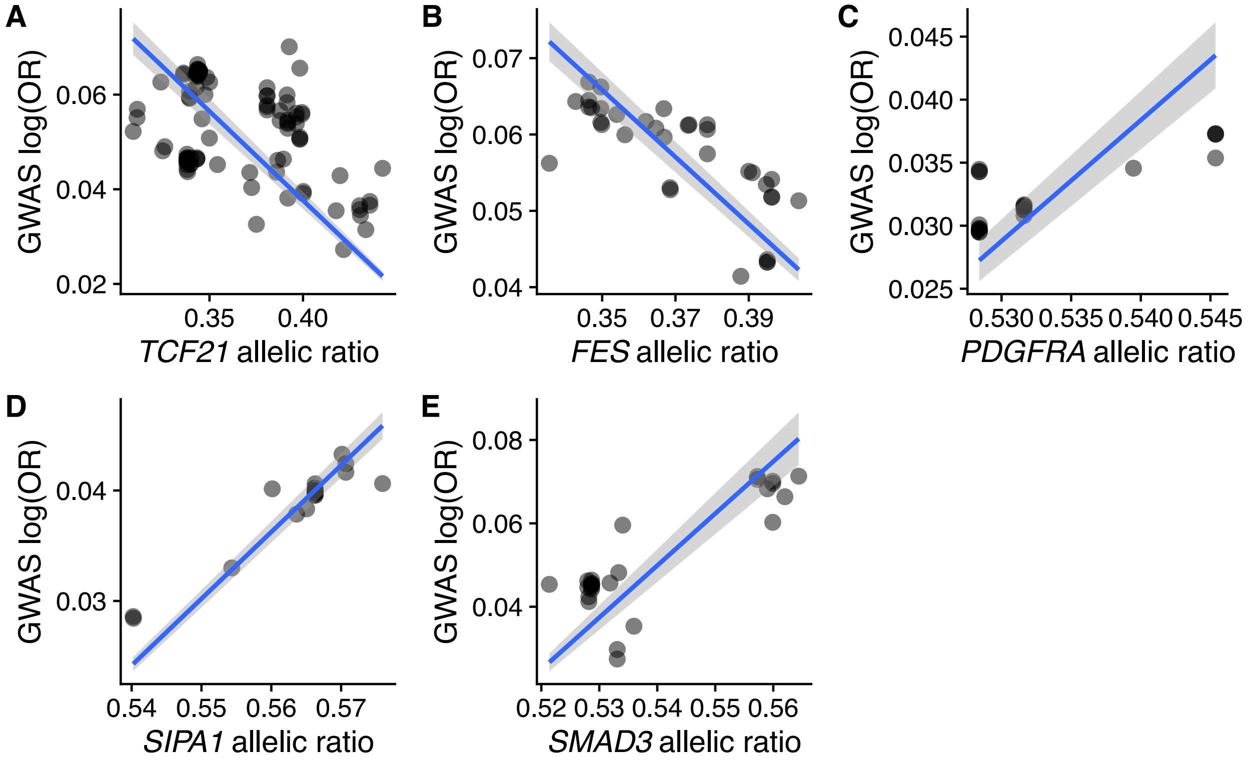


Fig. S17 Direction of effect for potential causal genes. We determined the direction of effect, i.e. whether gene expression upregulation increases risk, using the correlation between the GWAS and the eQTL study effect sizes on SNPs with p-value < 1x10^{-3} in both datasets. Upregulation of genes *TCF21* and *FES* may protect against CAD risk, and upregulation of *SMAD3, PDGFRA,* and *SIPA1* may increase CAD risk. We regressed through the point (0.5, 0) because an allelic ratio of 0.5 indicates equal expression from both alleles (no eQTL effect), and a log odds-ratio of 0 indicates no GWAS effect.

Supplementary Materials and Methods

1. Cell culture

Primary human coronary artery smooth muscle cell (HCASMC) lines derived from non-diseased human donor hearts were purchased from Cell Applications, Inc. (catalog # 350-05a), PromoCell (catalog # C-12511), Lonza (catalog # CC-2583), ATCC (catalog # PCS-100-021), and Lifeline Cell Technology (catalog # FC-0031). All cell lines were cultured in smooth muscle growth medium (Lonza catalog # CC-3182) supplemented with hEGF, insulin, hFGF-b, and 5% FBS, according to Lonza instructions (Table S1). All data presented are from HCASMC expanded to passage 5-6 prior to extraction.

Vendors	# lines	Website
Promocell	19	https://www.promocell.com/
Cell Applications	25	https://www.cellapplications.com/
Lonza	3	https://www.lonza.com/
Lifeline Cell Technology	3	https://www.lifelinecelltech.com/
ATCC	2	https://www.atcc.org/

2. Whole-genome sequencing (Genomics)

2.1 Library preparation and sequencing

We performed paired-end whole-genome sequencing (WGS) on 62 samples to an average depth of 30X. Briefly, genomic DNA was isolated from HCASMC lines using Qiagen DNeasy Blood & Tissue Kit (catalog # 69506) and quantified using NanoDrop 1000 Spectrophotometer (Thermo Fisher) and Agilent Bioanalyzer 2100. DNA libraries were prepared by Macrogen using Illumina's TruSeq DNA PCR-Free Library Preparation Kits and subjected to 150bp paired-end sequencing on an Illumina HiSeq X Ten System.

2.2 Data processing

Whole-genome sequencing data were processed with the GATK best practices pipeline[1,2]. We trimmed Illumina adapters using Cutadapt v1.9[3]. Trimmed reads were aligned to the hg19 reference genome with Burrows-Wheeler Aligner (BWA) v0.7.12[4] using its bwa-mem module[5] with parameters "-M -t 8 -R '@RG\tID:{sample}\tSM:{sample}\tPL:illumina\tLB:lib1\tPU:unit1'“. Duplicate reads in alignment result were marked with Picard v1.92. We performed indel realignment and base recalibration with GATK v3.4[6]. The GATK HaplotypeCaller was used to generate gVCF files, which were fed into GenotypeGVCFs for joint genotype calling. We recalibrated variants using the GATK VariantRecalibrator module. Since subsequent eQTL calling (see Section 12) with RASQUAL[7] required phased variants, we phased our call set with Beagle v4.1[8]. We first used the Beagle conform-gt module to correct any reference genotypes if they are different from hg19. We then phased and imputed against 1000 Genomes project phase 3 version 5a[9]. Variants with imputation allelic r^2 less than 0.8 and Hardy-Weinberg Equilibrium p-value less than $1x10^{-6}$ were filtered out.

2.3 Quality control

To eliminate cross contamination, we used VerifyBamID v1.1.2[10] to estimate the percentage of sample contamination. Since we did not have genotyping microarray data as a reference, we used allele frequencies from 1000 Genomes as a reference input for VerifyBamID. Three samples (1848, 1858, and 24635) with large proportion of contamination (> 0.1) were removed. We next compared genotypes across all pairs of individuals and found two samples (313605 and 317155) were duplicates of each other. We kept the sample with greater coverage (317155). A total of 58 WGS samples remained after filtering.

3. RNA sequencing (Transcriptomics)

3.1 Library preparation and sequencing

We performed RNA sequencing library preparation on 58 samples. Briefly, total RNA was extracted using the Qiagen miRNeasy Mini Prep Kit (catalog # 217004). Quality of RNA was assessed on the Agilent 2100 Bioanalyzer. Samples with RNA Integrity Number (RIN) greater than or equal to 8 were sent to the Next-Generation Sequencing Core at the Perelman School of Medicine at the University of Pennsylvania. Libraries were constructed using the Illumina TruSeq Stranded Total RNA Library Prep Kit (catalog # 20020597), multiplexed using D501-D508 i5 adapters and D701-D712 i7 adapters, and subjected to 125bp paired-end sequencing on a HiSeq 2500 platform with a median depth of 51.3 million reads (interquartile range 45.5 – 58.7 million reads), generating over 2.7 billion reads in total.

3.2 Data processing/Quality control

Demultiplexing was performed with bcl2fastq script from Illumina. Base quality control of demultiplexed sequences was done using FastQC v0.11.4 quality control tool. Fastq files that correspond to the unique sample were merged using a custom script. Mapping of the reads was performed with STAR v2.4.0i[11]. In accordance to the GATK Best Practices for RNA-Seq we used the STAR 2-pass alignment pipeline[1,2]. First, reads contained in the raw fastq files were mapped to GRCh37/hg19 human genome using STAR and during the first alignment pass splice junctions were discovered with high stringency. Second pass mapping with STAR was then performed using a new index that was created with splice junction information contained in the file SJ.out.tab from the first pass STAR mapping. Splice junctions from the first pass were used as annotation in a second pass to permit lower stringency alignment, and therefore higher sensitivity. Prior to gene expression quantification, we used WASP[12] to filter out reads that are prone to mapping bias. Read counts and RPKM were calculated with RNA-SeQC v1.1.8[13] using default parameters with additional flags "-n 1000 -noDoC -strictMode" using GENCODE v19 annotation[14]. Allele-specific read counts were generated with the createASVCF module in RASQUAL[7]. We quantified intron excision levels using LeafCutter[15]. In brief, we converted bam files to splice junction files using the bam2junc.sh script, and defined intron clusters using leafcutter_cluster.py with default parameters. This requires at least 30 reads supporting each cluster and at least 0.1% of reads supporting each intron within the cluster, and allows intron to have a maximum size of 100kb.

3.3 Quality control

To eliminate low-quality RNA sequencing samples, we checked RIN prior to sequencing, and mapping rate, duplication, and cross contamination after sequencing. Four samples were removed during library prep due to low RIN (1497, 1923, 2161, and 2477), one sample (2115) was removed after alignment due to extremely low mapping rate (<2% mapped reads), and one sample was removed due to contamination (24156). With the remaining 52 samples, We used VerifyBamID v1.1.2[10] to check for RNA and DNA sample concordance, and did not find any mislabeling.

4. ATAC sequencing (Epigenomics)

4.1 Library preparation and sequencing

We assessed chromatin accessibility with ATAC-seq with slight modifications to the published protocol, as previously described[16]. Briefly, HCASMCs (passages 5–6) were cultured in normal media until ~75% confluence. Approximately 5×10^4 fresh cells were collected by centrifugation at 500 g and washed twice with cold 1 × PBS. Nuclei-enriched fractions were extracted with cold lysis buffer containing 10 mM Tris–HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPAL (octylphenoxypolyethoxyethanol), and the pellets were resuspended in transposition reaction buffer containing Tn5 transposases (Illumina Nextera). Transposition reactions were incubated at 37 °C for 30 min, followed by DNA purification using the DNA Clean-up and Concentration kit (Zymo). Libraries were initially PCR amplified using Nextera barcodes and High Fidelity polymerase (NEB). The number of cycles was empirically determined from an aliquot of the PCR mix, by calculating the Ct value at 25–30% maximum Rn for each library preparation. The final amplified library was again purified using the Zymo DNA Clean-up and Concentration kit, and the DNA was evaluated by TBE gel electrophoresis and quantified using Bioanalyzer, nanodrop and quantitative PCR (KAPA Biosystems). Libraries were multiplexed and paired-end 50-bp sequencing was performed using an Illumina NextSeq 500. Raw output files were demultiplexed to generate individual FASTQ files, which were then evaluated using a modification of the FastQC 0.11.4 pipeline to generate per base and per sequence-level summary statistics. The libraries had a median of 45.0 million reads (interquartile range: 38.5 – 79.7 millions reads).

4.2 Data processing

We used the ENCODE ATAC-seq pipeline to perform alignment and peak calling (https://github.com/kundajelab/atac_dnase_pipelines)[17]. FASTQ files were trimmed with Cutadapt v1.9[3] and aligned with Bowtie2 v2.2.6[18] with default parameters. Duplicate reads were marked with Picard v1.126. The alignment were converted to ENCODE tagAlign format. Records were shifted +4 and -5 for positive-strand and minus-strand reads. MACS2 v2.0.8[19] was used to call peaks with default parameters. Each alignment was split into two pseudo-replicate (subsample of reads) and peaks were called independently. Irreproducible Discovery Rate (IDR)[20] analyses were performed based on pseudo-replicates with a cutoff of 0.1 to output an IDR call set, which was used for downstream analysis. We used WASP[12] to filter out reads that are prone to mapping bias.

4.3 Quality control

The libraries were sequenced to a median of 44.6 million reads (interquartile range: 37.7 – 69.8 millions reads), among which a median of 37.1 millions reads (84%) were mapped (interquartile range: 33.3 – 67.7 millions reads, Fig. S13A). We plotted the insert size distribution of ATAC-seq reads (Fig. S13B). The plot shows distinct peaks at length of the linker sequence (10-80 bp) as well as peaks spaced one nucleosome apart (~150 bp). Further, we plotted the distribution of distance to TSS and observed that the ATAC-seq datasets are centered symmetrically around TSS (Fig. S13C). An example V-plot (1346-rep1) is shown in Fig. S13D. We used deepTools v2.5.7[21] to calculate spearman correlation in coverage across all pairs of samples. The correlation ranges between 0.58 (2356 and 1508-rep2) to 0.9 (1508-rep1 and 1508-rep2). Note that the correlation between replicates are high (>0.88 for 2305, and 0.9 for 1508, Fig. S13E). The median number of peaks with FDR < 0.05 is $2.8x10^5$ (interquartile range: $1.7x10^5$ – $3.0x10^5$, Fig. S13F). We used all samples for downstream analysis.

5. External datasets

5.1 Genotype-Tissue Expression dataset

We used the GTEx v6p dataset (https://www.gtexportal.org/home/datasets), which sampled 53 tissues across 569 individuals[22]. We used all tissues for multidimensional scaling analysis (Section 6), and tissue-specific gene expression analysis (Section 10). Not all tissues were sampled in each individual, and 44 tissues are sampled across more than 70 individuals. These tissues were previously selected by GTEx for eQTL analysis. Therefore, we used the eQTL callset across these 44 tissues for both METASOFT analysis (Section 12) and tissue-specific GWAS colocalization analysis (Section 14).

5.2 ENCODE dataset

ENCODE datasets were downloaded from the data portal (https://www.encodeproject.org/). We downloaded DNAseI-seq datasets for all human tissues and primary cells, which totaled 577 datasets across 83 unique tissues and cell types. We concatenated across datasets for tissues with multiple replicates. To mitigate differences in sequencing and peak calling, we standardized all peaks using a unified framework. We reasoned that peak summits have higher read counts and are more readily defined than peak boundaries. We therefore identified the peak summit, and defined each peak as +/- 75bp around the summit.

6. Multidimensional scaling in transcriptomic space (Fig. 1A)

To determine whether the HCASMC transcriptome is unique from those in GTEx and to understand which GTEx tissues neighbor HCASMC in the transcriptomic space, we performed non-parametric clustering analysis. To remove genes lowly expressed in most tissues, we filtered for genes with at least 0.1 RPKM in at least 10 individuals. We applied Kruskal's non-metric multidimensional scaling on log-transformed RPKM values across 53 GTEx tissues as well as HCASMC. This method tries to minimize the square root of the ratio between the sum of the squared differences between the input distances and those of the low-dimensional configuration and the of sum the configuration distance squared[23].

$$S = \sqrt{\frac{\sum_{i<j}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}}$$

Where d_{ij} represent the distance between point *i* and point *j* in high dimensions, and $\hat{d}_{ij}$ represent the distance in lower dimensions. For the convenience of visualization, we specified the desired dimension to be two.

7. Jaccard similarity in epigenomic space (Fig. 1B)

To understand which ENCODE cell types neighbor HCASMC in terms of chromatin structure, we calculated the epigenetic distance defined by Jaccard similarity:

$$Jaccard(A, B) = \frac{length(A \cap B)}{length(A \cup B) - length(A \cap B)}$$

Where A and B represents the open chromatin peaks measured by either ATAC-seq or DNaseI-seq. Since each cell and tissue type in the ENCODE dataset may have multiple samples and replicates, we concatenated peaks across samples and replicates for each cell and tissue type according to Table S3. We used a standardized peak boundary defined as +/- 75bp around the peak summit (approximately the size of one nucleosome). We stratified cell lines into adult and fetal tissue groups to reflect the difference in chromatin structure across developmental stages.

8. Differential expression and pathway enrichment (Fig. 1C)

8.1 Differential expression

To identify the transcriptomic differences between HCASMC and neighboring tissues in GTEx, we performed differential expression analysis. To match the sample size of HCASMC (n = 52), we sampled 52 individuals for each GTEx tissue. To minimize sharing of individuals between tissues, we prioritized individuals that have not been sampled before. The following steps were carried out:

1) Rank all tissues by sample size in increasing order.
2) Sample 52 individuals for the tissue with smallest sample size (or use all individuals if N < 52).
3) Repeat for the tissue with next smallest sample size.
   a) Sample previously unselected individuals.
   b) If not enough, sample from other individuals.

To detect differentially expressed genes in HCASMC relative to other relevant cardiovascular tissues, we compared HCASMC with the following tissue groups:

Tissue group	GTEx Tissues

Heart	Heart - Atrial Appendage (n=194) Heart - Left Ventricle (n=218)
Arteries	Artery - Aorta (n=224) Artery - Coronary (n=133) Artery - Tibial (n=332)
Fibroblast	Cells - Transformed fibroblasts (n=284)

We found that the sample sizes are sufficient such that each individual is sampled at most once, and therefore there is no overlap across tissues.

We performed differential expression between HCASMC and each GTEx tissue with DESeq2 v1.10.1[24] using read count as input, controlling for sex, ancestry and hidden confounders (surrogate variables). The surrogate variables were extracted using the svaseq function in sva v3.18.0[25] by protecting sex, ancestry and tissue. We performed the same comparison with NOISeq v2.14.1[26] using RPKM as input. In general, we find that NOISeq is more conservative than DESeq2, and the correlation of effect size are high (r^2~0.9) for significantly DE genes (FDR<0.001). To be conservative, we called a gene differentially expressed if and only if it is significant by both methods. For DESeq2, we used FDR < 0.05, 0.01, and 0.001 as the significance cutoff. For NOISeq, we used q > 0.95, 0.99 and 0.999 as the significance cutoff (q = 1 - FDR). Overall, we found that thousands of genes are differentially expressed between HCASMC and its close neighbors, on par with cross-tissue variability previously observed[27].

8.2 Pathway enrichment

To obtain differential expression profiles between HCASMC and broad tissue groups, we performed meta-analysis across tissues within each tissue group using effect size and standard error estimated by DESeq2. Since a given gene can have different expression levels across tissues, we used a random effect model to account for the difference in effect sizes. We used the R package metafor v2.0[28] to conduct meta-analysis. To control for multiple hypothesis testing, we corrected p-values using the Benjamini-Hochberg FDR[29]. Under serum treatment, HCASMC undergo epithelial-mesenchymal transition from the contractile state to the synthetic state, in which it enters the cell cycle and secretes extracellular matrix proteins. As expected, we found that genes related to proliferation (such as mTORC1, Myc, mitotic spindle, G2M checkpoint), epithelial-mesenchymal transition, and protein secretion are upregulated in HCASMC, compared with fibroblast, artery and heart (Table S2).

8.3 The effect of age and sex on HCASMC transcriptome

To determine whether the transcriptome vary by age and sex, we performed differential expression with DESeq2 using age, sex and ancestry as covariates. We found two genes that are differentially expressed with age (FDR < 0.05), *KRT16* (keratin 16), which are intermediate filament proteins guarding the structural integrity of epithelial cells, and ENSG00000254337 (a lincRNA), whose function has not been studies in depth.

We found 90 genes differentially expressed with sex. Among these genes, 36 are on sex chromosomes (chrX = 15, chrY = 21). One notable gene is *DAZL*, which has been known to express in germ cells of males and females. Another example is *PPP1R14A*, which is an inhibitor of smooth muscle myosin phosphatase. Sex difference in this gene may contribute to differential susceptibility to CAD.

9. HCASMC-specific open chromatin peaks (Fig. 1D and E)

9.1 Tissue-specific peak calling

Using the entire ENCODE open chromatin dataset, we determined the number of HCASMC-specific open chromatin regions. For HCASMC, we used the sample with the largest number of biological replicates (n=3) to obtain the most robust call set after IDR analysis (Section 4.2). For each ENCODE sample with more than one replicate, we merged all replicates with bedtools merge[30] command, without using IDR analysis. As a result, the merged ENCODE datasets for each tissue have a larger number of peaks than HCASMC, and this will result in a more conservative estimate of HCASMC-specific peaks. We used bedtools v2.26.0[30] multi-intersect module to determine the peak intersection across all ENCODE tissues and primary cell cultures (Table S3). We defined HCASMC-specific peaks as those that appeared only in HCASMC and not any other tissue or cell type in the ENCODE dataset.

9.2 Transcription factor binding motif enrichment

To identify transcription factor binding sites enriched in HCASMC-specific open chromatin regions we employed the HOMER tool[31] (Hypergeometric Optimization of Motif EnRichment, http://homer.salk.edu/homer/ngs/). HOMER's findMotifsGenome.pl script was used to search for known TRANSFAC motifs and to generate *de novo* motifs. All software parameters were set to default values, with the addition of the ‘-size’ command to define the width around the peak center. Selected were three parameters of 50bp, for establishing the primary motif bound, and 200bp and 1000bp for finding both primary and "co-enriched" motifs. Motifs discovered by HOMER were validated with MEME-ChIP. The motifs identified by MEME-ChIP were further compared with the binding motifs of known TFs. To validate that found motifs were enriched in HCASMC-specific peaks (and not just any tissue-specific peaks), we performed TFBS enrichment analysis for tissue-specific peaks across HCASMC, fibroblast, heart, and brain (as a negative control). This yielded 24,372 specific peaks for brain, 7,332 for HCASMC, 6,388 for lung fibroblasts, and 11,249 for heart. Density plots were generated using ChIP-Cor Analysis Module Feature Correlation Tool v1.5.3 (https://ccg.vital-it.ch/chipseq/chip_cor.php). Input range was defined as +/- 10000 bp from the centered features. Window width was set to 500 bp and counts cut-off value was set to 999999. Normalization was global, i.e. histogram entries were normalized by the total number of reference and feature counts and the window width. FoxP1, FoxA1, FoxO1, Atoh1 and NFIC PWM matrices were derived from the JASPAR CORE vertebrates 2018 database[32]. PWMScan - Genome-wide position weight matrix (PWM) scanner was used to scan the GRCh37/h19 version of the human genome using the following parameters: p-value cut-off 0.0001, background base composition 0.29,0.21,0.21,0.29, search strand both.

10. Heritability enrichment around tissue-specific genes (Fig. 2A)

To understand whether CAD variants are enriched around HCASMC-specific genes, we compared disease heritability between HCASMC and GTEx tissues using stratified LD score regression[33]. We obtained median gene expression for each tissue. Since stratified LD score regression requires SNP annotations to be uncorrelated, we selected for independent tissues using the following algorithm:

1. Select a starting tissue (in our case HCASMC).
2. Remove tissues with median expression Pearson's r > 0.96 with the given tissue.
3. Select a random tissue from the remaining tissues and repeat the process.

Further, we removed the following tissues that primarily consist of smooth muscle to avoid correlation with HCASMC (Cervix - Endocervix, Colon - Sigmoid, Esophagus - Mucosa, Vagina, Stomach).

After this procedure, 16 tissues remained:
HCASMC, Adipose - Subcutaneous, Adrenal Gland, Artery - Coronary, Brain - Caudate (basal ganglia), Cells - EBV-transformed lymphocytes, Cells - Transformed fibroblasts, Liver, Lung, Minor Salivary Gland, Muscle - Skeletal, Pancreas, Pituitary, Skin - Not Sun Exposed (Suprapubic), Testis, Whole Blood

We used z-scores to define tissue-specific genes. In particular, we first calculated the median expression of each gene across individuals for each tissue. For each gene, we calculated the mean and standard deviation of median expression across tissues, from which we derive a z-score.

$$\tilde{e}_t = median(\boldsymbol{e_t})$$
$$z = \frac{\tilde{e}_t - E(\tilde{e}_t)}{Var(\tilde{e}_t)}$$

Where $\boldsymbol{e_t}$ is a vector that contains gene expression across all individuals in tissue *t*. We ranked each gene based on the z-score with a higher z-score corresponding to higher expression specificity. We selected the top 1000, 2000, and 4000 genes as tissue-specific genes. We assigned a given SNP to a gene if it falls into the union of exon +/- 1kbp of that gene. We reasoned that the variants affecting protein function and the strongest regulatory variant generally fall within or near exons[34]. Although a 1kb window may not capture all regulatory variants, we decided to keep a small window to capture gene-specific effects.

We estimated the heritability enrichment using stratified LD score regression on a joint SNP annotation across all 16 tissues against the CARDIoGRAMplusC4D GWAS meta-analysis[35]. As expected, coronary artery has the highest heritability enrichment. Subcutaneous adipose tissue has the second highest enrichment, reflecting the well-known link between lipid metabolism and coronary artery disease. We found that HCASMC has the third highest enrichment, suggesting that it is highly relevant to coronary artery disease risk. Estimate of heritability enrichment is robust to the number of top genes (n = 1000, 2000 and 4000, Fig. S8). It is worth noting that heritability enrichment estimate decreases as the number of tissue-specific genes increases, further indicating that top tissue-specific genes have higher heritability enrichment.

11. Open chromatin and GWAS overlap (Fig. 2B)

To investigate whether CAD risk variants are enriched in the open chromatin regions in any particular tissue or cell type, we estimated the likelihood of observing given number of GWAS variants falling into open chromatin regions of each tissue using the CARDIoGRAMplusC4D GWAS meta-analysis[35].

We used a recently published method specifically designed to assess the likelihood of GWAS-chromatin overlap, GREGOR[36], to assess the significance of overlap. In addition, we modified GREGOR (https://github.com/boxiangliu/vsea) to estimate the odds ratio between GWAS vs. background variants in terms of open chromatin overlap. Because of linkage disequilibrium and finite sample sizes, many lead GWAS variants are likely to be tagging the true causal variants. We therefore expanded given loci to include all variants in LD (r^2>0.7) with the lead variant. Given a set of GWAS variants, we selected 500 background variants matched by 1) number of variants in LD, 2) distance to the nearest gene, and 3) minor allele frequency, and 4) gene density in a 1Mb window. We calculated the odds ratio between GWAS variants and background variants and used the bootstrap to obtain confidence intervals by comparing HCASMC against ENCODE adult tissues and primary cell lines.

12. Expression quantitative trait (eQTL) analysis

12.1 Inferring Genotype Principal Components

The HCASMC donors come from diverse background, including Caucasian, Hispanic, African, and Asian. We inferred ancestry PCs using the R package SNPRelate[37]. We used PLINK v1.9[38] to convert VCF files to BED, FAM and BIM. Before inferring ancestry PCs, we pruned SNP using Hardy-Weinberg equilibrium (HWE > 1e-6), LD (R^2 > 0.2) and minor allele frequency (MAF > 0.05) using SNPRelate[37]. The final ancestry PCs are plotted in Fig. S4.

12.2 Inferring hidden confounders with PEER

RNA-seq experiments are often confounded by unrecorded (hidden) technical artifacts. To correct such hidden confounders, we extracted covariates that have global influence over a large number of genes using Probabilistic Estimation of Expression Residuals (PEER)[39]. Following the GTEx standard pipeline for estimating hidden factors, we applied PEER to the 10,000 most highly expressed genes across all samples. To ensure each sample followed the same distribution we quantile normalized RPKM values. Since PEER assumes normality in gene expression, we performed rank-based inverse normal transformation. We used the following PEER parameters to extract the top 15 hidden factors: MaxFactorsN=15, MaxIterations=10000, BoundTol=0.001, VarTol=0.00001, e_pa=0.1, e_pb=10, a_pa=0.001, a_pb=0.1. We plotted the pairwise correlation for hidden factors, and observed that factors 1 to 9 showed pairwise independence but factors 10 to 15 were correlated (Fig. S14). In the next section, we used cross-validation to select the number of factors which maximized the power to detect eQTL. The cross-validation analysis demonstrated that using the first 8 PEER factors was sufficient to obtain the highest power, therefore circumventing the correlation in factors 10 to 15.

12.3 Single-tissue eQTL calling

We mapped eQTLs using both FastQTL v2.184_gtex[40]and RASQUAL[7]. FastQTL uses total read count information, whereas RASQUAL integrates total read count with allele-specific expression (ASE). RASQUAL requires GC-corrected library sizes and therefore we calculated GC content based on the GTEx v6p distribution of GENCODE v19 by taking the average GC content of all exons of a given gene. The library sizes were calculated based on read count output from RNA-SeQC v1.1.8[13]. To select a combination of covariates that maximize the power to detect eQTLs, we tested combinations of 3, 4 and 5 genotype principal components with 1 to 15 PEER factors. To avoid overfitting, we only used chromosome 20 as a training set. We found that the combination of 4 genotype PCs with 8 PEER factors provided the most power to detect eQTLs. We then used sex, the top four genotype principal components, and the top eight PEER covariates to map eQTLs with both FastQTL and RASQUAL. Mathematically, the eQTL model is:

$$E(e|g, sex, PC, PEER) = \beta_0 + \beta_g \cdot g + \beta_s \cdot sex + \sum_{i=1}^{4} \beta_{a,i} \cdot PC + \sum_{i=1}^{8} \beta_{p,i} \cdot PEER$$

Where *e* stands for gene expression, and *g* stands for the genotype of the test SNP.

12.4 Multiple-hypothesis testing

For FastQTL, we calculated per-gene eQTL p-values using its permutation mode with --permute 1000 100000. We used the q-value package[41] to obtain adjusted p-values per gene using FDR < 0.05 as the cutoff. For RASQUAL, it was not computationally feasible to perform permutation testing on the gene level. Therefore, we used a hierarchical multiple-hypothesis correction procedure, TreeQTL, which was designed specifically for eQTL discoveries[42]. Note that TreeQTL is more conservative than permutation[22]. We chose level 1 (gene level) and level 2 (SNP level) FDR to be less than 0.05.

12.5 Enrichment of eQTLs in genomic features

As a quality control, we estimated enrichment for the top 1000 eQTLs within 11 functional annotations, including downstream-gene variant, exonic variant, intronic variant, missense variant, splice-acceptor variant, splice-donor variant, splice-region variant, synonymous variant, upstream-gene variant, 3' UTR variant, and 5' UTR variant.

The estimation was done with a modified version of GREGOR (Section 11). In brief, for each eQTL, we obtained 200 background variants matched for distance to TSS, LD SNPs, and MAF. We then calculated the enrichment, which is defined as the odds ratio between the probability of QTL overlapping the genomic annotation versus that of background variants. We used the bootstrap to obtain the confidence interval. As expected, we observed that 5' UTR, nearest to the TSS, was most enriched for eQTLs (Fig. S7A).

We validated that eQTLs are enriched in open chromatin regions measured by ATAC-seq. We selected the top SNP per gene, and intersected these SNPs with HCASMC open chromatin regions. We plotted the expected versus the observed p-values and found that eQTLs in ATAC-seq regions have a more pronounced upward trend, suggesting that strong eQTLs are enriched in open chromatin regions (Fig. S5).

12.6 Tissue-specific eQTL calling with METASOFT

To determine which genes are influenced by HCASMC-specific regulation, we compared the HCASMC eQTLs against GTEx eQTLs. To achieve a fair comparison between tissues we subsampled GTEx tissues to 52 individuals to match the number of HCASMC samples, and used FastQTL to call both GTEx and HCASMC eQTLs. For each HCASMC eGene, we selected the top eSNP, and performed multi-tissue eQTL calling using METASOFT[43] (with options -mvalue true and -mvalue_method mcmc to speed up computation). For each eSNP, we obtained 45 m-values for 44 GTEx tissue plus HCASMC. METASOFT authors recommend using m-value > 0.9 for eQTL and m-value < 0.1 for non-eQTL. Therefore, we defined HCASMC-specific eQTLs as having m-value > 0.9 in HCASMC and m-value < 0.1 in all GTEx tissues. Using this method, we found four HCASMC-specific eQTLs, RPAIN, CFB, FAM180A, and LINC01018 (Fig. S6).

13. Splicing quantitative trait loci (sQTL) analysis

13.1 Mapping sQTL with FastQTL

To determine whether any CAD variants act through splicing mechanisms, we decided to map splicing QTLs (sQTLs) genome-wide using a recently published tool called LeafCutter[15]. To correct for known and hidden confounders, we included sex, genotype PCs, and splicing PCs as covariates. We sought to find a set of covariates for best statistical power by testing combinations of 2 to 4 genotype PCs and 1 to 15 splicing PCs with a grid search. The search was performed with only chromosome 22 to avoid overfitting. We found that 3 genotype PCs and 6 splicing PCs returned the largest number of discoveries after at FDR < 0.05. To map sQTLs, we used FastQTL with sex, genotype PCs, and splicing PCs as covariates, and tested SNPs within a 100kbp window. As quality control, we visualized the p-value on a QQ-plot, which indicated enrichment of significant p-values (Fig. S15A). Further, we plotted the number of significant sQTLs (p-value < $1x10^{-4}$) versus their distance to the splice donor and acceptor sites. As expected, sQTL SNPs were enriched around the TSS (Fig. S15B). In addition, we visualized the number of sQTLs with respect to intron boundaries. As expected, sQTL SNPs are slightly enriched around the boundaries as compared to the intron center (Fig. S15C).

13.2 Enrichment of sQTL in genomic elements

To understand the genomic architecture underlying sQTLs, we estimated the enrichment of sQTLs across various genomic annotations and compared them against the enrichment of eQTLs. We took the top 1000 sQTLs and estimated their enrichment within a set of 11 annotations (including downstream-gene variant, exonic variant, intronic variant, missense variant, splice-acceptor variant, splice-donor variant, splice-region variant, synonymous variant, upstream-gene variant, 3' UTR variant, and 5' UTR variant). The estimation was done with a modified version of GREGOR (see Section 11). In brief, for each sQTL, we took 200 background variants matched for distance to splice donor or acceptor site, LD SNPs, and MAF. We then calculated the enrichment, which is defined as the odds-ratio between probability of QTL overlapping the genomic annotation versus that of background variants. We use bootstrap to get the confidence interval.

In contrast to eQTLs (section 12.3), the sQTL are highly enriched in splice donor site, splice acceptor site, and splice region annotations (Fig. S7B). Note that the confidence interval of splice donor and acceptor variants are wider than other annotations because there are fewer of splice donor and accpetor variants.

14. GWAS Colocalization (Fig. 3)

14.1 Summary-data based Mendelian Randomization (SMR)

We used summary-data-based Mendelian Randomization (SMR)[44] to identify GWAS signals that can be explained by cis-variants that moderate expression and splicing. We performed colocalization tests for 3,379 genes with cis-eQTL p-value < 5×10^{-5} for the top variant, and 2,439 splicing events with cis-sQTL p-value < 5×10^{-5} for the top variant in HCASMC against against the latest CARDIoGRAMplusC4D and UK Biobank GWAS meta-analysis[45]. We identified genome-wide significant eQTL and sQTL colocalizations based on their SMR p-values after controlling false discoveries (5% FDR, Benjamini-Hochberg). The equivalent p-value was 2.96×10^{-5} and 2.05×10^{-5} for eQTL and sQTL, respectively. SMR uses a reference population to determine linkage between variants; we used genetic data from individuals of European ancestry from 1000 Genomes as the reference population in our analyses. At loci that showed colocalization between HCASMC eQTLs and CAD GWAS associations, we repeated the colocalization tests using eQTLs from each of 44 GTEx tissues to identify shared and tissue-specific signals. To determine whether the reference panel would influence SMR result, we repeated SMR analysis using HCASMC genotypes as the reference. The results of a 33 genes were affected. For instance, *ADAMS7* had p-value < 0.41 now has p-value < 0.20. The results for most genes were not affected. In particular, *TCF21* and *FES* were still genome-wide significant.

14.2 Bayesian hierarchical colocaization test

We also used an orthogonal approach to perform colocalization testing. For every significant eGene, we then tested all variants within 500kb of the lead eQTL SNP for colocalization with CAD summary statistics from the latest CARDIoGRAMplusC4D and UK Biobank GWAS meta-analysis[45]. At each candidate locus (eGene), we ran FINEMAP[46] twice to compute the posterior probability that each individual SNP at the locus was a causal SNP for 1) the GWAS phenotype (CAD), and 2) HCASMC eQTL. We then processed the FINEMAP results to compute a colocalization posterior probability (CLPP) using the method described by Hormozdiari *et al.*[47]. Intuitively, the CLPP score represents the probability that the GWAS and HCASMC eQTL have the same causal variant at the locus, given the assumption that the GWAS trait and eGene each have exactly one causal SNP at the locus. In Hormozdiari *et al.*, the author recommend using CLPP > 0.01 as the cutoff for colocalization. In this study, we used a more conservative cutoff and defined a GWAS and e/sQTL pair to colocalize if CLPP > 0.05. At loci that showed colocalization between HCASMC eQTLs and CAD GWAS associations, we repeated the colocalization tests using eQTLs from each of 44 GTEx tissues. We compared the resulting CLPP scores to identify the full set of eQTL tissues with which each GWAS locus exhibited a colocalized signal.

14.3 Direction of effect determination

We determined the direction of effect, i.e. whether gene upregulation increase risk, using the correlation of effect sizes in the CARDIoGRAMplusC4D and UK Biobank GWAS meta-analysis[45] and the eQTL studies (Fig. 4). We first merged two datasets by rsIDs, correcting for any differences in reference and alternative allele designations. We subset to all SNPs with p-value < 1e-3 in both the GWAS and eQTL datasets (since other SNPs carry mostly noise). We fit a regression through the point (0.5, 0) using the eQTL effect sizes (allelic ratio) as predictor and the GWAS effect sizes as the response, and determined the direction of effect as the sign of the slope. We used the point (0.5, 0) because an allelic ratio of 0.5 indicates equal expression from both alleles (no eQTL effect), and a log odds-ratio of 0 indicates no GWAS effect. In Fig. 4, upward arrows indicate that upregulation of gene expression increases disease risk, and downward arrows indicate that upregulation of gene expression decreases risk.

Supplementary References

1. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2002). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline (Hoboken, NJ, USA: John Wiley & Sons, Inc.).

2. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491–498.

3. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17, pp.10–pp.12.

4. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

5. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997.

6. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

7. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet 48, 206–213.

8. Browning, B.L., and Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am. J. Hum. Genet. 85, 847–861.

9. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68–74.

10. 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.

11. 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.

12. 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 Meth 12, 1061–1063.

13. Deluca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 28, 1530–1532.

14. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774.

15. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and

Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet 50, 151–158.

16. Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol 109, 21.29.1–.29.9.

17. Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Research 44, D726–D732.

18. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357–359.

19. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, R137.

20. Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics 5, 1752–1779.

21. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Research 42, W187–W191.

22. Consortium, G., analysts, L., Laboratory, Data Analysis & Coordinating Center (LDACC):, management, N.P., collection, B., Pathology, group, E.M.W., Battle, A., Brown, C.D., Engelhardt, B.E., et al. Genetic effects on gene expression across human tissues. Nature 550, 204–213.

23. Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 1–27.

24. 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.

25. Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3, 1724–1735.

26. Tarazona, S., Garcia, F., Ferrer, A., Dopazo, J., and Conesa, A. (2012). NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. EMBnet.Journal 17, 18.

27. Mele, M., Ferreira, P.G., Reverter, F., Deluca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al. (2015). The human transcriptome across tissues and individuals. Science 348, 660–665.

28. Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software 36, 1–48.

29. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A Practical and powerful approach to multiple testing. J. Roy. Statist. Soc. 57, 289–300.

30. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

31. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription

Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell 38, 576–589.

32. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research 46, D260–D266.

33. 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. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 47, 1228–1235.

34. Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science 352, 600–604.

35. Consortium, T.C. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet 47, 1121–1130.

36. Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., and Willer, C.J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics 31, 2601–2606.

37. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328.

38. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaSci 4, 559.

39. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature Protocols 7, 500–507.

40. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics 32, 1479–1485.

41. Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. PNAS 100, 9440–9445.

42. Peterson, C.B., Bogomolov, M., Benjamini, Y., and Sabatti, C. (2016). TreeQTL: hierarchical error control for eQTL findings. Bioinformatics 15, 2556-8.

43. Han, B., and Eskin, E. (2012). Interpreting Meta-Analyses of Genome-Wide Association Studies. PLoS Genet. 8, e1002555.

44. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 48, 481–487.

45. Nelson, C.P., Goel, A., Butterworth, A.S., Kanoni, S., Webb, T.R., Marouli, E., Zeng, L., Ntalla, I., Lai, F.Y., Hopewell, J.C., et al. (2017). Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat Genet 49, 1385–1391.

46. Benner, C., Spencer, C.C.A., Havulinna, A.S., Salomaa, V., Ripatti, S., and Pirinen, M.

(2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493–1501.

47. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. The American Journal of Human Genetics 99, 1245–1260.