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[Δ], Daniel J. Rader[Δ], Susannah Elwyn[Δ],Erik Ingelsson[@], Stephen B. Montgomery[*#$], Clint L. Miller[*&], and Thomas Quertermous[*@]

Stanford University Dept. of [+]Biology, [#]Genetics, [$]Pathology, [^]Bioengineering, [@]Medicine, and the Cardiovascular Institute, Stanford School of Medicine, 300 Pasteur Drive, Stanford, CA 94305;
[%]Biomedical Informatics Training Program, Stanford School of Medicine, 300 Pasteur Drive, Stanford, CA 94305;
[Δ]University of Pennsylvania, Philadelphia, PA; and [&]Center for Public Health Genomics, Department of Public Health Sciences, Biochemistry and Genetics, and Biomedical Engineering, University of Virginia, Charlottesville, VA 22908

*These authors contributed equally to this work

Materials and Correspondence

Thomas Quertermous
300 Pasteur Dr.
Falk CVRC
Stanford, CA 94305
tomq1@stanford.edu
Tel: 650-723-5012
Fax: 650-725-2178

Abstract

Coronary artery disease (CAD) is the leading cause of death globally. Genome-wide association studies (GWAS) have identified more than 95 independent loci that influence CAD risk, most of which reside in non-coding regions of the genome. To interpret these loci, we generated transcriptome and whole-genome datasets using human coronary artery smooth muscle cells (HCASMC) from 52 unrelated donors, as well as epigenomic datasets using ATAC-seq on a subset of 8 donors. Through systematic comparison with publicly available datasets from GTEx and ENCODE projects, we identified transcriptomic, epigenetic, and genetic regulatory mechanisms specific to HCASMC. We assessed the relevance of HCASMC to CAD risk using transcriptomic and epigenomic level analyses. By jointly modeling eQTL and GWAS datasets, we identified five genes (*SIPA1*, *TCF21*, *SMAD3*, *FES*, and *PDGFRA*) that modulate CAD risk through HCASMC, all of which have relevant functional roles in vascular remodeling. Comparison with GTEx data suggests that *SIPA1* and *PDGFRA* influence CAD risk predominantly through HCASMC, while other annotated genes may have multiple cell and tissue targets. Together, these results provide new tissue-specific and mechanistic insights into the regulation of a critical vascular cell type associated with CAD in human populations.

Introduction

Atherosclerotic coronary artery disease (CAD) is the leading cause of death in both developed and developing countries worldwide, and current estimates predict that more than 1 million individuals will suffer from new and recurrent CAD this year in the U.S. alone[1]. Like most polygenic diseases, both genetic and environmental factors influence an individual's lifetime risk for CAD[2]. Early Swedish twin studies and more recent genome-wide association studies (GWAS) have estimated that about 50% of CAD risk is explained by genetic factors[3,4]. To date, GWAS have reported more than 95 replicated independent loci and numerous additional loci that are associated at an FDR<0.05[5-8]. A majority of these loci reside in non-coding genomic regions and are expected to function through regulatory mechanisms. Also, approximately 75% of CAD loci are not associated with classical risk factors, suggesting that at least part of them function through mechanisms intrinsic to the vessel wall.

Smooth muscle cells (SMC) constitute the majority of cells in the coronary artery wall. In response to vascular injury (e.g. lipid accumulation, inflammation), SMCs undergo phenotypic switching, and ultimately contribute to both atherosclerotic plaque formation and stabilization[9-12]. Recent lineage tracing studies in mice have revealed that although 80% of plaque-derived cells lack traditional SMC markers, roughly half are of SMC origin[13,14]. Thus, genetic studies of human coronary artery smooth muscle cells (HCASMC) have the potential to shed new light on their diverse functions in the vessel wall relevant to human atherosclerosis. In a few cases, the underlying mechanisms have been identified for CAD loci in vascular SMC models[10,15-18]. Large-scale expression quantitative trait loci (eQTL) mapping efforts such as the Genotype Tissue Expression (GTEx) project have helped refine these mechanisms for multiple traits across human tissues[19]. However, due to the lack of HCASMC in both GTEx and other studies, the overall contribution of this cell type towards heritable CAD risk remains unknown.

Herein, we performed whole-genome sequencing and transcriptomic profiling of 52 HCASMC donors to quantify the effects of *cis*-acting variation on gene expression and splicing associated with CAD. We evaluated the tissue specificity and disease relevance of our findings in HCASMC by comparing to publicly available GTEx and ENCODE datasets. We observed significant colocalization of eQTL and GWAS signals for five genes (*FES*, *SMAD3*, *TCF21*, *PDGFRA* and *SIPA1*), which all have the capacity to perform relevant functions in vascular remodeling. Further, comparative analyses with GTEx datasets reveals that *SIPA1* and *PDGFRA* act primarily in HCASMC. Together, these findings demonstrate the power of leveraging genetics of gene regulation for a critical cell type to uncover new risk-associated mechanisms for CAD.

Material and Methods

Sample acquisition and cell culture. A total of 62 primary human coronary artery smooth muscle cell (HCASMC) lines collected from donor hearts were purchased, and 52 lines remained after stringent filtering (see **Supplementary Note**). These 52 lines were from PromoCell (catalog # C-12511, n = 19), Cell Applications (catalog # 350-05a, n = 25), Lonza (catalog # CC-2583, n = 3), Lifeline Cell Technology (catalog # FC-0031, n = 3), and ATCC (catalog # PCS-100-021, n = 2). All lines were stained with smooth muscle alpha actin to check for smooth muscle content and all lines tested negative for mycoplasma (Table S1). 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. All HCASMC lines were expanded to passage 5-6 prior to extraction.

Library preparation and sequencing. Whole Genome Sequencing: Genomic DNA was isolated using Qiagen DNeasy Blood & Tissue Kit (catalog # 69506) and quantified using NanoDrop 1000 Spectrophotometer (Thermo Fisher). Macrogen performed library preparation using Illumina's TruSeq DNA PCR-Free Library Preparation Kit and 150 bp paired-end sequencing on Illumina HiSeq X Ten System. RNA Sequencing: RNA was extracted using Qiagen miRNeasy Mini Prep Kit (catalog # 74106). Quality of RNA was assessed on the Agilent 2100 Bioanalyzer. Samples with 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 made using Illumina TruSeq Stranded Total RNA Library Prep Kit (catalog # 20020597) and sequenced using 125bp paired-end on HiSeq 2500 Platform. ATAC Sequencing: We used ATAC-seq to assess chromatin accessibility with slight modifications to the published protocol[57]. Approximately $5x10^4$ fresh cells were collected at 500 g, washed in PBS, and nuclei extracted with cold lysis buffer. Pellets were subjected to transposition containing Tn5 transposases (Illumina) at 37 °C for 30 min, followed by purification using the DNA Clean-up and Concentration kit (Zymo). Libraries were PCR amplified using Nextera barcodes, with the total number of cycles empirically determined using SYBR qPCR. Amplified libraries were purified and quantified using bioanalyzer, nanodrop and qPCR (KAPA) analysis. Libraries were multiplexed and 2x75 bp sequencing was performed using an Illumina NextSeq 500.

Alignment and quantification of genomic, transcriptomic and epigenomic features. Whole-genome sequencing data were processed with the GATK best practices pipeline with hg19 as the reference genome[20,58], and VCF records were phased with Beagle v4.1[59]. Variants with imputation allelic r^2 less than 0.8 and Hardy-Weinberg Equilibrium p-value less than $1x10^{-6}$ were filtered out (see **Supplementary Note**). De-multiplexed FASTQ files were mapped with STAR version 2.4.0i in 2-pass mode[60] over the hg19 reference genome. Prior to expression quantification, we filtered our reads prone to mapping bias using WASP[61]. Total read counts and RPKM were calculated with RNA-SeQC v1.1.8[62] using default parameters with additional flags

"-n 1000 -noDoC -strictMode" over GENCODE v19 reference. Allele-specific read counts were generated with the createASVCF module in RASQUAL[27]. We quantified intron excision levels using LeafCutter intron-spanning reads[63]. 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, which requires at least 30 reads supporting each intron and allows intron to have a maximum size of 100kb. We used the ENCODE ATAC-seq pipeline to perform alignment and peak calling (https://github.com/kundajelab/atac_dnase_pipelines)[64]. FASTQ files were trimmed with Cutadapt v1.9[65] and aligned with Bowtie2 v2.2.6[66]. MACS2 v2.0.8[67] was used to call peaks with default parameters. Irreproducible Discovery Rate (IDR)[68] analyses were performed based on pseudo-replicates (subsample of reads) with a cutoff of 0.1 to output an IDR call set, which was used for downstream analysis. We used WASP[61] to filter out reads that are prone to mapping bias.

Mapping of cis-acting quantitative trait loci (QTL). Prior to QTL mapping, we inferred ancestry principal components (PCs) using the R package SNPRelate[69] on a pruned SNP set (Fig. S4). We filtered out SNPs based on Hardy-Weinberg equilibrium (HWE < $1x10^{-6}$), LD (r^2 < 0.2) and minor allele frequency (MAF < 0.05)[69]. To correct for hidden confounders, we extracted 15 covariates using PEER[70] on quantile normalized and rank-based inverse normal transformed RPKM values. The number of hidden confounders to removed was determined by empirically maximize the power to discover eQTLs on chromosome 20 (for computational speed and to avoid overfitting). We tested combinations of 3 to 5 genotype principal components with 1 to 15 PEER factors. We found that the combination of 4 genotype PCs with 8 PEER factors provides the most power to detect eQTLs. We then used sex, the top four genotype principal components, and the top eight PEER factors in both FastQTL and RASQUAL to map cis-eQTL with a 2Mb window centered at transcription start sites. Mathematically, the model is the following:

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

Where *e* stands for gene expression, and *g* stands for the genotype of the test SNP. We used LeafCutter[31] to quantify intron excision levels, and FastQTL[26] to map cis-sQTLs within a 200 kbp window around splice donor sites, controlling for sex, genotype PCs, and splicing PCs. Using a similar approach, we found that 3 genotype PCs and 6 splicing PCs maximized the power to map sQTLs. To control for multiple hypothesis testing, we calculated per-gene eQTL p-values using FastQTL with permutation, and controlled transcriptome-wide false

discovery rate with the q-value package[71]. For RASQUAL, it was not computationally feasible to perform gene-level permutation testing. Instead, we used TreeQTL to simultaneously control for SNP-level FDR and gene-level FDR[72]. Note that TreeQTL is more conservative than permutation {Consortium:jn}.

Quantifying tissue- and cell-type specific contribution to coronary artery disease (CAD) risk

We used stratified LD score regression[33] to estimate the enrichment of heritability for SNPs around tissue- and cell-type specific genes as described previously[32]. We defined tissue-specific genes by first selecting for independent tissues and removing tissues primarily composed of smooth muscle to avoid correlation with HCASMC (see **Supplementary note**). After filtering, 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 defined tissue-specific genes using gene expression z-score. For each gene, we determined the mean and standard deviation of median RPKM across tissues, from which the z-score is derived.

$$\tilde{e}_t = median(\boldsymbol{e_t})$$

$$z = (\tilde{e}_t - E(\tilde{e}_t))/Var(\tilde{e}_t)$$

Where $\boldsymbol{e_t}$ is the RPKM across all individuals in tissue *t*. We ranked each gene based on the z-scores (a higher z-score indicates more tissue specificity), and defined tissue-specific genes as the top 1000, 2000, and 4000 genes. A given SNP was assigned to a gene if it fell into the union of exon +/- 1kbp of that gene. 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[73]. To determine whether CAD risk variants are enriched in the open chromatin regions tissue- and cell-type specific fashion, we used a modified version of GREGOR[34] to estimate the likelihood of observing given number of GWAS variants falling into open chromatin regions of each tissue and cell type (see **Supplementary Note**). We first defined a GWAS locus as all variants in LD (r^2>0.7) with the lead variant. Given a set of GWAS loci, 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 p-values and odds ratios between GWAS variants and background variants across HCASMC and all ENCODE tissues and primary cell lines.

Colocalization between molecular QTL and CAD genome-wide association study (GWAS). We used summary-data-based Mendelian Randomization (SMR)[37] to determine GWAS loci that can be explained by *cis*-acting QTLs. We performed colocalization tests for 3,379 genes with cis-eQTL p-value < $5x10^{-5}$ for the top variant and 2,439 splicing events with cis-sQTL p-value < $5x10^{-5}$ for the top variant in HCASMC against the latest CARDIoGRAMplusC4D and UK Biobank GWAS meta-analysis[6]. We identified genome-wide significant eQTL and sQTL colocalizations based on adjusted SMR p-values (Benjamini-Hochberg FDR < 0.05). The equivalent p-value was $2.96x10^{-5}$ and $2.05x10^{-5}$ 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. We also used a modified version of eCAVIAR[35] to identify colocalized signals (see **Supplementary Note**). We calculated colocalization posterior probability (CLPP) using all SNPs within 500kb of the lead eQTL SNP against CAD summary statistics from CARDIoGRAMplusC4D and UK Biobank GWAS meta-analysis[6]. For computational feasibility, the GWAS and eQTL loci were assumed to have exactly one causal SNP. We defined colocalization events using CLPP > 0.05. Note that this is more conservative than the default eCAVIAR cutoff (CLPP > 0.01). We determined the direction of effect, namely whether gene upregulation increases risk, using the correlation of effect sizes in the GWAS and the eQTL studies. We selected SNPs with p-value < $1x10^{-3}$ in both the GWAS and eQTL datasets (since other SNPs carry mostly noise), and fitted a regression using the GWAS and eQTL effect sizes as the predictor and the response, respectively. We defined the direction of effect as the sign of the regression slope.

Results

HCASMC-specific genomic architecture

We obtained and cultured 62 primary HCASMC lines, and 52 lines remained for analysis after stringent quality control (**Supplementary Note** and Table S1). We performed whole-genome sequencing to an average depth of 30X, and jointly called genotypes using the GATK best practices pipeline[20], producing a total of ~15.2 million variants after quality control (see **Methods**). For RNA, we performed 125bp paired-end sequencing to a median depth of 51.3 million reads, with over 2.7 billion reads in total. After quantification and quality control, 19,607 genes were expressed in sufficient depth for downstream analysis (Table 1). To confirm that HCASMC

derived from tissue culture reflect *in vivo* physiology, we first projected their transcriptomes onto the 53 tissues profiled in GTEx[19] (Fig. 1A). Using multi-dimensional scaling (MDS) to visualize the similarity of HCASMC to GTEx tissues, we observed that HCASMC forms a distinct cluster and closely neighbors fibroblasts, skeletal muscle, arteries, heart and various smooth muscle-enriched tissues (vagina, colon, stomach, uterus and esophagus). These results were expected given that HCASMC are predicted to be similar to skeletal muscle, smooth muscle-enriched tissues as well as tissues representing the same anatomical compartment (e.g. heart and artery)[21]. In addition, HCASMC resemble fibroblast as both can be differentiated from mesenchymal cells from the dorsal mesocardium[22]. We also computed the epigenetic similarity between HCASMC and ENCODE cell types[23]. Consistent with the transcriptomic findings, the closest neighbors to HCASMC using epigenomic data were fibroblasts, heart, lung and skeletal muscle (Fig. 1B).

Next, we determined the pathways that may be selectively upregulated in HCASMC compared to closely related tissues. We performed differential expression analysis of HCASMC against fibroblasts and coronary artery in GTEx after correcting for batch effects and other hidden confounders (see **Methods**). Overall, 2,610 and 6,864 genes were found to be differentially expressed, respectively (FDR < $1x10^{-3}$, Fig. 1C and Fig. S1), affecting pathways involved in cellular proliferation, epithelial-mesenchymal transition (EMT) and extracellular matrix (ECM) secretion (Table S2). Next, we sought to identify HCASMC-specific epigenomic signatures by comparing HCASMC open chromatin profiles, as determined with ATAC-seq, against DNaseI hypersensitivity (DHS) sites across all ENCODE primary cell types and tissues (Table S3). We processed HCASMC ATAC-seq data with the ENCODE pipeline and standardized peaks as 75 bp around the peak summit for all tissues and cell lines to mitigate batch effect (see **Methods**). A total of 7332 peaks (2.1%) were not previously identified in ENCODE and represent HCASMC-specific sites (Fig. 1D). For example, an intronic peak within the *LMOD1* gene was found to be unique to HCASMC (Fig. 1E). This gene is expressed only in vascular and visceral smooth muscle cells where it is involved in actin polymerization, and has been mapped as a candidate causal CAD gene[11]. We then sought to identify transcription factor binding sites overrepresented in HCASMC-specific peaks. Motif enrichment analyses indicated that HCASMC-specific open chromatin sites are enriched with binding sites for members of the forkhead box (FOX) transcription factor family (see **Methods**). We performed motif enrichment analysis using 50-, 200-, and 1000-bp regions flanking HCASMC-specific peaks, and found

that the enrichment was robust to selection of window size, indicating the result is not simply due to selection bias (Fig. S2). The FOX transcription factors are known to regulate tissue- and cell-type specific gene transcription[24], and a subgroup of this family includes those with the ability to serve as pioneer factors[25]. To validate that FOX motif enrichment is specific to HCASMC, we performed similar analyses for brain-, heart-, and fibroblast-specific open chromatin regions and observed a depletion of FOX motifs (Fig. S3). Together these results suggest that HCASMC-specific transcriptomic and epigenomic profiles provide new regulatory mechanisms previously lacking in large publicly available datasets.

Expression and splicing quantitative trait locus discovery

In order to investigate the genetic regulatory mechanisms of gene expression in HCASMC, we conducted genome-wide mapping of eQTLs using both FastQTL[26] and RASQUAL[27] on the 52 donor samples from diverse ethnic backgrounds (Table S1 and Fig S4). RASQUAL has been previously shown to increase the *cis*-eQTL discovery power in small sample sizes by leveraging allele-specific information[27]. Indeed, using a threshold of FDR < 0.05, RASQUAL increased the number of eQTLs discovered approximately seven-fold as compared to FastQTL (RASQUAL:1220 vs. FastQTL:167, Table 1). We next evaluated whether these eQTLs were enriched in regions of open chromatin using data from a subset of individuals with ATAC-seq profiles. We observed that eQTLs within HCASMC open chromatin regions had more significant p-values compared to all eQTLs (Fig. S5, two-sided rank-sum test p-value < 9.2×10^{-5}). This is consistent with putative effects of *cis*-acting variation, potentially functioning through altered TF binding around these accessible regions. Next, using a Bayesian meta-analytic approach[28], we sought to identify HCASMC-specific eQTLs using GTEx tissues as a reference. Under the most stringent criteria (eQTL posterior probability > 0.9 for HCASMC and < 0.1 for all GTEx tissues, see **Methods**), we identified four HCASMC-specific eQTLs (Fig. S6). For example, rs1048709 is the top eQTL-SNP and confers HCASMC-specific regulatory effects on Complement Factor B (Fig. S6B), a gene that has been previously implicated in atherosclerosis and other inflammatory diseases[29]. In addition to regulatory effects on gene expression, previous studies have identified splicing as a major source of regulatory impact of genetic variation on complex diseases[30]. Therefore, we mapped splicing QTLs (sQTLs) using LeafCutter[31] and identified 581 sQTLs associated at FDR < 0.05 (Table 1). As a quality control, we estimated the enrichment of sQTLs and eQTLs against a matched set of background variants. As expected, eQTLs were enriched around

the 5' UTR (Fig. S7A), whereas sQTLs were enriched in splicing regions, particularly splice donor and acceptor sites (Fig. S7B).

Overall CAD genetic risk mediated by HCASMC

We next examined the heritable contribution of HCASMC towards the risk of CAD. Previous reports have suggested that disease-associated SNPs are often enriched in genes expressed in the relevant tissue types[32]. Thus, we estimated the contribution to CAD risk from SNPs in or near genes showing patterns of tissue-specific expression and identified the top 2000 tissue-specific genes for HCASMC and GTEx tissues (see **Methods**). We then applied stratified LD score regression[33] to estimate CAD heritability explained by SNPs within 1kb of tissue-specific genes. We found that HCASMC, along with coronary artery and adipose tissues, contribute substantially towards CAD heritability (Fig. 2A). These enrichment results were robust to the tissue-specificity cutoff (top 1000, 2000, or 4000 genes), suggesting that they were not simply due to selection bias (Fig. S8). Complementary epigenomic evidence previously demonstrated that risk variants for complex diseases are often enriched in open chromatin regions in relevant tissue types[23,33,34]. Thus, we estimated the degree of overlap between CAD variants and open chromatin in HCASMC and ENCODE cell types using a modified version of GREGOR[34] (see **Methods**). We observed that open chromatin regions in HCASMC, as well as vascular endothelial cells, monocytes, uterus (smooth muscle) and B-cells, are enriched for CAD risk variants (Fig. 2B). These findings support the role of HCASMC as an appropriate cellular model to map the genetic basis of CAD, which may be supplemented by the contribution of other vessel wall cell types.

Fine-mapping CAD risk variants

Whole-genome sequencing of our HCASMC population sample provides the opportunity to fine-map CAD risk loci. Several studies have used colocalization between GWAS and eQTL signals as a fine-mapping approach to identify candidate causal regulatory variants[35-38], and in several cases pinpointing single causal variants[39,40]. Given the global overlap between CAD risk variants and genetic regulation in HCASMC, we sought to find evidence for colocalization between GWAS and eQTL signals. We thus compiled publicly available genome-wide summary statistics from the latest meta-analysis[6]. We then applied two methods with different statistical assumptions, eQTL and GWAS CAusal Variants Identification in Associated Regions (eCAVIAR)[35] and

Summary-data-based Mendelian Randomization (SMR)[37] to identify colocalizing variants and genes across all CAD loci, and focused on the union of results from the two independent methods. We used FDR < 0.05 and colocalization posterior probability (CLPP) > 0.05 as cutoffs for SMR and eCAVIAR, respectively (Note that CLPP > 0.05 is more conservative than the CLPP > 0.01 recommended in the publication of the eCAVIAR method). From this approach, we identified five high-confidence genes, namely *FES*, *SMAD3*, *TCF21*, *PDGFRA* and *SIPA1* (Fig. 3). Although the top genes found by two methods differed, we observed that the SMR p-values and eCAVIAR CLPPs positively correlate (Fig. S9), and that two of the three genes found only by eCAVIAR achieved nominal significance in SMR (Table S4). We then investigated whether these colocalizations were unique to HCASMC by conducting colocalization tests across all GTEx tissues. For *SIPA1* and *PDGFRA*, colocalization appears to be HCASMC-specific (Fig. 3G; Fig. S10A and S10D). For *SMAD3*, both HCASMC and thyroid have strong colocalization signals (Fig. S10B). *TCF21* and *FES* colocalization were found to be shared across multiple tissues (Fig. S10C and Fig. S11D). Next, we conducted colocalization analysis between sQTL and GWAS summary statistics with both eCAVIAR and SMR. We identified colocalization with four genes (Table S4, and Fig. S12). The most significant colocalization event is at the *SMG9* locus. Interestingly, the top sQTL variant, rs4760, is a coding variant located in the exon of the *PLAUR* (plasminogen activator urokinase receptor) gene and is also a GWAS variant for circulating cytokines and multiple immune cell traits[41,42]. By correlating eQTL and GWAS effect sizes, we observed that increased *TCF21* and *FES* expression levels are associated with reduced CAD risk, while increased *PDGFRA*, *SIPA1*, and *SMAD3* expression levels are associated with increased CAD risk (Fig. 4A-E). These results provide genetic evidence that pathways promoting SMC phenotypic transition during atherosclerosis can be both protective and detrimental depending on the genes implicated (Fig. 4F).

Discussion

In this study, we have integrated genomic, transcriptomic, and epigenetic datasets to create the first map of genetic regulation of gene expression in human coronary artery smooth muscle cells. Comparison with publicly available transcriptomic and epigenomic datasets in GTEx and ENCODE revealed regulatory patterns specific to HCASMC. By comparing against neighboring tissues in GTEx, we found thousands of differentially

expressed genes, which were enriched in pathways such as EMT, protein secretion and cellular proliferation, consistent with our current understanding of HCASMC physiology *in vivo*. In comparison with ENCODE, we found 7332 (~2.1%) open chromatin peaks unique to HCASMC, and showed that these peaks are enriched with binding motifs for Forkhead box family proteins, which are known to regulate cell-type-specific gene expression[43]. FOXP1 in particular has been shown to increase collagen production in smooth muscle cells[44], supporting a potential role in extracellular matrix remodeling in the vessel wall.

Using both transcriptomic and epigenomic profiles, we established that HCASMC represent an important cell type for coronary artery disease. On a tissue-level, we demonstrated that genes highly expressed in HCASMC, coronary artery and adipose tissue are enriched for SNPs associated with CAD risk. While the proximal aortic wall is also susceptible to atherosclerosis, the coronary arteries represent the primary origin of ischemic coronary artery disease in humans[9]. Given that the majority of coronary arteries in the epicardium are encapsulated by perivascular adipose tissue in individuals with disease, one would expect these tissues to share gene responses involved in both vascular inflammation and lipid homeostasis[45]. Further, we demonstrated that HCASMC, endothelial cells, and immune cells also contribute towards the genetic risk of coronary artery disease. Recent –omic profiling of human aortic endothelial cells (HAECs) isolated from various donors identified a number of genetic variants and transcriptional networks mediating responses to oxidized phospholipids and pro-inflammatory stimuli[46]. Likewise, systems approaches investigating resident macrophages and other immune cells involved in vessel inflammation have provided additional insights into context-specific disease mechanisms[47,48].

Our integrative analyses identified a number of CAD-associated genes that may offer clues into potentially targetable HCASMC-mediated disease mechanisms. Although two of these associated genes, *TCF21* and *SMAD3*, have established roles in regulating vascular remodeling and inflammation during disease[12,16,49], the other identified genes, *PDGFRA*, *FES* and *SIPA1*, appear to be novel SMC associated genes. While the role for *PDGFRB* mediated signaling has been well documented in atherosclerosis and modulation of SMC phenotype, the possible involvement of *PDGFRA* has not been investigated in detail[50,51]. Interestingly, *FES* and *SIPA1* were found to harbor CpGs identified in current smokers in the Rotterdam Study, based on targeted

methylation profiling of CAD loci in whole blood[52]. The two identified CpGs in *FES* were located near the transcription start site, while the one CpG identified in *SIPA1* was located in the 5'-UTR, suggesting potential environmental influences on gene expression levels. *SIPA1* encodes a mitogen induced GTPase activating protein (GAP), specifically activating Ras and Rap GTPases[53]. *SIPA1* may be a unique mitogen response signal in HCASMC undergoing phenotypic transition in the injured vessel wall; however, these hypotheses should be explored in relevant functional models. Another HCASMC eQTL variant, rs2327429, located in the *TCF21* promoter region, was also the lead SNP in this locus in a recent CAD meta-analysis and has been identified as an mQTL for *TCF21* expression in two separate studies[54,55]. These data suggest that regulation of methylation is a novel molecular trait that may mediate risk for CAD. Splicing QTL colocalization analysis reveals that alternative splicing in *SMG9* also influences CAD risk. *SMG9* has been shown to regulate the non-sense mediated decay (NMD) pathway in human cells, and has been implicated in several developmental disorders such as brain malformations and congenital heart disease[56].

In summary, the current study confirms the value of detailed genomic and genetic analyses of disease-related tissues and cell types, which when analyzed in the context of publicly available data can provide deep insights into the physiology of human traits and pathophysiology of complex human disease. We expect that these findings will provide a rich resource for the community and prompt detailed functional investigations of candidate loci for preclinical development.

Supplemental Data

Supplemental Data include fifteen figures and five tables.

Acknowledgement

We thank Professor Nicolas Mermod for providing biological material and Normal Cyr for making illustrations. B.L. is supported in part by the Stanford Center for Computational, Evolutionary and Human Genetics Fellowship. T.Q is supported by NIH grants R01HL109512(NIH), R01HL134817 (NIH), R33HL120757 (NIH), R01DK107437 (NIH), and R01HL139478 (NIH). C.L.M is supported by R00HL125912 (NIH). S.B.M. is supported by R33HL120757 (NHLBI), U01HG009431 (NHGRI; ENCODE4), R01MH101814 (NIH Common

Fund; GTEx Program), R01HG008150 (NHGRI; Non-Coding Variants Program) and the Edward Mallinckrodt Jr. Foundation.

Declaration of Interests

The authors declare no competing financial interests.

Web Resources

Data and code availability. RNA sequencing data has been deposited at Gene Expression Omnibus (GEO), accession number GSE113348. All eQTL and sQTL summary statistics are accessible through the website http://montgomerylab.stanford.edu/resources.html. All code used to perform analyses and generate figures are in the GitHub repository: https://github.com/boxiangliu/hcasmc_eqtl

URLs. GATK, https://software.broadinstitute.org/gatk/; BWA, https://github.com/lh3/bwa; STAR, https://github.com/alexdobin/STAR; Picard, https://broadinstitute.github.io/picard/; RASQUAL, https://github.com/natsuhiko/rasqual; Beagle, https://faculty.washington.edu/browning/beagle/beagle.html; 1000 Genomes, http://www.internationalgenome.org/; VerifyBamID, https://genome.sph.umich.edu/wiki/VerifyBamID; FastQC, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/; WASP, https://github.com/bmvdgeijn/WASP; RNA-SeQC, http://archive.broadinstitute.org/cancer/cga/rna-seqc; GENCODE, https://www.gencodegenes.org/; ENCODE, https://www.encodeproject.org/; Bowtie2, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml; MACS2, https://github.com/taoliu/MACS; ENCODE ATAC-seq/DNAse-seq pipeline, https://github.com/kundajelab/atac_dnase_pipelines; DESeq2, https://bioconductor.org/packages/release/bioc/html/DESeq2.html; sva, https://bioconductor.org/packages/release/bioc/html/sva.html; NOISeq, https://bioconductor.org/packages/release/bioc/html/NOISeq.html; bedtools, http://bedtools.readthedocs.io/en/latest/; JASPAR, http://jaspar.genereg.net/; LD score regression, https://github.com/bulik/ldsc; LeafCutter, https://github.com/davidaknowles/leafcutter; GREGOR, https://genome.sph.umich.edu/wiki/GREGOR; PLINK, https://www.cog-genomics.org/plink2; FastQTL,

http://fastqtl.sourceforge.net/; TreeQTL, http://www.bioinformatics.org/treeqtl/; METASOFT, http://genetics.cs.ucla.edu/meta/; SMR, http://cnsgenomics.com/software/smr/#Overview; eCAVIAR, http://zarlab.cs.ucla.edu/tag/ecaviar/; FINEMAP: http://www.christianbenner.com/#.

Reference

1. Benjamin, E.J., Blaha, M.J., Chiuve, S.E., Cushman, M., Das, S.R., Deo, R., de Ferranti, S.D., Floyd, J., Fornage, M., Gillespie, C., et al. (2017). Heart Disease and Stroke Statistics—2017 Update: A Report From the American Heart Association. Circulation 135, CIR.0000000000000485–e603.

2. Khera, A.V., Emdin, C.A., Drake, I., Natarajan, P., Bick, A.G., Cook, N.R., Chasman, D.I., Baber, U., Mehran, R., Rader, D.J., et al. (2016). Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. N. Engl. J. Med. 375, 2349–2358.

3. Won, H.-H., Natarajan, P., Dobbyn, A., Jordan, D.M., Roussos, P., Lage, K., Raychaudhuri, S., Stahl, E., and Do, R. (2015). Disproportionate Contributions of Select Genomic Compartments and Cell Types to Genetic Risk for Coronary Artery Disease. PLoS Genet. 11, e1005622.

4. Zdravkovic, S., Wienke, A., Pedersen, N.L., Marenberg, M.E., Yashin, A.I., and De Faire, U. (2002). Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. J. Intern. Med. 252, 247–254.

5. Howson, J.M.M., Zhao, W., Barnes, D.R., Ho, W.-K., Young, R., Paul, D.S., Waite, L.L., Freitag, D.F., Fauman, E.B., Salfati, E.L., et al. (2017). Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. Nat Genet 385, 117.

6. 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.

7. Klarin, D., Zhu, Q.M., Emdin, C.A., Chaffin, M., Horner, S., McMillan, B.J., Leed, A., Weale, M.E., Spencer, C.C.A., Aguet, F., et al. (2017). Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. Nat Genet 388, 1459.

8. van der Harst, P., and Verweij, N. (2017). The Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circulation Research 122, CIRCRESAHA.117.312086–CIRCRESAHA.117.312443.

9. Khera, A.V., and Kathiresan, S. (2017). Genetics of coronary artery disease: discovery, biology and clinical translation. Nature Reviews Genetics 18, 331–344.

10. Pu, X., Xiao, Q., Kiechl, S., Chan, K., Ng, F.L., Gor, S., Poston, R.N., Fang, C., Patel, A., Senver, E.C., et al. (2013). ADAMTS7 Cleavage and Vascular Smooth Muscle Cell Migration Is Affected by a Coronary-Artery-Disease-Associated Variant. The American Journal of Human Genetics 92, 366–374.

11. Miller, C.L., Pjanic, M., Wang, T., Nguyen, T., Cohain, A., Lee, J.D., Perisic, L., Hedin, U., Kundu, R.K., Majmudar, D., et al. (2016). Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. Nature Communications 7, 12092.

12. Braitsch, C.M., Combs, M.D., Quaggin, S.E., and Yutzey, K.E. (2012). Pod1/Tcf21 is regulated by retinoic acid signaling and inhibits differentiation of epicardium-derived cells into smooth muscle in the developing he...

- PubMed - NCBI. Developmental Biology 368, 345–357.

13. Shankman, L.S., Gomez, D., Cherepanova, O.A., Salmon, M., Alencar, G.F., Haskins, R.M., Swiatlowska, P., Newman, A.A.C., Greene, E.S., Straub, A.C., et al. (2015). KLF4-dependent phenotypic modulation of smooth muscle cells has a key role in atherosclerotic plaque pathogenesis. Nature Medicine 21, 628–637.

14. Cherepanova, O.A., Gomez, D., Shankman, L.S., Swiatlowska, P., Williams, J., Sarmento, O.F., Alencar, G.F., Hess, D.L., Bevard, M.H., Greene, E.S., et al. (2016). Activation of the pluripotency factor OCT4 in smooth muscle cells is atheroprotective. Nature Medicine 22, 657–665.

15. Nurnberg, S.T., Cheng, K., Raiesdana, A., Kundu, R., Miller, C.L., Kim, J.B., Arora, K., Carcamo-Oribe, I., Xiong, Y., Tellakula, N., et al. (2015). Coronary Artery Disease Associated Transcription Factor TCF21 Regulates Smooth Muscle Precursor Cells That Contribute to the Fibrous Cap. PLoS Genet. 11, e1005155.

16. Miller, C.L., Haas, U., Diaz, R., Leeper, N.J., Kundu, R.K., Patlolla, B., Assimes, T.L., Kaiser, F.J., Perisic, L., Hedin, U., et al. (2014). Coronary Heart Disease-Associated Variation in TCF21 Disrupts a miR-224 Binding Site and miRNA-Mediated Regulation. PLoS Genet. 10, e1004263.

17. Srivastava, R., Zhang, J., Go, G.-W., Narayanan, A., Nottoli, T.P., and Mani, A. (2015). Impaired LRP6-TCF7L2 Activity Enhances Smooth Muscle Cell Plasticity and Causes Coronary Artery Disease. Cell Reports 13, 746–759.

18. Kim, J.B., Pjanic, M., Nguyen, T., Miller, C.L., Iyer, D., Liu, B., Wang, T., Sazonova, O., Carcamo-Orive, I., Matic, L.P., et al. (2017). TCF21 and the environmental sensor aryl-hydrocarbon receptor cooperate to activate a pro-inflammatory gene expression program in coronary artery smooth muscle cells. PLoS Genet. 13, e1006750.

19. 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.

20. 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.).

21. Wang, G., Jacquet, L., Karamariti, E., and Xu, Q. (2015). Origin and differentiation of vascular smooth muscle cells. - PubMed - NCBI. J Physiol 593, 3013–3030.

22. Vrancken Peeters, M.P., Gittenberger-de Groot, A.C., Mentink, M.M., and Poelmann, R.E. (1999). Smooth muscle cells and fibroblasts of the coronary arteries derive from epithelial-mesenchymal transformation of the epicardium. Anat. Embryol. 199, 367–378.

23. Consortium, R.E., 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.

24. Li, S., Weidenfeld, J., and Morrisey, E.E. (2004). Transcriptional and DNA binding activity of the Foxp1/2/4 family is modulated by heterotypic and homotypic protein interactions. Mol. Cell. Biol. 24, 809–822.

25. Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J.A., Mahony, S., Pugh, B.F., Lee, D., Kaestner, K.H., and Zaret, K.S. (2016). The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. Molecular Cell 62, 79–91.

26. 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.

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

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

29. Hovland, A., Jonasson, L., Garred, P., Yndestad, A., Aukrust, P., Lappegård, K.T., Espevik, T., and Mollnes, T.E. (2015). The complement system and toll-like receptors as integrated players in the pathophysiology of atherosclerosis. Atherosclerosis 241, 480–494.

30. 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.

31. 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.

32. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169, 1177–1186.

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. 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.

35. 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.

36. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. 10, e1004383.

37. 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.

38. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. PLoS Genet. 6, e1000895.

39. Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. Http://Dx.Doi.org/10.1056/NEJMoa1502214 373, 895–907.

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

41. Ahola-Olli, A.V., Würtz, P., Havulinna, A.S., Aalto, K., Pitkänen, N., Lehtimäki, T., Kähönen, M., Lyytikäinen, L.-P., Raitoharju, E., Seppälä, I., et al. (2017). Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. - PubMed - NCBI. The American Journal of Human Genetics 100, 40–50.

42. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. - PubMed - NCBI. Cell 167, 1415–1429.e1419.

43. Golson, M.L., and Kaestner, K.H. (2016). Fox transcription factors: from development to disease. Development 143, 4558–4570.

44. Bot, P.T., Grundmann, S., Goumans, M.J., 2011 Forkhead box protein P1 as a downstream target of transforming growth factor-β induces collagen synthesis and correlates with a more stable plaque phenotype. Atherosclerosis.

45. Berg, A.H., and Scherer, P.E. (2005). Adipose Tissue, Inflammation, and Cardiovascular Disease. Circulation Research 96, 939–949.

46. Hogan, N.T., Whalen, M.B., Stolze, L.K., Hadeli, N.K., Lam, M.T., Springstead, J.R., Glass, C.K., and Romanoski, C.E. (2017). Transcriptional networks specifying homeostatic and inflammatory programs of gene expression in human aortic endothelial cells. eLife Sciences 6, e22536.

47. Ghattas, A., Griffiths, H.R., Devitt, A., Lip, G.Y.H., and Shantsila, E. (2013). Monocytes in Coronary Artery Disease and Atherosclerosis: Where Are We Now? Journal of the American College of Cardiology 62, 1541–1551.

48. Kinlay, S., Libby, P., and Ganz, P. (2001). Endothelial function and coronary artery disease. Curr. Opin. Lipidol. 12, 383–389.

49. Turner, A.W., 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.

50. Andrae, J., Gallini, R., and Betsholtz, C. (2008). Role of platelet-derived growth factors in physiology and medicine. Genes Dev. 22, 1276–1312.

51. He, C., Medley, S.C., Hu, T., Hinsdale, M.E., Lupu, F., Virmani, R., and Olson, L.E. (2015). PDGFRβ signalling regulates local inflammation and synergizes with hypercholesterolaemia to promote atherosclerosis. Nature Communications 6, 7770.

52. Steenaard, R.V., Ligthart, S., Stolk, L., Peters, M.J., van Meurs, J.B., Uitterlinden, A.G., Hofman, A., Franco, O.H., and Dehghan, A. (2015). Tobacco smoking is associated with methylation of genes related to coronary artery disease. - PubMed - NCBI. Clin Epigenet 7, 820.

53. Kurachi, H., Wada, Y., Tsukamoto, N., Maeda, M., Kubota, H., Hattori, M., Iwai, K., and Minato, N. (1997). Human SPA-1 Gene Product Selectively Expressed in Lymphoid Tissues Is a Specific GTPase-activating Protein for Rap1 and Rap2 SEGREGATE EXPRESSION PROFILES FROM A rap1GAP GENE PRODUCT. J. Biol. Chem. 272, 28081–28088.

54. Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., et al. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife Sciences 2, e00523.

55. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. PLoS Genet. 6, e1000952.

56. Shaheen, R., Anazi, S., Ben-Omran, T., Seidahmed, M.Z., Caddle, L.B., Palmer, K., Ali, R., Alshidi, T., Hagos, S., Goodwin, L., et al. (2016). Mutations in SMG9, Encoding an Essential Component of Nonsense-

Mediated Decay Machinery, Cause a Multiple Congenital Anomaly Syndrome in Humans and Mice. Am. J. Hum. Genet. 98, 643–652.

57. 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.

58. 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.

59. 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.

60. 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.

61. van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. (2014). WASP: allele-specific software for robust discovery of molecular quantitative trait loci. bioRxiv.

62. 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.

63. Li, Y.I., Knowles, D.A., and Pritchard, J.K. (2016). LeafCutter: Annotation-free quantification of RNA splicing. bioRxiv.

64. 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.

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

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

67. 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.

68. 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.

69. 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.

70. 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.

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

72. Peterson, C.B., Bogomolov, M., Benjamini, Y., and Sabatti, C. (2016). TreeQTL: hierarchical error control for eQTL findings. Bioinformatics btw198.

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

74. Schunkert, H., König, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., Preuss, M., Stewart, A.F.R., Barbalic, M., Gieger, C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 43, 333–338.

75. The CARDIoGRAMplusC4D Consortium, DIAGRAM Consortium, CARDIOGENICS Consortium, MuTHER Consortium, Wellcome Trust Case Control Consortium, Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., et al. (2012). Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet 45, 25–33.

76. Acharya, A., Baek, S.T., Huang, G., Eskiocak, B., Goetsch, S., Sung, C.Y., Banfi, S., Sauer, M.F., Olsen, G.S., Duffield, J.S., et al. (2012). The bHLH transcription factor Tcf21 is required for lineage-specific EMT of cardiac fibroblast progenitors. Development (Cambridge, England) 139, 2139–2149.

77. Greer, P. (2002). Closing in on the biological functions of fps/fes and fer. Nat Rev Mol Cell Biol 3, 278–289.

78. Hattori, M., Tsukamoto, N., Nur-e-Kamal, M.S., Rubinfeld, B., Iwai, K., Kubota, H., Maruta, H., and Minato, N. (1995). Molecular cloning of a novel mitogen-inducible nuclear protein with a Ran GTPase-activating domain that affects cell cycle progression. Mol. Cell. Biol. 15, 552–560.

79. Verrecchia, F., and Mauviel, A. (2002). Transforming Growth Factor-β Signaling Through the Smad Pathway: Role in Extracellular Matrix Gene Expression and Regulation. Journal of Investigative Dermatology 118, 211–215.

Figure Legends

Fig. 1 The relationship between HCASMC and GTEx and ENCODE cell and tissue types. (A) The multidimensional scaling plot of gene expression shows that HCASMC form a distinct cluster, which neighbors fibroblast, skeletal muscle, heart, blood vessel and various types of smooth muscle tissues such as esophagus and vagina (inset). **(B)** Jaccard similarity index between HCASMC and ENCODE cell and tissue types reveals that fibroblast, skeletal muscle, heart and lung are most closely related to HCASMC. **(C)** Thousands of genes are differentially expressed between HCASMC and its close neighbors, fibroblast, as well as the tissue of origin, coronary artery. **(D)** A total of 344284 open chromatin peaks are found in HCASMC, of which 7332 (2.1%) are HCASMC-specific. **(E)** An example of a HCASMC-specific peak located within the intron of *LMOD1*, which is an HCASMC-specific gene.

Fig. 2 Tissue- and cell-type specific contribution to CAD risk. (A) Tissue-specific enrichment of CAD heritability. We used stratified LD score regression to estimate the CAD risk explained by SNPs close to tissue-specific genes, defined as the 2000 genes with highest expression z-scores (see **Methods**). Genes whose

expression is specific to coronary artery, adipose, and HCASMC harbors SNPs with large effects on CAD. **(B)** Overlap between CAD risk variants and tissue- and cell-type specific open chromatin regions. We used a modified version of GREGOR (see **Methods**) to estimate the probability and odds ratio (compared with matched background SNPs) of overlap between CAD risk variants and open chromatin regions in HCASMC and across ENCODE tissues. HCASMC, arterial endothelial cells, monocytes, B cell, uterus (composed primarily of smooth muscle), and pons (possibly through regulation of blood pressure) showed the highest degrees of overlap.

Fig. 3 Colocalization between HCASMC eQTL and coronary artery disease GWAS. (A-C) Three potential causal genes identified by eCAVIAR. **(A)** Platelet-derived growth factor alpha (*PDFGRA*) eQTL signal colocalized with the *KDR* GWAS locus, which reached FDR < 0.05 significance in the latest CARDIoGRAMplusC4D and UK Biobank GWAS meta-analysis[6]. **(B)** Signal-Induced Proliferation-Associated 1 (*SIPA1*) eQTL signal colocalized with the *PCNX3* locus, which reached genome-wide significance (p-value < 9.71×10^{-9}) in Howson *et al.*[5] Note that this study only genotyped selected loci but have a larger sample size than the UK Biobank study. **(C)** *SMAD3* eQTL signal colocalized with the *SMAD3* locus, which is newly identified in the UK Biobank meta-analysis[6]. **(D)** Transcriptome-wide colocalization signals between HCASMC eQTL and CAD GWAS. We used eCAVIAR (top) and SMR (bottom) to fine-map GWAS causal variants and to identify eQTL signals that can explain CAD risk variants (see **Methods**). We found five genes whose eQTL signals show significant colocalization with CAD GWAS signal (SMR FDR < 0.05 or eCAVIAR colocalization posterior probability > 0.05). **(E-F)** Two potential causal genes identified by SMR. **(E)** Transcription factor 21 (*TCF21*) eQTL signal colocalized with the *TCF21* locus, which was identified by Schunkert *et al.*[74] and replicated in the UK Biobank meta-analysis. **(F)** *FES* eQTL signal colocalized with the *FURIN-FES* locus, which was identified by Deloukas *et al.*[75] and replicated in the UK Biobank meta-analysis. **(G)** *SIPA1* colocalization is strongest in HCASMC, indicating that this gene influences CAD risk primarily through HCASMC.

Fig. 4 Causal genes are involved in HCASMC-related vascular remodeling. (A-E) 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 < $1\text{x}10^{-3}$ in both datasets. Upregulation of genes

TCF21 and *FES* are protective against CAD risk, and upregulation of *SMAD3, PDGFRA,* and *SIPA1* increases CAD risk. **(F)** Hypothetical functions of five potential causal genes. Upregulation of *TCF21* facilitates the transition of smooth muscle cells from a contractile to a synthetic state[76]. Upon phenotypic transition, *FES* assists in smooth muscle cell migration to the neo-intima[77]. Both *SIPA1* and *PDGFRA* promotes HCASMC proliferation[50,78]. *SMAD3* induces synthetic smooth muscle re-differentiation into the synthetic phenotype for vessel wall repair[79]. Upward arrows indicate genetic upregulation increases CAD risk, and downward arrows indicate genetic upregulation is protective against CAD risk.

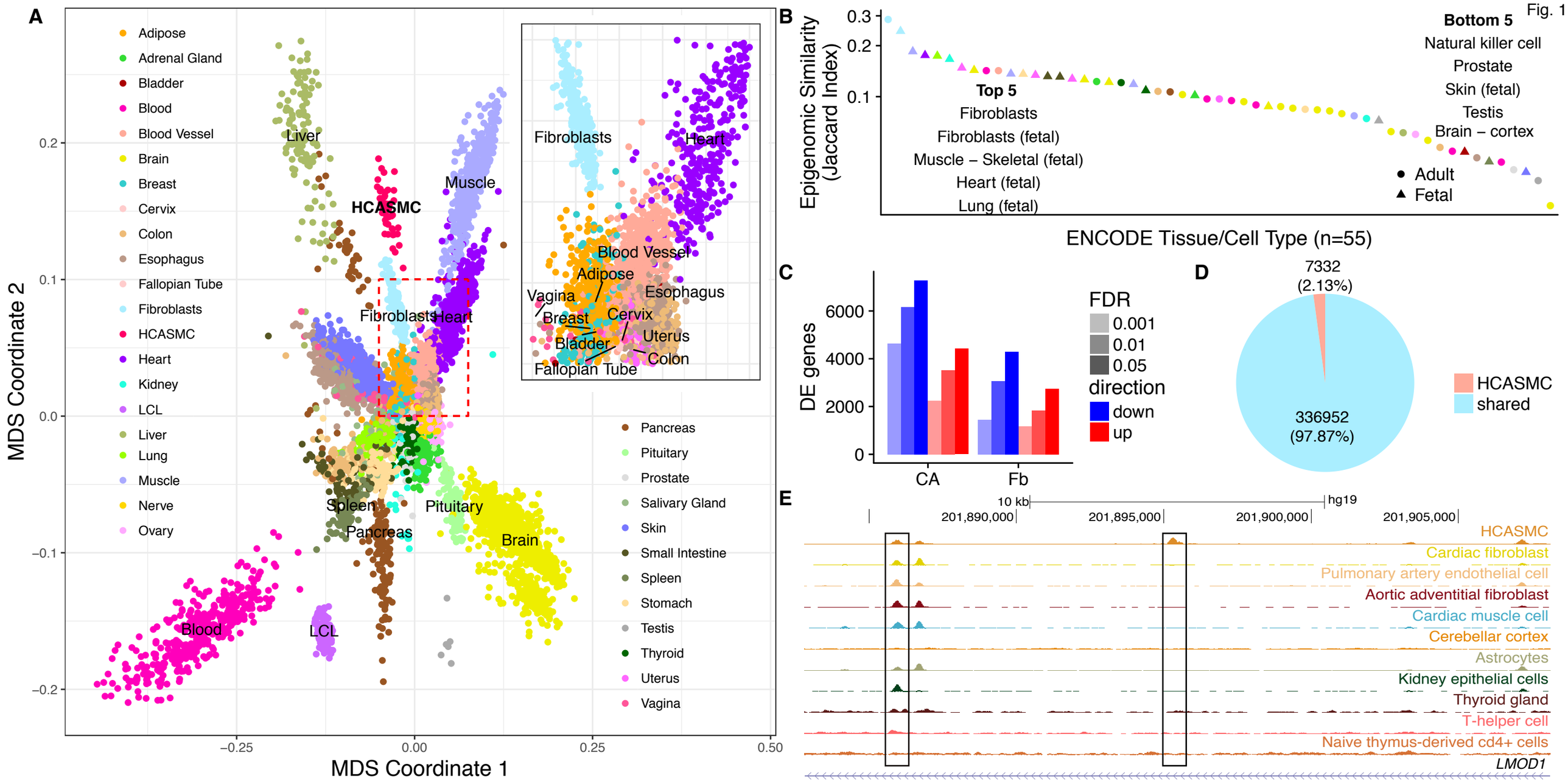
Table Legends

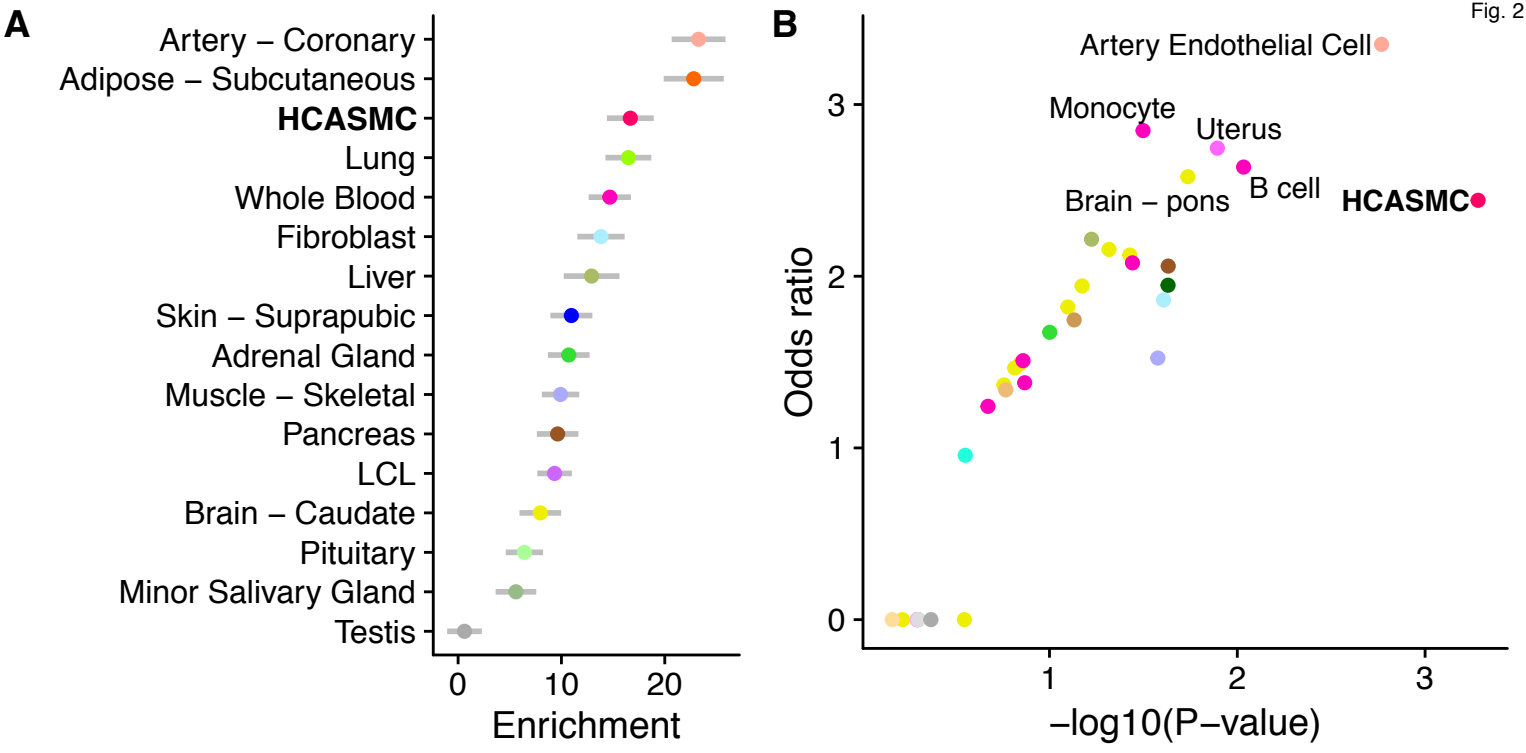
Table 1 Molecular quantitative trait loci discoveries. We report the number of tests performed and the number of significant loci at FDR < 0.05, 0.01, and 0.001 for eQTL and sQTL stratified by molecular trait type. We used permutation and the Benjamini-Hochberg adjustment for sQTL discovery, and a multi-level FDR correction procedure (TreeQTL[72]) for eQTL discovery, where permutation was not computationally feasible (see **Methods**).

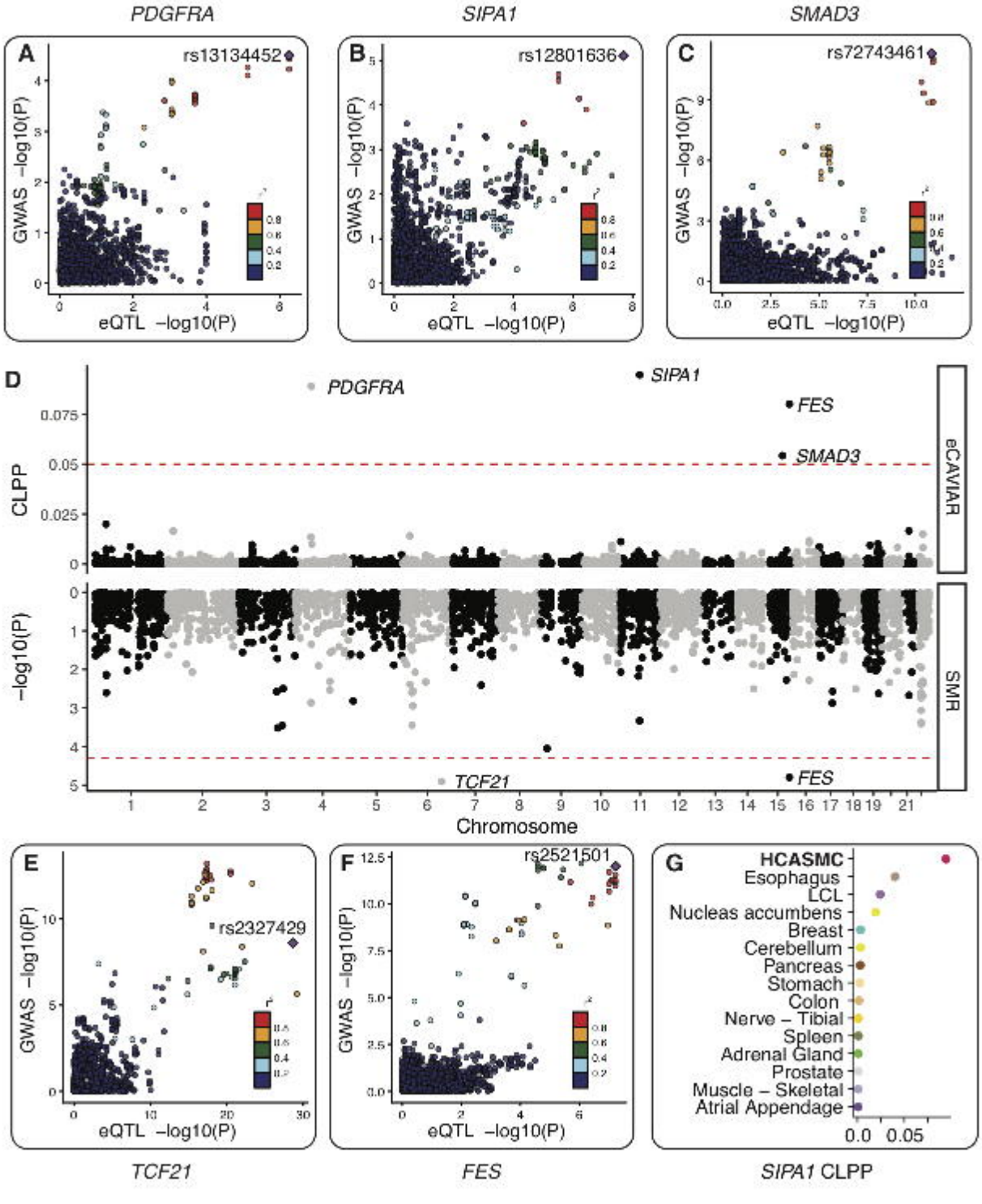
Tables

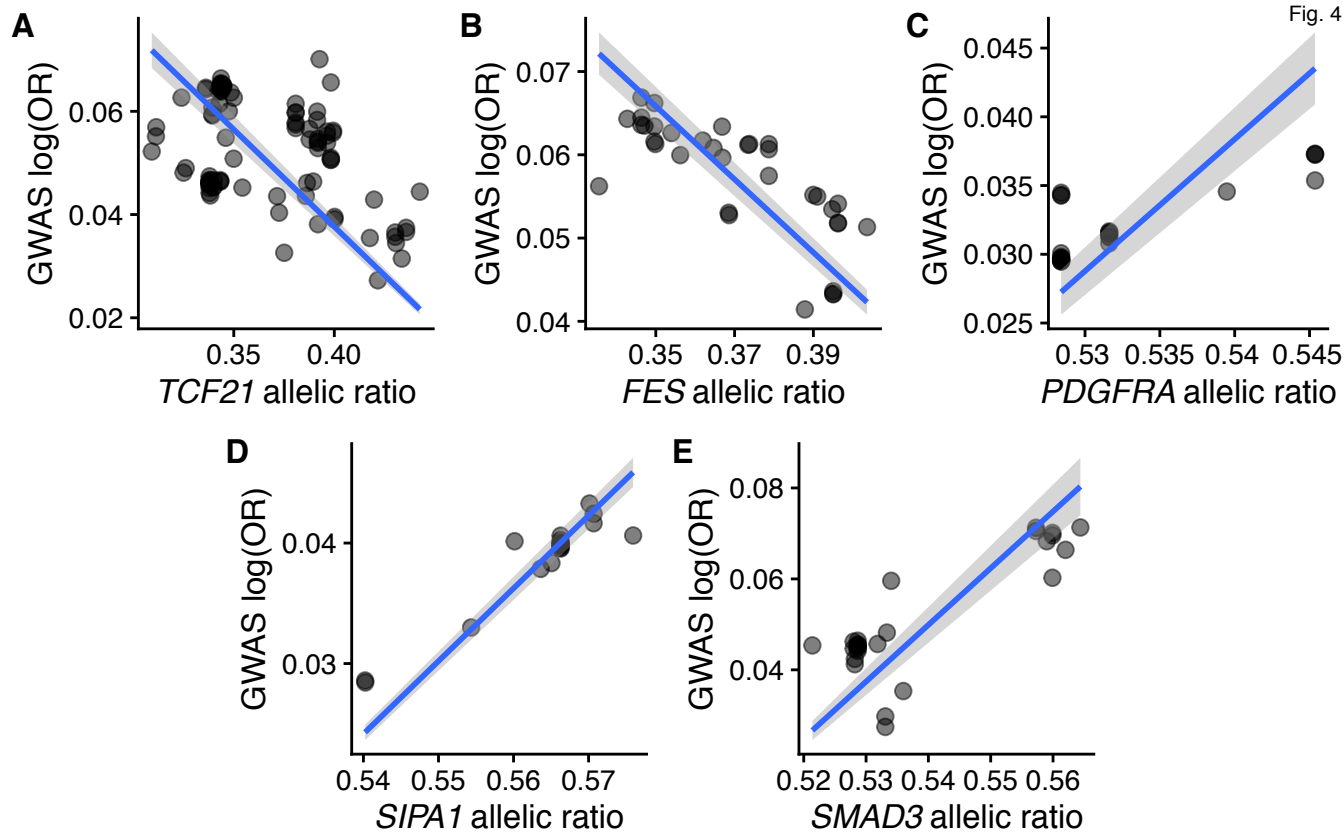
Table 1. Molecular quantitative trait loci discoveries.

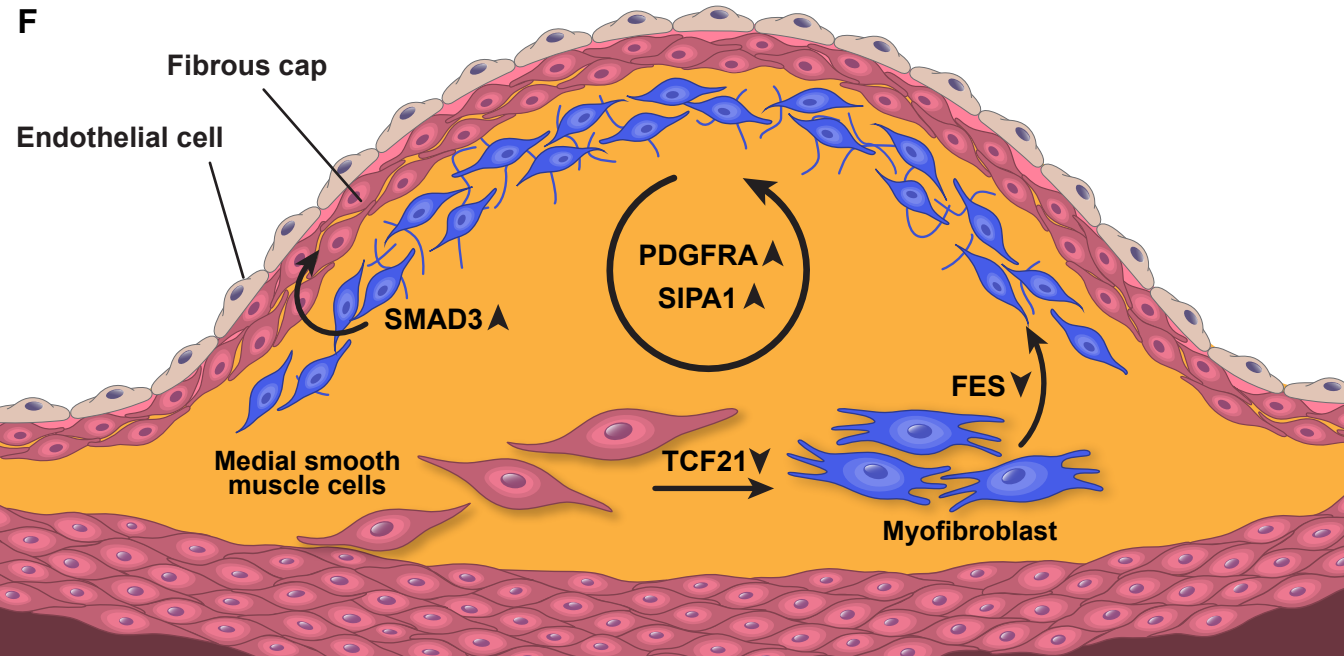
Molecular Phenotype	Trait type	# of traits tested	# of traits with at least one QTL		
			FDR = 0.05	FDR= 0.01	FDR = 0.001
Gene expression	Protein coding	13624	1048 (7.69%)	841 (6.17%)	636 (4.67%)
	lincRNA	1266	51 (4.03%)	41 (3.24%)	33 (2.61%)
	Pseudogene	2616	50 (1.91%)	34 (1.3%)	25 (0.96%)
	Other	2101	71 (3.38%)	56 (2.67%)	44 (2.09%)
	Total	19607	1220 (6.22%)	972 (4.96%)	738 (3.76%)
Splicing	Protein coding	24461	519 (2.12%)	349 (1.43%)	245 (1%)
	lincRNA	300	11 (3.67%)	7 (2.33%)	5 (1.67%)
	Pseudogene	376	22 (5.85%)	15 (3.99%)	12 (3.19%)
	Other	541	29 (5.36%)	19 (3.51%)	17 (3.14%)
	Total	25678	581 (2.96%)	390 (1.99%)	279 (1.42%)


A
Adipose
Adrenal Gland
Bladder
Blood
Blood Vessel
Brain
Breast
Cervix
Colon
Esophagus
Fallopian Tube
Fibroblasts
HCASMC
Heart
Kidney
LCL
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Salivary Gland
Skin
Small Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
MDS Coordinate 1
MDS Coordinate 2
B
Epigenomic Similarity (Jaccard Index)
Top 5
Fibroblasts
Fibroblasts (fetal)
Muscle – Skeletal (fetal)
Heart (fetal)
Lung (fetal)
Bottom 5
Natural killer cell
Prostate
Skin (fetal)
Testis
Brain – cortex
Adult
Fetal
ENCODE Tissue/Cell Type (n=55)
C
DE genes
FDR
0.001
0.01
0.05
direction
down
up
CA
Fb
D
7332
(2.13%)
336952
(97.87%)
HCASMC
shared
E
10 kb
hg19
201,890,000
201,895,000
201,900,000
201,905,000
HCASMC
Cardiac fibroblast
Pulmonary artery endothelial cell
Aortic adventitial fibroblast
Cardiac muscle cell
Cerebellar cortex
Astrocytes
Kidney epithelial cells
Thyroid gland
T-helper cell
Naive thymus-derived cd4+ cells
LMOD1


A
Artery – Coronary
Adipose – Subcutaneous
HCASMC
Lung
Whole Blood
Fibroblast
Liver
Skin – Suprapubic
Adrenal Gland
Muscle – Skeletal
Pancreas
LCL
Brain – Caudate
Pituitary
Minor Salivary Gland
Testis
0
10
20
Enrichment
B
Artery Endothelial Cell
Monocyte
Uterus
B cell
Brain – pons
HCASMC
Odds ratio
0
1
2
3
1
2
3
–log10(P–value)


PDGFRA
SIPA1
SMAD3
A
rs13134452
B
rs12801636
C
rs72743461
GWAS −log10(P)
eQTL −log10(P)
r²
0.8
0.6
0.4
0.2
D
PDGFRA
SIPA1
FES
SMAD3
TCF21
CLPP
eCAVIAR
−log10(P)
SMR
Chromosome
E
rs2327429
F
rs2521501
G
HCASMC
Esophagus
LCL
Nucleas accumbens
Breast
Cerebellum
Pancreas
Stomach
Colon
Nerve – Tibial
Spleen
Adrenal Gland
Prostate
Muscle – Skeletal
Atrial Appendage
TCF21
FES
SIPA1 CLPP


A
GWAS log(OR)
0.06
0.04
0.02
0.35
0.40
TCF21 allelic ratio
B
GWAS log(OR)
0.07
0.06
0.05
0.04
0.35
0.37
0.39
FES allelic ratio
C
GWAS log(OR)
0.045
0.040
0.035
0.030
0.025
0.53
0.535
0.54
0.545
PDGFRA allelic ratio
D
GWAS log(OR)
0.04
0.03
0.54
0.55
0.56
0.57
SIPA1 allelic ratio
E
GWAS log(OR)
0.08
0.06
0.04
0.52
0.53
0.54
0.55
0.56
SMAD3 allelic ratio


F
Fibrous cap
Endothelial cell
SMAD3
PDGFRA
SIPA1
FES
TCF21
Medial smooth muscle cells
Myofibroblast