SUPPLEMENTAL MATERIAL

SUPPLEMENTAL METHODS

GWAS Summary Level Data

We retrieved GWAS summary statistics of an interim release of UK Biobank (UKBB) data,[1] where the sample size was up to 336,924 including 10,801 CAD cases. We also retrieved summary level data on the CARDIoGRAMplusC4D (Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics) consortium[2] based on 60,801 CAD cases and 123,504 controls. Key datasets used in this study are summarized in Supplemental Table I.

STARNET

Subject recruitment and tissue collection in STARNET were performed as previously described.[11] Briefly, patients with CAD who were eligible for open-thorax surgery at the Department of Cardiac Surgery, Tartu University Hospital in Estonia as well as control subjects without CAD were enrolled into this approved protocol and after informed consent. From each STARNET subject, venous blood (BLOOD) as well as biopsies from atherosclerotic aortic wall (AOR), pre/early-atherosclerotic mammary artery (MAM), liver (LIV), skeletal muscle (SKLM), subcutaneous fat (SF) and visceral fat (VAF) were obtained and RNA was extracted as described.[11,12] BLOOD was also used to obtain macrophages (MP) and foam cells (FC), as well as DNA which was isolated for genotyping using the Illumina Infinium assay. Additional details are available in the original[11] and subsequent manuscripts.[12,32]

Analysis of eQTLs in STARNET was performed as described.[11] Briefly, genotyping data was analyzed using GenomeStudio 2011.1 (Illumina) which produced 951,117 genomic markers (genome build 37). Quality control was performed using PLINK v.1.07, and for genotype imputation IMPUTE2 v.2.3.0 was used to increase the power of the analysis.[11] Whole transcriptome sequencing reads were mapped with STAR v.2.3.0e onto the human genome (Genome Reference Consortium GRCh37).[33] After quality control, the expression matrix was normalized and adjusted for covariates. *cis*- regulated eQTLs were identified with the R package Matrix eQTL v.2.1.1.[34] Only bi-allelic markers were included and all *cis*-regulatory SNPs located within 1MB of the gene were tested using a linear model. For single locus eQTL queries, eQTLs with nominal P value < 0.05 per gene were considered statistically significant.

Multi-tissue weighted gene co-expression networks were constructed as described.[35] Correlations between genes were calculated using weighted gene co-expression network analysis (WGCNA) to identify co-expression network modules within and across tissues.[36] Regulatory networks were reconstructed using GENIE3, an algorithm that infers gene regulatory networks from expression data based on feature selection with tree-based Random Forest ensemble methods.[37] Key driver analyses were performed using weighted key driver analysis (wKDA) in Mergeomics,[38] a computational pipeline that overlays disease-associated processes onto molecular interaction networks to pinpoint hubs as potential key regulators. Network and key driver analysis results were imported into Cytoscape v3.7.0 to generate a network visualization.[39]

LD Score Regression (LDSC) Analysis

LD Score Regression[40,41] version 1.0.0 was employed to test the enrichment of CAD heritability for specific genomic annotations (ie, 143 GTEx genomic annotations and 53 BLUEPRINT epigenomic annotations of various tissue/cell types). The European subjects in the 1000Genome dataset served as an LD score reference.

Integrative genomics analysis (IGA)

This IGA incorporated two sources of data: GWAS summary statistics from an interim release of UKBB data[1] or CARDIoGRAMplusC4D,[2] and tissue/cell-specific eQTLs from STARNET.[11] As a validation, the same IGA pipeline was applied using the same GWAS datasets but substituting GTEx (V7)[18] for STARNET. Our IGA pipeline employed three methods: MetaXcan,[8] SMR[6,7] and Coloc,[9] which belong to two broad classes. Class 1 methods (e.g. MetaXcan and SMR) test for significant genetic correlations between cis expression and GWAS. To improve statistical power, we employed both MetaXcan and SMR, and results were combined by taking the union. Class 2 methods (e.g. Coloc) estimate the posterior probability of colocalization, defined as shared causal variant(s) between gene expression and GWAS. We intersected the results of Class 1 and 2 methods to identify a set of likely causal genes for CAD.

In greater detail, we applied MetaXcan[8] to integrate each pair of a GWAS summary dataset and the eQTL dataset of a given STARNET tissue/cell type to identify putative causal genes for CAD. We built the prediction model for each STARNET tissue which consisted of SNP weights and tissue-specific covariance matrices. The prediction models were built using

methods published by the PrediXcan/MetaXcan authors which are available at https://github.com/hakyimlab/PredictDB-Tutorial. The prediction models we applied for GTEx tissues were built by the PrediXcan/MetaXcan authors and are available at https://zenodo.org/record/3572799#.YR3SOt8pAtJ.

In parallel, we applied SMR[6,7,42] to CAD GWAS and eQTLs to identify candidate causal genes. SMR also uses linkage disequilibrium (LD) information to differentiate pleiotropy from causality by reporting the results of a heterogeneity test (findings of $P_{HEIDI} < 0.05$ were filtered out). In total, 36 sets of tests were conducted over two tests (MetaXcan and SMR), two CAD GWASs (UKBB and CARDIoGRAMplusC4D) and eQTLs of 9 STARNET tissue/cell types (Supplemental Table IV). We pooled all the tests and calibrated the FDR according to this burden of multiple testing, with findings at FDR ≤ 5% considered as significant.

Next, we applied Coloc to further filter the significant findings of MetaXcan and SMR, and removed genes of Coloc $H_4 < 0.8$ to identify the final set of candidate causal genes for CAD. To achieve this, we first defined "intervals" around each gene, which include gene body ± 500 kb flanking, then colocalization analysis was performed within each interval using COLOC version 2.3-6 in R.[9] This method assesses whether two association signals, GWAS summary statistics and eQTL statistics, are consistent with shared causal variant(s).[9] In total, five hypotheses were evaluated. H0: No association with either CAD risk or gene expression; H1: Association with CAD, not with gene expression; H2: Association with gene expression, not with CAD; H3: Association with both CAD and gene expression, and multiple independent SNPs influencing the two traits; H4: Association with both CAD and gene expression, one shared SNP. We consider posterior probability (PP.H4) ≥ 0.8 as evidence that CAD risk and gene expression of a given tissue/cell type were controlled by the same genetic variant.

STARNET eQTLs modified by CAD disease status

The STARNET eQTLs of 7 tissue/cell types (AOR, MAM, LIV, SKLM, BLOOD, SF and VAF) were from CAD cases only. The eQTLs of MP and FC were derived from both CAD cases and controls. To explore if it was necessary to account for this, we explicitly tested if MP and FC eQTLs were modified by disease status and affected our results. We fitted a linear model with interaction terms GEX ~ SNP + D + SNP*D, where GEX denotes the expression level of a candidate CAD causal gene detected in MP or FC, and D denote the disease status. The

significance level of the interaction term was determined using permutation-based empirical FDR, and findings < 5% FDR would indicate that the eQTL is influenced by disease status, otherwise, the eQTL would be considered as comparable in cases and controls. At 5% FDR, none of the candidate CAD causal genes' eQTLs were affected by CAD status. Because of this negative result, we did not purse any stratification based on case vs. control status.

CAD GWAS loci nominated to be associated with candidate causal genes

To compare our list of candidate causal CAD genes with existing lists of putative causal CAD genes, we adopted a list of 163 independent CAD association peaks that was recently compiled by Erdmann et al.[4] Around each peak SNP, we constructed an interval of ± 200kb, and considered the locus nominated for candidate gene(s) if the gene body overlaps with the interval.

Validation study using GTEx eQTLs

Seven CAD relevant tissues of GTEx (V7)[18] (AOR, SKLM, BLOOD, LIV, SF, VAF and coronary artery [COR]) were used as a validation dataset. Because GTEx did not include MP, FC or MAM, these tissues/cell types were not included in this validation analysis. The same IGA pipeline as applied to the STARNET datasets was applied to integrate the two CAD GWASs (UKBB and CARDIoGRAMplusC4D) and GTEx eQTLs. In total, 28 sets of test results were generated (2 CAD GWASs vs. eQTLs of 7 GTEx tissues vs. 2 methods [MetaXcan and SMR]). This gave 108,755 tests in total (Supplemental Table IV), and we then used this number to calibrate the FDR. Findings at FDR ≤ 5% were considered as significant and subsequently filtered by Coloc H_4 < 0.8 criteria, and compared to candidate CAD causal genes identified using STARNET eQTLs.

Hypergeometric testing

Hypergeometric testing with Benjamini-Hochberg correction was performed for the 224 STARNET GRNs in relation to the 162 candidate causal genes identified by the IGA, as previously described.[35] The same analysis was also run on the 224 STARNET GRNs using sets of 162 randomly selected genes (from https://www.molbiotools.com/randomgenesetgenerator.html).

Linkage disequilibrium determinations

SNPs 5000bp upstream and downstream of rs9349379 were queried from the Ensembl database and linkage disequilibrium (LD) scores between the SNPs were calculated using LDlinkR.[43] An LD score of 0.8 was used to determine SNPs that were in LD.

SUPPLEMENTAL TABLES

Only Supplemental Table I is presented in this document. Supplemental Tables II – XIII are presented online in a separate file.

Supplemental Table I. Summary of cohorts and datasets used in this study. Abbreviations: CAD, coronary artery disease; ChIP sequencing, chromatin immunoprecipitation sequencing; RNAseq, RNA sequencing; WES, whole exome sequencing; WGS, whole genome sequencing. *For all tissues in STARNET except foam cells (FC) and macrophages (MP), only CAD cases were used in this analysis.

Cohort name	Patient/case features	Number of cases or patients	Control features	Number of controls	Sample (s) collected	Techniques applied to samples	Reference
STARNET: Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task*	Patients with angiographically-proven CAD requiring CABG	~600	Patients without CAD having open-thorax surgery (i.e. valve replacement)	~150	Venous blood (BLOOD); biopsies from atherosclerotic aortic wall (AOR), pre/early-atherosclerotic mammary artery (MAM), liver (LIV), skeletal muscle (SKLM), subcutaneous fat (SF) and visceral fat (VAF). BLOOD was also used to obtain MP and FC.	DNA SNP-based genotyping with imputation, and RNAseq on all tissue samples	11,12,32
GTEx; Genotype-Tissue Expression project	Samples collected from ~900 deceased donors who were identified through organ and tissue transplant programs. There are no specific cases nor controls in GTEx.				48 non-diseased tissue sites including SF, VF, AOR, LIV, SKLM, BLOOD and coronary artery (COR).	Primarily molecular assays including WGS and WES of DNA, and RNASeq of tissues	18
BLUEPRINT	Blood samples were drawn from 200 blood donors from a local blood donor population, ascertained to be free of disease and representative of the United Kingdom (UK) population at large (54% females, mean age 55 years)				From blood, three major human immune cell types were isolated and studied: CD14+ monocytes; CD16+ neutrophils; and naive CD4+ T cells	WGS, RNAseq, genome-wide DNA methylation, and ChIPseq at 2 histone modification marks (H3K4me1, H3K27ac)	16,17

UK Biobank (UKBB)	“Soft” but inclusive CAD definition incorporating self-reported angina or other evidence of chronic CAD, including more stringently defined phenotypes such as myocardial infarction and/or revascularization	10,801	UKBB participants not meeting the “soft” CAD criteria	326,924	DNA	DNA SNP-based genotyping with imputation	1
CARDIoGRAMplusC4D (Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics) consortium	Case status was defined by an inclusive CAD diagnosis (e.g. myocardial infarction, acute coronary syndrome, chronic stable angina or coronary stenosis of >50%).	60,801 Note that this was a meta-analysis that combined data from 48 studies	Controls were generally free of clinical CAD, cerebrovascular disease and peripheral arterial disease.	123,504	DNA	DNA SNP-based genotyping with imputation	2

SUPPLEMENTAL LEGENDS FOR SEPARATE EXCEL FILE OF TABLES

Supplemental Table II. Stratified linkage disequilibrium score regression (LDSC) analysis to identify causal tissues/cell types for CAD using Nelson et al (UKBB).[1] This analysis leveraged data from BLUEPRINT,[16] GTEx[18] and UKBB[1] and integrated GWAS data, eQTL data, gene expression and epigenetic marks to identify disease-relevant tissues. The equivalent results are also presented using GWAS data from CARDIoGRAMplusC4D[2] rather than UKBB[1] in Supplemental Table III. Abbreviations: MaxCPP, maximum value of causal posterior probability; 95CredibleSet, SNPs that belong to the 95% credible set of QTL fine-mapping.

Supplemental Table III. Stratified LDSC analysis to identify causal tissues/cell types for CAD using CARDIoGRAMplusC4D. This analysis leveraged data from BLUEPRINT,[16] GTEx[18] and CARDIoGRAMplusC4D[2] and integrated GWAS data, eQTL data, gene expression and epigenetic marks to identify disease-relevant tissues. The equivalent results are also presented using UKBB[1] rather than CARDIoGRAMplusC4D[2] in Supplemental Table II. Abbreviations: MaxCPP, maximum value of causal posterior probability; 95CredibleSet, SNPs that belong to the 95% credible set of QTL fine-mapping.

Supplemental Table IV. Number of tests conducted for our IGA pipeline using STARNET or GTEx. This table shows the number of tests performed when applying two methods (MetaXcan and SMR), to integrate two CAD GWASs (UKBB[1] and CARDIoGRAMplusC4D[2]), with the eQTLs of either the 9 STARNET tissue/cell types or of the 7 GTEx tissues used in this study.

Supplemental Table V. Integrative genomics analysis (IGA) integrating GWAS summary statistics and tissue/cell-specific eQTLs from UKBB[1] with STARNET.[12] A “1” in columns C – K indicates the tissue in which putative causality was identified. Column L indicates the total number of putative causal tissues identified for each gene. Columns M – CF are the detailed results supporting the results in columns C – L.

Supplemental Table VI. IGA integrating GWAS summary statistics and tissue/cell-specific eQTLs from CARDIoGRAMplusC4D[2] with STARNET.[12] A “1” in columns C – K indicates the tissue in which putative causality was identified. Column L indicates the total number of putative causal tissues identified for each gene. Columns M – CF are the detailed results supporting the results in columns C – L.

Supplemental Table VII. Combined, prioritized results based on Supplemental Tables V and VI, representing all unique candidate causal genes for CAD identified using either UKBB[1] with STARNET,[12] or, CARDIoGRAMplusC4D[2] with STARNET.[12] Causal genes were prioritized based on the smallest P value for the class 1 analyses (MetaXcan or SMR). The top 25 causal genes listed here are also listed in Table 1.

Supplemental Table VIII. List of CAD GWAS loci summarized by Erdmann et al,[4] and candidate causal genes nominated by this IGA for each of these loci. The IGA also pinpoints the tissue specificity of the genes’ causal effects (shown in columns J – R).

Supplemental Table IX. IGA integrating GWAS summary statistics and tissue/cell-specific eQTLs from UKBB[1] with GTEx.[18] A “1” in columns C – I indicates the tissue in which putative causality was identified. Column J indicates the total number of putative causal tissues identified for each gene. Columns K – BN are the detailed results supporting the results in columns C – J.

Supplemental Table X. IGA integrating GWAS summary statistics and tissue/cell-specific eQTLs from CARDIoGRAMplusC4D[2] with GTEx.[18] A “1” in columns C – I indicates the tissue in which putative causality was identified. Column J indicates the total number of putative causal tissues identified for each gene. Columns K – BN are the detailed results supporting the results in columns C – J.

Supplemental Table XI. Common genes identified using both STARNET and GTEx as being causal for CAD. Six tissues were common to both STARNET and GTEx and are

represented by the columns. Results are shown using either UKBB[1] or CARDIoGRAMplusC4D[2] GWAS datasets.

Supplemental Table XII. Representation of candidate causal CAD genes in gene regulatory co-expression networks (GRNs). For each causal gene identified in the IGA from the master list of 162 candidate causal CAD genes in Supplemental Table VII, GRNs curated from the STARNET datasets are shown where the tissue of causality from the IGA matched the tissue of effect for that gene in the GRN. Note that STARNET does not yet have curated GRNs for MP and FC. Therefore this table only considered AOR, MAM, LIV, BLOOD, VAF, SF, SKLM. Abbreviations: KD, key driver.

Supplemental Table XIII. Isoform-specific *PHACTR1* eQTLs at rs9349379. We queried STARNET for isoform-specific *PHACTR1* eQTLs at rs9349379 using Ensembl grch38.p13 release 99. Because we only queried isoform-specific eQTLs at rs9349379 and multiple testing was not performed, all nominally significant eQTLs are shown. “No protein” indicates this is a non-protein coding isoform. Transcript Support Level (TSL) is a method to highlight the level of evidence for differing transcripts, with TSL1 representing the best supported transcripts and TSL5 the least supported.