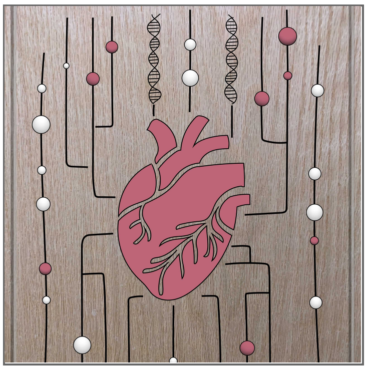

Variant to gene to function workflow for endothelial cell programs related to CAD V.1

Gavin R. Schnitzler[1,2], Helen Kang[3,4], Shi "House" Fang[2], ronghao[3,4], Vivian Lee[2], Rosa Ma[3,4], Ramcharan Angom[5], Jesse Engreitz[3,4], Rajat M. Gupta[2]

[1]Broad Institute of Harvard & MIT, Cambridge, MA;
[2]Divisions of Genetics and Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Boston MA;
[3]Department of Genetics, Stanford University School of Medicine, Stanford, CA;
[4]BASE Initiative, Betty Irene Moore Children’s Heart Center, Lucile Packard Children’s Hospital, Stanford, CA;
[5]Department of Biochemistry and Molecular Biology, Mayo Clinic College of Medicine and Science, Jacksonville, FL

Gavin R. Schnitzler: For Perturb-seq methods, dialout analysis, bulk RNAseq, TeloHAEC microscopic analyses, differential expression analysis, co-IP assays;
Helen Kang: For cNMF, G2P and V2G2P analyses, internal & external benchmarking, application to other traits & cell types.;
Shi "House" Fang: For assays under laminar flow, nucleofection of TeloHAEC;
ronghao: For FLOW-Fish;
Vivian Lee: For TEER/ECIS assay of endothelial barrier function
Rosa Ma: For V2G analysis (linking variants to genes using ABC).
Ramcharan Angom: For zebrafish methods
Jesse Engreitz: Co-corresponding author of manuscript
Rajat M. Gupta: Co-corresponding author of manuscript

RguptaLab Engreitz Lab

Gavin R. Schnitzler

VERSION 1

JAN 12, 2024

External link: https://www.biorxiv.org/content/10.1101/2022.11.01.514606v1.article-info

Protocol Citation: Gavin R. Schnitzler, Helen Kang, Shi "House" Fang, ronghao, Vivian Lee, Rosa Ma, Ramcharan Angom, Jesse Engreitz, Rajat M. Gupta 2024. Variant to gene to function workflow for endothelial cell programs related to CAD. **protocols.io** https://protocols.io/view/variant-to-gene-to-function-workflow-for-endotheli-c7f3zjqn

MANUSCRIPT CITATION:
Title: **Convergence of coronary artery disease genes onto endothelial cell programs**

Publication status: In press at **Nature**

Preprint available on BioRxiv: https://www.biorxiv.org/content/10.1101/2022.11.01.514606v1.article-info

Authors & affiliations:
Gavin R. Schnitzler1,2,5*, Helen Kang3,4,*, Shi Fang1,5, Ramcharan S. Angom6, Vivian S. Lee-Kim1,5, X. Rosa Ma3,4, Ronghao Zhou3,4, Tony Zeng3,4, Katherine Guo3,4, Martin S. Taylor15, Shamsudheen K. Vellarikkal1,5, Aurelie E. Barry1,5, Oscar Sias-Garcia1,5, Alex Bloemendal1,2, Glen Munson1, Philine Guckelberger1, Tung H. Nguyen1, Drew T. Bergman1,7, Stephen Hinshaw16, Nathan Cheng1, Brian Cleary1,8, Krishna Aragam1,9, Eric S. Lander1,10,11, Hilary K. Finucane1,12,13,14, Debabrata Mukhopadhyay6, Rajat M. Gupta1,2,5,†, Jesse M. Engreitz1,2,3,4,17†

1. Broad Institute of MIT and Harvard, Cambridge, MA
2. The Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute, Cambridge, MA
3. Department of Genetics, Stanford University School of Medicine, Stanford, CA
4. BASE Initiative, Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Stanford, CA
5. Divisions of Genetics and Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Boston MA
6. Department of Biochemistry and Molecular Biology, Mayo Clinic College of Medicine and Science, Jacksonville, FL
7. Geisel School of Medicine at Dartmouth, Hanover, NH
8. Faculty of Computing and Data Sciences, Departments of Biology and Biomedical Engineering, Biological Design Center, and Program in Bioinformatics, Boston University, Boston, MA
9. Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA
10. Department of Biology, MIT, Cambridge, MA
11. Department of Systems Biology, Harvard Medical School, Boston, MA
12. Department of Medicine, Massachusetts General

ABSTRACT

Linking variants from genome-wide association studies (GWAS) to underlying mechanisms of disease remains a challenge1,4,6. For some diseases, a successful strategy has been to look for cases where multiple GWAS loci contain genes that act in the same biological pathway[1–6]. However, our knowledge of which genes act in which pathways is incomplete, particularly for cell-type specific pathways or understudied genes. Here we introduce a method to connect GWAS variants to functions, which links variants to genes using epigenomic data, links genes to pathways de novo using Perturb-seq, and integrates these data to identify convergence of GWAS loci onto pathways. We apply this approach to study the role of endothelial cells in genetic risk for coronary artery disease (CAD), and discover that 43 CAD GWAS signals converge on the cerebral cavernous malformations (CCM) signaling pathway. Two regulators of this pathway, CCM2 and TLNRD1, are each linked to a CAD risk variant, regulate other CAD risk genes, and affect atheroprotective processes in endothelial cells. These results suggest a model where CAD risk is driven in part by the convergence of causal genes onto a particular transcriptional pathway in endothelial cells, highlight shared genes between common and rare vascular diseases (CAD and CCM), and identify TLNRD1 as a new, previously uncharacterized member of the CCM signaling pathway. This approach will be widely useful for linking variants to functions for other common polygenic diseases.

Note that authors listed with emails for this protocol are those who are best to contact for questions regarding the methods, and do not represent all contributions to the manuscript (for that see the Manuscript Citation section)

IMAGE ATTRIBUTION

Helen Kang & Gavin Schnitzler

GUIDELINES

INTRODUCTION TO THE V2G2P APPROACH FOR CAD

Genetic variants that influence complex traits are thought to regulate genes that work together in biological pathways. Identifying convergence on particular pathways can help in discovering genes and cellular functions that causally influence disease risk[1–6]. However, it is often challenging to identify such convergence: complex traits involve contributions from multiple cell types; most risk variants are noncoding and can regulate multiple nearby genes; and it remains unclear which genes work together in which pathways in which cell types[7–9]. GWAS for coronary artery disease have discovered over 300 independent signals[10–12]). 75% of these signals are not associated with circulating lipids, indicating the presence of undiscovered disease mechanisms that may function through cells in the coronary artery where atherosclerosis that causes CAD

Hospital, Boston, MA
13. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA
14. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA
15. Department of Pathology, Massachusetts General Hospital, and Harvard Medical School, Boston, MA
16. Department of Chemical and Systems Biology, ChEM-H, and Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA
17. Stanford Cardiovascular Institute, Stanford University, Stanford, CA, USA
*Equal contribution.
†Equal contribution.

Protocol status: Working
We use this protocol and it's working

Created: Jan 11, 2024

Last Modified: Jan 12, 2024

PROTOCOL integer ID: 93403

Keywords: Perturb-seq, scRNA-seq, endothelial cells, coronary artery disease, GWAS, variant-to-function, consensus non-negative matrix factorization

develops. Endothelial cells (ECs) are one of the most important of these arterial cells, controlling cholesterol uptake and efflux, smooth muscle cell responses, blood clotting and inflammatory immune cell recruitment[13,14], and are highly enriched for CAD heritability[15]. At a few individual CAD GWAS loci, noncoding risk variants have been shown to regulate the expression of key EC genes such as endothelial nitric oxide synthase (NOS3), endothelin 1 (EDN1), and others[16]. It remains unclear, however, which other genes in CAD GWAS loci might work together in which EC pathways to modulate disease risk.

To address these challenges, we have developed a new approach that systematically and unbiasedly links GWAS variants to genes and identifies their convergence onto specific disease-associated transcriptional programs. The 5 steps of this Variant-to-Gene-to-Program (**V2G2P**) approach, and their application to EC functions in CAD, are summarized below:

1. Identify a cell type and cellular model relevant to disease genetics, through enrichment of disease risk variants in enhancers in that cell type. Here, we focused on human arterial ECs, using telomerase-immortalized human aortic ECs (teloHAEC) as a model.
2. Build a map of variant-to-gene (V2G) links in that cell type, to link disease-associated variants to potential target genes. Here, we consider evidence from variants in EC enhancers, as well as coding regions and splice sites.
3. Build a map of gene-to-program (G2P) links in that cell type, by using Perturb-seq[17–20] to systematically knock down all possible candidate disease genes and identify sets of genes that act together in biological pathways. Here, we knock down all expressed genes within ±500kb of 306 CAD GWAS signals, read out the effects of each perturbation with single cell RNA-seq, and use unsupervised machine learning to define gene "programs," unbiased by prior knowledge of gene sets or pathways.
4. Identify "disease-associated programs", by developing a statistical test to determine whether the genes with links to risk variants are enriched in (that is, converge on) particular programs. Here, we find that many CAD GWAS loci converge on 5 gene programs identified de novo with Perturb-seq, which appear to correspond to branches of the cerebral cavernous malformations (CCM) signal transduction pathway.
5. Study the genes in disease-associated programs. Here, we nominate 41 genes likely to influence CAD risk through effects in ECs, and dissect two in detail: showing that knockdown of TLNRD1 or CCM2 mimics the effects of atheroprotective laminar blood flow, and that the poorly-characterized gene, TLNRD1, is a novel regulator in the CCM pathway.

In summary, the V2G2P approach defines cellular programs de novo using Perturb-seq, intersects these programs with enhancer-to-gene maps from the same cell type, and provides an interpretable, systematic, and unbiased framework for tracing the path from variant to gene to disease program simultaneously for all GWAS loci for a given disease and cell type.

GENERAL CONSIDERATIONS FOR APPLYING THE V2G2P APPROACH.

Funders Acknowledgement:
NHLBI
Grant ID: R01HL159176
NHLBI
Grant ID: R01HL164811
NHGRI
Grant ID: UM1HG011972
NHGRI
Grant ID: R35HG011324
NHGRI
Grant ID: K99HG009917
NHGRI
Grant ID: R00HG009917
Novo Nordisk Foundation
Grant ID: NNF21SA0072102
NHLBI
Grant ID: DP2HL152423
NHLBI
Grant ID: U01HL166060
NHLBI
Grant ID: HL70567
NIDDK
Grant ID: K08DK129824
Broad Institute
Grant ID: Variant-to-Function Initiative

We aimed to create an approach to identify genes and programs relevant to disease risk that was cell-type specific, interpretable, unbiased with respect to prior information, and generally applicable to many cell types and complex traits. We and others have previously shown that combining both “top-down” information from gene programs and “bottom-up” approaches linking variants to genes can achieve higher specificity than either category of information alone[3,31,32]. By combining GWAS, epigenomic, and Perturb-seq data, the variant-to-gene-to-program (V2G2P) approach expands upon these previous approaches by (i) generating variant-to-gene and gene-to-program maps in the same cell type; (ii) generating gene-to-program maps using Perturb-seq, providing a unbiased approach not dependent on previously known biological pathways or gene sets; and (iii) providing interpretable, testable hypotheses linking a specific variant to a gene to a program in a given cell type.

To implement this approach, we selected a cellular model enriched for heritability for the disease of interest. We constructed genome-wide enhancer-to-gene maps in endothelial cells by applying the Activity-by-Contact (ABC) model, which we recently showed performs well at linking noncoding variants to target genes in specific cell types[9,22]. ABC outperforms other methods at predicting the effects of enhancers on target genes [9,22], and requires minimal data inputs (e.g., ATAC-seq and H3K27ac ChIP-seq), allowing us, here, to apply the approach to link variants to candidate target genes in multiple endothelial cell states. We next created a catalog of gene programs and their regulators by applying Perturb-seq to systematically study all expressed genes in all GWAS loci for CAD. Perturb-seq, which involves knocking down hundreds to thousands of genes in parallel and measuring their effects on gene expression using single-cell RNA-seq, has previously been shown to provide a high-content, unbiased view of cellular programs as represented in gene expression[17–19]. Finally, we developed a simple statistical test to determine whether candidate disease genes might converge on particular gene programs by integrating gene-to-program information from Perturb-seq with variant-to-gene linking approaches.

Our approach to building a gene-to-program map using CRISPRi-Perturb-seq involved particular design and analysis considerations:

(i) We aimed to identify cellular programs and their related genes in an unbiased manner, such that we could look for enrichment of candidate CAD genes across a range of different endothelial cell pathways. This is in contrast to the approach of selecting a particular cellular phenotype (such as endothelial cell adhesion) that may or may not be important for the genetics of disease. Accordingly, we selected Perturb-seq due to its ability to perturb many hundreds or thousands of genes in parallel, and its ability to read out the effects on all genes in the genome, thereby providing a high-throughput and high-content readout of cell states.

(ii) Targeting all candidate genes near GWAS signals was important for the V2G2P approach. Specifically, we designed our Perturb-seq study to include all expressed genes within 500 Kb on either side of each CAD GWAS signal, as well as the two closest genes on either side if they were further than 500Kb, rather than selecting just a prioritized subset of genes. In practice, this resulted in us testing a median of

8 genes per CAD GWAS locus. This unbiased approach to selecting genes was essential for conducting the V2G2P enrichment test, which examines whether particular programs contain more genes with CAD variant-to-gene (V2G) links than expected by chance. This assessment would have been impossible if we had pre-selected only genes with V2G links to include in the Perturb-seq experiment. As such, the V2G2P enrichment test is applicable to experiments that perturb all expressed genes in all GWAS loci, or all genes in the genome.

(iii) The CRISPRi Perturb-seq approach was designed to read-out long term transcriptomic effects of gene knockdowns in the expected range of effect for common disease variants. We aimed to perturb genes in a way consistent with presumed effects of noncoding variants, which are thought to lead to quantitative changes in the expression of genes (rather than completely eliminating expression), and which might act over long periods of time to affect disease risk. Accordingly, we used CRISPRi to quantitatively knock down gene expression (average: 40% reduction). We then read out the effects after 5 days of doxycycline induction, to allow perturbations to propagate through the network and identify how perturbations affect stable gene expression programs.

(iv) We defined gene “programs” using an unsupervised machine learning approach (consensus non-negative matrix factorization), allowing us to identify sets of genes with similar properties (here, co-expression across single cells). This approach is independent of and unbiased by prior knowledge about endothelial cell pathways — allowing us to avoid bias toward rediscovering or over-emphasizing known pathways, and identify new pathways if they exist. We did indeed discover gene programs that appeared to correspond to a wide range of biological pathways active in endothelial cells. Many of the 50 programs appeared to correspond to housekeeping pathways active in all cell types, because these genes/pathways (as well as EC-specific ones) are expressed and functional in endothelial cells.

MATERIALS

Required materials are described under Steps.

General cell culture

1 Telomerase-immortalized human aortic endothelial cells (TeloHAEC) were purchased from ATCC (CRL-4052), and grown in Lifeline VEGF endothelial cell media (LL-0005) with 1x Penn/Strep. Cells were plated at a density of 0.5-1.0 x 10^6 cells per 10 cm plate and split before reaching 4 x 10^6/plate (3 to 4 days). Eahy926 cells (a HUVEC + A549 hybrid line) and HEK293T cells were purchased from ATCC (CRL-2922 and CRL-3216, respectively), and grown in DMEM + 10% FBS.

To study responses to CAD-associated cytokines, cells were untreated (control), or treated with 10 ng/ml recombinant human IL-1β (Millipore IL038), 10 ng/ml recombinant TNFα (Millipore GF023), or with normal media lacking VEGF (for TeloHAEC) or supplemented with VEGF (1x concentration from LifeLine VEGF media, for Eahy926), for 24 hours.

Note: Be sure that cell lines being used are mycoplasma free and are the correct lines. For instance, in addition to authentication by the provider, we further authenticated each line by analysis of microscopic morphology (e.g. TeloHAEC displayed the characteristic EC cobblestone morphology and showed localization of VE-Cadherin to endothelial cell junctions), functionality (high transfectability and protein expression for HEK293T), mapping of ATAC-seq, ChIP-seq and RNAseq reads to the human genome, and RNAseq profiles and responses (e.g., for Eahy926 & TeloHAEC, expression of EC-specific genes and observation of previously-observed responses to stimuli, such as IL-1β).

Identification of active chromatin regions (ATAC-seq and H3K…

2 **ATAC-seq.**

For ATAC-seq, one well of a 12-well plate (~200,000 cells) was directly lysed using a custom TN5 buffer (33 mM Tris Acetate pH 7.8, 66 mM Potassium Acetate, 10 mM Magnesium Acetate, 16% dimethylformamide & 0.1% NP40). 47.5 µl of lysed cells was added to 2.5 µl Tn5 tagmentation enzyme (Illumina) & incubated at 37oC for 1 hr, and the reaction stopped by addition of 20 µl buffer RLT (Qiagen). Products were purified by addition of 1.8 volumes Ampure XP beads (Beckman-Coulter) & magnetic separation of beads, followed by two 80% ethanol washes, brief drying of pellets & resuspension in 23 µl water. Barcoded ATAC-seq libraries were then generated as described in [9,22], and sequenced to a depth of 10-20 million reads per library.

3 **H3K27ac ChIP-seq.**

Chromatin immunoprecipitation for histone H3 lysine 27 acetylation (H3K27ac) was performed as described in [9,22], using anti H3K27ac antibody (#39685, Active Motif) at 1:200 dilution. ChIP-seq libraries were prepared using the KAPA Hyper Prep Kit (KAPA Biosystems). ATAC-seq libraries were prepared in biological triplicate, and ChIP-seq libraries in biological duplicate. For both types of libraries, reads were mapped to the human genome (hg19 build) using Bowtie2, and peaks identified using MACS2, essentially as per [9,22].

Linking enhancers to genes using the Activity-by-Contact mo…

4 Enhancers and their predicted target genes were identified by applying the Activity-by-Contact (ABC) model to these data, using ATAC-seq and H3K27ac ChIP-seq as the measures of enhancer Activity, and using a cross-cell type average of Hi-C maps as the measure of 3D enhancer-promoter contact frequency (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction 7,8). We used an ABC fractional score threshold of ≥0.015[9].

Variant to Gene (V2G) analysis to identify genes plausibly-reg…

5 **Defining variants in CAD GWAS signals for variant-to-gene analysis**

CAD lead GWAS variants were derived from both Aragam et al.[12] and Harst et al.[10]. We excluded lead variants from Harst et al. if the variants were in strong LD (r2 ≥ 0.7) with an Aragam et al.[12] lead variant or were ≤5Kb away from an Aragam et al. lead variant. An LD-expansion was performed to include variants that are both within a 1 Mb window of, and are in strong LD (r2 ≥ 0.9) with the any of these lead GWAS variants in 1000 Genome European ancestry (plink --ld-window-kb 1000 --ld-window 99999 --ld-window-r2 0.9). For each lead variant, we also included variants prioritized through functionally informed fine-mapping (PIP ≥ 0.1) in either study[12][10]. We defined a "**GWAS Signal**" as this collection of variants around, and including, each lead variant.

6 **Identifying CAD variants associated with lipid levels**

We classified CAD GWAS signals as “lipid” or “non-lipid” based on their association with lipid levels in other GWAS studies, because the CAD GWAS signals also associated with lipids are presumed to act through non-endothelial cells such as hepatocytes. For lead signals included in Aragam et al.[12], we defined a CAD GWAS signal to be associated with lipids if the lead variant was linked to “LDL-direct”, “Triglycerides”, “Cholesterol”, “HDL-cholesterol”, “Apolipoprotein A”, “Apolipoprotein B”, “HDLC” or “LDLC” in the phenome-wide association scan (PheWas) conducted by Aragam et al.[12] For GWAS signals exclusively nominated by Harst et al[10], we used a different procedure in which we considered a signal to be associated with lipids if its lead variant was associated ($P < 5 \times 10{-8}$) with HDLC, LDLC, TG, ApoA, or ApoB based on GWAS from the UK Biobank (Hilary Finucane and Jacob Ulirsch: https://www.finucanelab.org/data). We refer to the remaining GWAS signals not associated with lipid levels as “non-lipid CAD GWAS signals”, and focused on this subset of signals as cases where CAD variants might plausibly act in endothelial cells.

7 **Linking variants to genes**

We used a combination of variant-to-gene methods to identify a list of genes linked to CAD variants that could plausibly act in endothelial cells. At each CAD GWAS signal, we considered as candidate genes at least two genes upstream or downstream of the lead GWAS SNP, and all the genes within +/- 500Kb of the lead variant to be potentially regulated by the GWAS signal. We focused our analysis on protein-coding genes and excluded long noncoding RNAs (“^LINC”), gene isoforms (“-AS”), microRNAs (“^MIR”), small nuclear RNAs (“RNU”), and genes of uncertain functions ("^LOC"). To link CAD variants to genes, we intersected the variants with ABC enhancers[9] in endothelial cells to identify the top two genes most likely to be regulated by each variant (highest 2 ABC fractional scores over 0.015). Specifically, we used ABC data, for enhancers and predicted target genes, from TeloHAEC and Eahy926 (control, or treated with IL1β, TNFα or VEGF, this study), and from prior ABC analysis of HUVEC (‘endothelial_cell_of_umbilical_vein_Roadmap’, ‘endothelial_cell_of_umbilical_vein_VEGF_stim_12_hours-Zhang2013’, and ‘endothelial_cell_of_umbilical_vein_VEGF_stim_4_hours-Zhang2013’ datasets from [9]). To account for cell state-specific regulation that was not predicted by ABC, we also intersected candidate CAD variants at each signal with ATAC peaks and considered the 2 genes closest to variant-containing peaks as plausibly regulated. We also linked variants to genes if the variant was in a coding sequence or within 10 bp of a splice site annotated in the RefGene database (downloaded from UCSC Genome Browser on 24 June 2017)[68]. We identified 254 candidate CAD genes, defined as “**genes with V2G (variant-to-gene) links”**, at 125 of 228 non-lipid CAD GWAS signals.

We have created a snakemake pipeline for V2G analysis, which is available at: https://github.com/EngreitzLab/V2G (DOI: 10.5281/zenodo.10357646) (V2G)

Using epigenetic and expression data to validate disease rele…

8 **General considerations for selection of a cellular model for Perturb-seq.** The cellular model should be relevant to the GWAS trait of interest. Here, we chose an endothelial cell model as particularly relevant to the genetics of coronary artery disease, because: 1) endothelial cells play several key roles that are directly relevant to coronary artery atherosclerosis that leads to CAD, including: control of cholesterol influx from the blood, control of immune cell recruitment, and regulation of smooth muscle cell functions through release of vasoactive molecules, such as EDN1 and nitric oxide [13,14,47], 2) previous studies

have demonstrated strong enrichment of CAD heritability in endothelial cells (e.g. 131), and 3) detailed studies of individual CAD GWAS loci have identified likely causal genes that are clearly related to endothelial cell functions, including NOS3[132], EDN1[14], JCAD[133], ARHGEF26[134], PLPP3[135], and AIDA[136].

We chose telomerase immortalized human aortic endothelial cells (teloHAEC) for these studies, because, while immortalized, they maintain important in vitro EC functions such as tubing, lipid transport and response to inflammatory stimuli 21,137. We confirmed that teloHAEC enhancers were enriched for heritability for CAD using S-LDSC (described below). We also compared their gene expression profiles to those of primary coronary artery endothelial cells in vivo 69 and found that genes near GWAS signals were similarly expressed (also described below). We expect that similar analyses will be useful for future applications of the V2G2P approach.

Notably, although we conducted our Perturb-seq experiment in resting, unstimulated conditions, we identified several programs related to specific stimulus responses. These included non-cell type-specific programs for unfolded protein response (UPR), DNA damage, heat shock, and inflammation, as well as endothelial-specific programs such as flow response and the endothelial to mesenchymal transition (endMT). Thus, knocking down genes with Perturb-seq in resting cells can, nonetheless, reveal gene programs relevant to various stimuli that may be informative for understanding disease mechanisms. It remains possible, however, that prioritization of certain disease-associated programs will require specific atherogenic stimuli (e.g. inflammatory cytokines or oxidized LDL), and further studies will be required to test this possibility.

9 **Use of S-LDSC to test for enrichment of variants in enhancers in the chosen model.** We used S-LDSC to estimate the enrichment of CAD heritability linked to enhancers in TeloHAEC. While the original implementations of S-LDSC linked variants to genes based on genomic distance[28,70], we additionally required that variants either overlap exonic regions of the gene or overlap nearby candidate enhancers in endothelial cells (as in 32,71). To estimate the enrichment of CAD heritability in TeloHAEC enhancers, we required the variants to overlap enhancers predicted by ABC from ATAC-seq and H3K27ac ChIP-seq data in TeloHAEC under control conditions or treated with IL1β, TNFα or VEGF (ABC score > 0.015). For each set of variants (programs or TeloHAEC enhancers) we ran S-LDSC using 1000G EUR Phase3 genotype data to estimate LD scores, baseline v2.2 annotations as recommended by the LDSC developers[73], and HapMap 3 SNPs excluding the MHC region as regression SNPs. We ranked programs by their enrichments and reported the p-values of these enrichments.

10 **Comparison to the same cell type in animal models of disease.** To confirm the validity of teloHAEC as a relevant model for endothelial cells in human coronary artery (where atherosclerosis that leads to CAD develops), we compared single cell RNA-seq gene expression from control guide carrying teloHAEC from our Perturb-seq screen to scRNAseq data from explanted human right coronary artery endothelial cells (RCAECs)[69]. We compared the gene expression at two levels: for all perturbed genes (2,285 genes) and for the 41 CAD associated genes. Among the perturbed genes in teloHAEC, 2,107 genes are expressed at TPM > 1 in healthy or disease RCAECs. We observed high correlation of gene expression in transcripts per million (TPM) between teloHAECs and RCAECs (Pearson correlation = 0.66, p-value = 6.45 x 10-280).

After completion of the V2G2P enrichment test (described below), and identification of V2G2P genes, we observed similar correlations of gene expression for the 41 CAD associated genes (Pearson correlation

= 0.63, p-value = 9.29 x 10-6). Furthermore, 40 out of 41 CAD associated genes are expressed at >1 TPM in RCAECs.

Generation of an inducible CRISPRi cell line

11 To create the TeloHAEC CRISPRi line, cells were transduced with lentiviral vectors containing 1) dox-inducible (tetracycline operator controlled) dCas9-KRAB-BFP (CRISPRi machinery, which targets epigenetic repressors to efficiently silence enhancers or promoters[53–55], Addgene #85449) and 2) rtTA (tetracycline activator) with a hygromycin marker (Addgene #66810).
Perform hygromycin selection (250 µg/ml for 4 days), then treat cells with 1 µg/ml doxycycline (dox, a stable tetracycline analogue) for 3 days.
FACS sort dox-treated cells and select the top 15% of BFP positive cells. After a period in culture without dox, treat cells for 3 days with dox and and re-sort. CRISPRi TeloHAEC were passaged for routine maintenance in the absence of dox.

12 Diagnostic FACS should show no leaky BFP expression in the absence of dox, and >90% BFP positive cells in the presence of dox. It is a good idea to perform diagnostic FACS with a sample of cells after expansion and dox induction for the "Perturb-seq library preparation" step, below, to confirm that these manipulations haven't reduced the efficacy of doxycycline control of the CRISPRi machinery.

Identification of gene targets for perturbation

13 We constructed a library of promoter-targeted CRISPRi guides to all potential causal CAD genes. First, we identified all coding genes within a 1 megabase window surrounding the lead SNPs from CAD loci identified in either or both of van der Harst et al.[10] and Aragam et al.[12] that were expressed in TeloHAEC (1+ TPM, from bulk RNA-seq). If fewer than 2 expressed genes were found within 500kb up- or downstream of the lead SNP, the window was expanded to include the closest 2 genes to each side (for a total of 1661 genes). Non-coding genes were generally excluded, unless there was strong evidence for regulatory functions, particularly in ECs. Selected genes with TPM <1 were included, particularly if they were known to be important for CAD in tissues where they were more highly expressed (e.g. PCSK9), or were regulated by IL1-beta in bulk RNA-seq data in TeloHAEC (FDR<0.05, fold change >1.3). As negative controls, we included guides targeting 48 coding genes expressed in other cell types but not detectably expressed in ECs, and the 132 expressed coding genes within 1 Mb of 16 randomly-selected lead SNPs associated with Inflammatory bowel disease, Crohn's disease or Ulcerative colitis [59], and which did not overlap with CAD loci. As positive controls, and to aid in connecting candidate CAD genes to known pathways in ECs, we targeted the promoters of an additional 284 genes with known roles in a wide range of CAD-relevant EC functions such as barrier formation, TGF-beta signaling and inflammation, as well as major classes of expressed transcription factors and common essential genes. We also targeted an additional 160 promoters of expressed genes predicted to be regulated by EC enhancers containing fine-mapped variants associated with other disease phenotypes expected to be modulated by ECs (migraine, blood clotting in leg, systolic blood pressure, diastolic blood pressure & mean arterial blood pressure, from UKBB). This gave a total of 2285 genes, some of which were members of more than one category.

Preparation of lentiviral CRISPRi gRNA library

14 sgRNA guides were designed to target promoters of the chosen CAD and control genes (15 guides spanning from -150 to +100 relative to the Transcription Start Site (TSS)), using our established pipeline ([9,22], https://github.com/EngreitzLab/CRISPRDesigner). We included 400 non-targeting guides (that do

not have close matches to any region in the human genome) and 600 safe targeting guides (targeting non-genic regions lacking enhancer marks) 53. Because TeloHAEC are puromycin resistant, we adapted the CROP-opti vector (20, Addgene, #106280) for Blasticidin resistance ("CROP-opti Blast"), by digesting the vector with BsiWI and MluI, PCR-amplifying the Blasticidin resistance gene from lenti-dCas-VP64_Blast (Addgene, #61425) with added homology arms, and performing Gibson Assembly (Gibson Master mix, New England Biolabs). To create "CROP-opti-BC-Blast", we added HyPR-Seq barcodes between the WPRE element and the U6 promoter of CROP-opti-Blast, as described in 60. A pool of oligos encoding the guide sequences, plus extensions with homology to the U6 promoter and downstream scaffold (TATCTTGTGGAAAGGACGAAACACCG & GTTTAAGAGCTATGCTGGAAACAGCATAG) was synthesized by Agilent Technologies, and cloned into Crop-Opti-BC-Blast by Gibson assembly and bacterial electroporation as described53, at an average coverage of 202 transformants per guide. Note that, since the vector was prepared from a single clone, diversity of the HyPR-seq barcodes (which were not required for Perturb-seq) was not preserved. The library was sequenced and shown to include all 37,637 designed guides with relatively equal coverage of each (the difference in count frequency between the top and bottom 10th percentiles of guides was 2.8). A lentiviral library was produced using a standard 3-plasmid protocol53, at a scale to yield 10 ml of virus, stored in aliquots at -80°C, with each aliquot thawed only once.

Perturb-seq library preparation

15 To transduce this library into CRISPRi TeloHAEC, cells were resuspended in media containing 10 µg/ml polybrene at a density of 1e6 cells per ml, mixed with virus and plated 4 ml per well to 6-well plates, centrifuged at 2000 rpm for 2 hrs at 30oC, and incubated at 37oC for 2 hrs before addition of another 4 ml media without polybrene. The next day, cells were harvested and plated to 15 cm plates and treated with 15 µg/ml blasticidin for 4 days. The effective viral titre was determined using this same protocol, and a volume of virus was chosen that gave a final measured 15.7% infection rate (such that most successfully transduced cells have only 1 guideRNA). For the Perturb-seq study, 127.5 million CRISPRi TeloHAEC were transduced and selected for blasticidin resistance, for a coverage of approximately 360 cells per guide (as back-calculated from yield at the first post-blasticidin split, using the 36.7 hr doubling time observed in routine culture) to 461 cells per guide (as estimated from initial number of cells and infection rate). After blasticidin selection, cells were treated with 2 µg/ml dox for 5 days (plating 18e6 cells at each split, to maintain complexity of the library). We reasoned that, since atherosclerotic plaques develop slowly, the longer-term transcriptional effects of causal CAD gene disruption would provide the greatest insights into disease mechanisms. Thus, while we have found that knock down of guide-targeted genes is near maximal after 2 days of doxycycline treatment (inducing the CRISPRi machinery), we treated guide-containing cells with 2 µg/ml doxycycline for 5 days, to measure the longer-term consequences of each perturbation.

The presence of guideRNAs in cells allows multiplets (droplets containing 2 or more cells) to be unambiguously identified, as droplets containing more than one guide. This allowed us to load ~10-fold more cells per 10X Genomics lane than the maximum number recommended in the manufacturer's protocol. Briefly, cells were harvested, resuspended in PBS with 1% BSA, counted, and loaded at 150,000 cells per lane on a 10X Genomics Chromium Controller using a 3' scRNA-seq V3 kit (20 lanes, for a total of 3 million cells). Cells were isolated in two batches, with 6 lanes for the first batch, and 14 lanes, across 2 cassettes, for the 2nd batch, 6 hours later. scRNA-seq libraries were generated using the 10X Genomics protocol, and given lane-specific indexes.

Dial out library preparation

16 From the initial amplified cDNA, we used a two stage PCR protocol to generate “dialout” libraries, for each lane. Because the CROP-seq vector expresses a Pol II polyadenylated transcript that ends just downstream of the guide sequence, the dialout libraries identify the guideRNA sequences associated with each droplet20. PCR1 oligos for the guide dialout PCR were: CTACACGACGCTCTTCCGATCT & GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGTGGAAAGGACGAAACACC, and PCR2 oligos were AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC & CAAGCAGAAGACGGCATACGAGAT-8bp index sequence-GTGACTGGAGTTCAG.

Identification of scRNAseq cells and assignment to guides

17 To get complete information about guide assignments, dialout libraries were sequenced to approximately 40-fold saturation. Guides were identified from read 1 sequences, using Bowtie2 to align dialout reads to a “genome” composed of all 37,637 guide sequences, requiring no-mismatches. Aligning read 1 and read 2 sequences linked gRNA sequences with cell barcodes (CBCs, unique to each bead/droplet) and unique molecular identifiers (UMIs). To avoid low-frequency PCR chimeras, we required that each CBC-UMI-guide combination be duplicated at least 4 times. We then identified the guides associated with each CBC, and the number of different UMIs for each CBC-guide combination. We selected 4 UMIs for any single guide as the threshold to call a cell as containing a guide. We defined singlets (one cell & one guide per CBC) as having ≥4 UMIs for the most frequent guide and ≥4x less than this for the 2nd most frequent guide (choosing these thresholds to give a good balance between power to detect transcriptional effects and accuracy in measuring the magnitude of these effects, as described under Selection of Singlet Thresholds, below). Doublets and higher multimers, were cells with ≥4 UMIs for the top guide, and one or more additional guides with more than 1/4 this number of UMIs.

Note: The CROPseq PolII transcript that reads through the guide cassette can be expressed at different levels in different cell lines. This expression was borderline OK in TeloHAEC, but may be too low to effectively assign guides in other cell lines. If so, one alternative is to use the "Feature barcoding" approache offered by 10X genomics.

18 **Considerations for selection of singlet thresholds**

Expression of the CROP-seq guide mRNA in TeloHAEC is lower than in some other cell lines, such as K562 & HEK293T12 resulting in the absence of a clear gap between noise (low UMI CBC-guide combinations that are likely PCR chimeras) and higher UMI-count true guide reads (**Extended Data Fig. 1h**). We hypothesized that reducing stringency for singlet calls could potentially reduce power to detect perturbation effects on transcription (due to increased noise from mis-calling some true doublets as singlets), or could increase power (by increasing the total number of called singlets analyzed). To test which of these was true, we measured the correlation between differential expression calls for cells with guides to a given target in the full Perturb-seq library versus a smaller pilot library tested in resting TeloHAEC, reasoning that parameters that improved the correlation between these separate studies would also increase the power of the full scale library to detect real transcriptional effects. Information about guides, as well as raw and processed data for the “200 gene” pilot library can be found on GEO, with accession number GSE212396. For the pilot library, we chose singlets with the very stringent threshold of 6 UMIs for the top guide and more than 5-fold less for the next most frequent guide (“6& <5x”). For the full Perturb-seq dataset we chose 4 UMIs for the top guide and equal to or more than 4-

fold less than the next most frequent guide ("4&<=4x", our final applied standard, yielding 214,449 singlets), or the relaxed thresholds "3&<=3x" (284,466 singlets) and "2&<=2x" (389,792 singlets). We identified 37 gene targets that were shared between libraries, and which also showed an FDR<0.1 effect on the transcriptome in the full Perturb-seq 4&<=4x dataset (measured as described above). We then ran EdgeR[58,62] for differential expression testing (cells with guides to each of these 37 targets versus cells with control guides), for each library and singlet definition (pilot 6&<5x, or full library 4&<=4x, 3&<=3x, and 2&<=2x). Then, for all genes called as differentially expressed in either the pilot library or the full library (raw p-value < 0.01), we measured the correlation in log2 fold changes between the pilot & full scale data, repeating this analysis for each singlet definition.

Lastly, we measured the difference in correlation coefficients (R) between the relaxed threshold comparisons (pilot v. full library 3&<=3x, and pilot v. full library 2&<=2x) and the base comparison (pilot vs. full library 4&<=4x). We found that the median correlation between pilot & full-scale studies significantly improved with the relaxed singlet thresholds (with significance assessed by two-sided t-test). This indicates that the increased number of called singlets with the relaxed thresholds increased the power to detect real transcriptional effects, despite an expected increase in doublets mis-assigned as singlets. Plotting change in R for each target for the 2&<=2x singlet definition ((R for pilot v. full library 2&<=2x) - (R for pilot v. full library 4&<=4x), y-axis) against the R value for the base correlation (between the pilot and the 4&<=4x full library singlet definition, x-axis), we found that in all 13 cases where R started high (>0.15, likely real correlations between strong transcriptional effects), R increased. R also increased in all but one case where it started out negative (correcting anti-correlations likely driven by noise). Weak positive base correlations were adjusted up or down, potentially improving true correlations and correcting spurious ones. As such, relaxed singlet thresholds might improve power to detect reproducible transcriptional changes more than is indicated by simple mean differences in R values. On the other hand, we found that lower stringencies reduced the apparent knock down effect on these target genes, themselves (median log2 fold changes: -0.53 for 4&<=4x, -0.41 for 3&<=3x and -0.42 for 2&<=2x), likely due to the fact that a mis-called singlet that was actually 2 cells with different guides would show half-magnitude transcriptional effects of each guide. Reduced singlet thresholds also decreased median log2-fold changes for target genes across all targets in the full-scale library (-0.368 for the 4&<=4x singlet definition, and -0.327 for the 2&<=2x singlet definition). Based on these observations, we chose the thresholds of 4 UMIs for the top guide and <=¼ this for the next (4&<=4x), to provide a good balance between overall power and accurate detection of the magnitude of effects.

scRNAseq library sequencing and assignment of reads to cells

19 scRNA-seq libraries were sequenced on two Illumina NovaSeq S4 flowcells, yielding 20,245,734,673 total reads, across all 20 libraries. The FASTQ files were processed on the 10X Cloud to run CellRanger count with the hg38 reference genome. We used the "filtered" features (i.e., cell barcodes corresponding to droplets that contain a cell), and combined the outputs from all twenty 10X lanes into a single genes x cell matrix. This analysis identified 822,156 cell-containing droplets . To measure the effects of individual guides on individual cells, we selected only those CBCs identified in the dialout analysis as corresponding to singlet cells. This identified 214,449 singlets (droplets containing one cell and one guide), defined as 4+ unique molecular identifiers (UMIs) for the top guide and ≥4-fold fewer UMIs for any other guide. This gave an average of 5.7 cells per guide and 85.5 cells per target promoter. Average sequencing depth was 10,870 transcriptome-mapped UMIs per singlet cell, and 929,000 transcript UMIs, across all 15 guides, for each target promoter.

Estimation of fitness effects of guides

20 To estimate the fitness effects of guides, we compared the relative frequency of all 15 guides to a given target in the original library to the frequency of the same guides in singlet cells, and estimated significance by Benjamini-Hochberg adjusted binomial tests. Essential genes were defined as those that scored as fitness reducing in 5 of 7 tested lines in 61.

Measurement of individual knock down efficacy and effects o…

21 To measure the differential effects of guides to specific target promoters on individual genes, we used edgeR58, with settings for scRNA-seq from 62, comparing all singlet cells with guides to each target to all singlet cells with any of the 1,000 non-targeting and safe targeting guides. Genes with fewer than 10 UMI counts across all singlet cells were excluded from the analysis. To control for possible batch effects, we included the 10X lane number as a covariate. For average knockdown efficacy for each perturbation (across all 15 guides), we used the log2 fold change and p-values reported by edgeR. To measure the knockdown efficacy of individual guides, we performed binomial tests on: the number of transcripts for the guide's target in singlet cells with that guide (hits), all transcripts in singlets with that guide (tests) versus a background frequency of (transcripts to the target in other singlet cells)/(all transcripts in other singlet cells). Note that with an average of 5.7 cells per guide, assigning significance for knockdown effects of individual guides was only possible for genes with high expression in unperturbed cells (e.g., TPM>100). To identify perturbations with a significant effect on the transcriptome, we used the edgeR results for the 48 negative control promoters (for genes not detectably expressed in TeloHAEC) to estimate the number of DE genes found by chance, at thresholds of nominal p-value < 0.01 and fold change > 1.15. Perturbations with a significant effect on the transcriptome (across all 15 guides to each target) were defined as having more DE genes, by these same thresholds, than the 48 non-expressed controls (using binomial tests with a background rate equal to the average DE gene count for controls over all genes tested, and multiple hypothesis correction by the Benjamini Hochberg method).

Consensus non-negative matrix factorization analysis of scRN…

22 **Introduction to cNMF and considerations for perturb-seq analysis**
To identify sets of genes that are co-expressed across single cells in a dataset, we used non-negative matrix factorization (NMF). NMF decomposes an input cell x gene UMI count matrix (X) into a cell x component matrix (W) and a component x gene matrix (H), such that X = W • H + E, where E is the error term. The cell x component matrix W represents the contribution of each component to the cell's transcriptional profile, and the component x gene matrix H encodes information about gene expression programs. The number of components (K) is a hyperparameter defined prior to performing matrix factorization (see below). To account for the fact that the NMF algorithm is a stochastic algorithm that depends on the initial seed, we used the consensus NMF (cNMF) method developed by Kotliar et al25. The cNMF method, after normalizing each gene's expression to unit standard deviation, factorizes the normalized matrix multiple times (here, 100 repeats); clusters the components from the repeat runs based on their pairwise Euclidean distances; removes the components that show low similarity to any other component (here, threshold on Euclidean distance = 0.2); defines "consensus components" as the median of each of the component clusters; and recomputes the cell x component matrix W using these

consensus components. As one technical note about applying the cNMF pipeline as described by Kotliar et al.[25], we found that including all genes, as opposed to the 2000 most variable genes, was important for finding certain programs observed only infrequently in the dataset (data not shown). This is because genes whose expression changes in only a small fraction of cells (e.g. cells with a particular perturbation) would not end up being included in the 2000 most variable genes.

23 **Data pre-processing**

To remove noncoding RNA from the analysis, we removed genes with names starting with “LINC” and gene names with patterns starting with two letters and six digits. We retained cells with a minimum of 200 unique detected genes and a minimum of 200 UMIs. We retained genes detected in a minimum of 10 cells.

24 **Running cNMF and defining "co-regulated gene sets"**

We have published a pipeline for running cNMF on Perturb-seq data that can be found here: https://github.com/EngreitzLab/cNMF_pipeline/ (DOI: 10.5281/zenodo.10357454) (G2P and V2G2P enrichment).

We defined ‘co-regulated genes’ for each cNMF component as the 300 marker genes with the highest z-score regression coefficient as defined by cNMF[25]. Essentially, cNMF uses a linear regression model to identify coefficients indicating the number of standard deviations each gene’s expression would change with the increased usage of a given component. A component’s marker genes, then, are those with the highest “marker gene regression coefficients” (or “specificity scores”) for that component, and we selected the top 300 of these marker genes as the set of “**co-regulated genes**” for each gene expression “program”.

In addition to descriptions of workflow steps on github, important considerations for cNMF analysis (choosing the number of components, accounting for batch effects, etc.), are also described below.

25 **Choosing the number of components for cNMF analysis**

To choose the free parameter K (number of components), we defined a set of benchmarking statistics and compared the results of cNMF run for K = [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 19, 21, 23, 25, 27, 29, 30, 35, 40, 45, 50, 55, 60, 100]. We ultimately chose K=60 for all downstream analyses.

We examined the following benchmarking statistics (each described in the sections below):
(i) Number of unique GO terms enriched in program co-regulated genes (see below)
(ii) Number of unique enriched TF motifs in the promoters or enhancers of program co-regulated genes (see below)
(iii) Number of perturbations significantly regulating any component (see below)
(iv) Error of cNMF (difference between the original normalized data and reconstructed data, calculated by taking the sum of squares of the element-wise difference of the data
(v) Stability of cNMF (a measure of consistency of the components output from repeated runs, represented by the silhouette score[25])
We chose K = 60 for further analysis, as the number of components that gave a low cNMF error value while near-maximizing each other metric.

26 **Determining enrichment of annotated gene sets in components.** To determine if the gene expression programs align with annotated and publicly available pathways, we tested whether the co-regulated genes in each component were enriched in gene sets from the Molecular Signatures Database (MSigDB). To do so, we used the clusterProfiler R package[66] and MSigDB gene sets[67] (here, the gene sets labeled

as “all” for all gene sets and ”c5” for GO terms only). We filtered the MSigDB gene sets to only those with more than 3 genes and less than 800 genes. We annotated each program with the gene sets that showed significant enrichment among the program genes (FDR < 0.05), and compared the number of gene sets showing significant enrichment as a function of the number of programs K.

27 Identifying motifs enriched in promoters and enhancers

To identify transcription factors that might regulate program co-regulated genes, we calculated enrichment of human transcription factor motifs in the sequences of the promoter and enhancers of the top 300 genes ranked by component specificity score.

We obtained promoter sequences by taking 500 bp surrounding the TSS as previously annotated[22] and enhancer regions from the Activity by Contact model at an ABC score threshold of 0.015 in the TeloHAEC control condition. For a gene that had multiple enhancers, we counted motif instances across all of its enhancers. To match motifs to sequences, we used HOCOMOCO v11 human full scan motifs (https://hocomoco11.autosome.ru/downloads_v11), and Find Individual Motif Occurrences (FIMO) (https://meme-suite.org/meme/meme_5.3.2/tools/fimo), with the default settings, and p-value thresholds of 10-6 for enhancers or 10-4 for promoters.

For a given motif and a given program, we counted the number of occurrences of a motif in the promoter sequences of either (i) the top 300 program co-regulated genes, or (ii) all expressed genes in teloHAEC, and compared these two vectors of motif counts using a two-sided t-test. We computed enrichment by dividing the program gene’s average motif match count by the rest of the expressed gene’s average motif match count. We tested all pairs of matched motifs (570 for promoter and 590 for enhancer) x 60 programs, and used the Benjamini-Hochberg method to account for multiple hypothesis testing on the t-test p-values.

28 Excluding components associated with batch effects

We examined whether some components identified by cNMF were likely to represent batch effects. To do so, we calculated the Pearson correlation between each of the 20 batches (i.e., 10X lanes) and the expression of each component across all cells. Based on the distribution of batch x program Pearson correlation, we assigned 10 components with Pearson correlation > 0.15 as likely representing batch effects. We chose the threshold of 0.15 because: 1) most components showed very low correlation with batch, 2) above this threshold sample correlation with batch increases greatly (indicating particularly strong association with batch for these 10 components), 3) 8 of 10 of these components were associated with mitochondrial or ribosomal genes - sets of genes commonly observed as batch effects in 10x scRNAseq, 4) these components showed abs(R)>.15 across multiple 10X lanes (average of 9), and 5) importantly, none of these components showed enrichment by the V2G2P test (described below), and so would not have been identified as significant for CAD risk. We used the remaining 50 components for further analysis. This approach (including batch as a covariate in the differential expression test) has theoretical advantages, in particular reducing bias when groups (here, perturbed genes) are not distributed evenly across batches[63].

29 Defining endothelial-cell-specific programs

To annotate programs as “endothelial-cell-specific”, we analyzed the degree to which program co-regulated genes were expressed in endothelial cells versus other cell types. We took gene expression transcript per million (TPM) data across all available cell types from FANTOM5 and calculated the expression z-score of each gene across all cell types. To give each gene an endothelial-cell specificity score, we calculated the average of all z-scores for a gene across endothelial cell samples. We defined endothelial-cell specificity scores for each program as the average of the 300 co-regulated genes' specificity scores, and selected 0.19 (90% percentile) as the threshold to call programs as "endothelial-cell-specific".

30 **Calculating Variance explained by all cNMF components**

To quantify the fraction of variance explained by all 60 programs jointly, we compared the residual variance in the dataset after subtracting the consensus matrix factorization to the total variance in the dataset: $V = 1 - Var(X - WH)/Var(X)$, where X is the (cell x gene) normalized data matrix input to cNMF, W is the (cell x program) usage matrix, H is the (program x gene) spectra or weight matrix, and matrix variance is defined by summing the column- or gene-level
variances: $Var(X) = \Sigma jVar(Xj)$. Note that cNMF normalizes the input data so each $Var(Xj) = 1$.

Calculating Variance explained by individual gene programs

To rank gene programs by variance explained, we devised a method to quantify variance explained by NMF or cNMF components separately. For the k'th program Hk, we consider the effective matrix decomposition given only this program; the effective usage matrix Bk in this case is given simply by orthogonal projection or ordinary least squares: $Bk = XH'k/||Hk||2$, where the prime indicates transposition. We then define the variance explained in terms of the residual fraction as above: $Vk = 1 - Var(X - BkHk)/Var(X)$. Our method may be generalized to any set of programs, but with more than one program the effective usage matrix must be obtained by nonnegative least squares (a single iteration of NMF).

Identification and annotation of transcriptional programs

31 **Defining regulators for each cNMF component**

We tested whether gRNAs targeting a given gene led to a significant change in expression of each component from the cNMF model. We used the Model-based Analysis of Single Cell Transcriptomics package (MAST)[50] to compare the expression of each component in cells carrying gRNAs targeting a given gene vs. cells carrying control gRNAs (1,000 safe-targeting and negative control guides), including 10X lane as a covariate to account for batch effects. We removed the guides present in fewer than 3 singlet cells and the perturbations with fewer than 2 guides. We used the Benjamini-Hochberg method to account for multiple hypothesis testing on the MAST p-values, and assigned '**regulators**' of a component as those genes whose perturbation affected component expression with FDR < 0.05 accounting for 140,760 total tests

.

To confirm that these FDRs were well-calibrated, we also conducted a simulation-based test. For each perturbed gene, we sampled from the control cells (all singlet cells with non-targeting or safe-targeting guides) the same number of cells, and compared these sampled cells to the rest of the control cells using the same MAST[50] procedure. We identified 0 significant regulators in this approach, indicating that our FDR < 0.05 threshold is a conservative estimate. We also performed the same procedure to estimate the background rate for perturbations called as having a significant effect on the transcriptome using EdgeR[58,64,65].

32 **Definition and annotation of gene expression programs**

We defined a gene expression **program** as the set of genes comprised of both the 300 "**co-regulated**

genes” and the significant “**regulators**” for each cNMF component. We annotated programs based on features of their co-regulated genes and regulators, including: by manual curation of genes with known biological functions, by enrichment of transcription factor (TF) motifs in the promoters and predicted enhancers of co-regulated genes, and by GO term enrichment (see below). The manual labels we assigned to each program (e.g., “Program 8 – Angiogenesis and osmoregulation”) represent our attempt to annotate the program based on functions of known genes, but we note each program includes many genes and regulators that have not been studied in combination before and may represent new, specific gene-pathway relationships that we do not currently have the vocabulary to describe.

Variant to Gene to Program (V2G2P) enrichment analysis.

33 **Identifying CAD-associated programs via variant-to-gene-to-program analysis**

We developed an approach to identify gene programs likely to affect CAD risk through functions in endothelial cells. To do so, we tested whether the 254 genes with V2G links (between CAD variants and enhancers/coding regions in endothelial cells) were enriched in each Perturb-seq program. Specifically, we performed a one-tailed Fisher exact test separately for co-regulated genes and for regulators. For co-regulated genes, we constructed a contingency table for whether a gene is a co-regulated gene (out of 17,472 expressed genes) and whether a gene has a V2G link. For regulators, we constructed a contingency table for whether a gene is a regulator (out of all perturbed genes) and whether a gene has a V2G link. We then multiplied the p-values from co-regulated gene and regulator Fisher exact tests together to get a final program enrichment p-value. We use Benjamini-Hochberg method for multiple hypothesis correction across all 50 non-batch programs. 5 programs showed significant enrichment by this method (FDR < 0.05: Programs 8, 35, 39, 47, 48), referred to as “**V2G2P programs for CAD**”.

Defining CAD-associated V2G2P genes

We defined “**V2G2P genes for CAD**” as those 41 genes that were both (i) a gene with a V2G link to a CAD variant and (ii) a member of one of the 5 CAD-associated programs (as a regulator and/or co-expressed gene). The 41 genes were linked to 43 GWAS signals due to cases where independent GWAS signals are linked to the same gene.

Identifying and annotating convergent programs. We found that many of the same perturbations affected the 5 V2G2P programs, and that many of these perturbed genes had been previously reported to interact physically or functionally with the CCM complex and/or downstream ERK5/MEK5 signaling[46,51,74–78], plus one additional gene (TLNRD1) that we identify here as a member of the CCM pathway. These genes were manually selected through an iterative process involving examining genes known to interact with the CCM complex and that were found to regulate the enriched programs in Perturb-seq.

Methods to test the association of V2G2P genes and program...

34 **Identifying enriched programs via MAGMA**

We tested whether the co-regulated genes in each program were significantly enriched near variants associated with CAD using MAGMA[2]. To do so, we took the CAD summary statistics from Aragam et al.[12] (https://data.mendeley.com/public-files/datasets/2zdd47c94h/files/5b4eb0d7-96e8-4c7e-b109-046107ebd480/file_downloaded), and used the MAGMA --annotate function to summarize CAD association p-values for variants within a 50 kb window of all human genes, using the 1000 genomes

European reference data for base allele frequencies (https://ctg.cncr.nl/software/MAGMA/ref_data/g1000_eur.zip). We then ran MAGMA to test for enrichment of CAD heritability within 50 kb of the top 300 program genes, and corrected for multiple testing (60 components) using the Benjamini-Hochberg method.

35 **Identifying programs enriched for CAD heritability via stratified LD score regression**

We used S-LDSC to estimate the enrichment of CAD heritability linked to program genes and to enhancers in TeloHAEC. While the original implementations of S-LDSC linked variants to genes based on genomic distance[28,70], we additionally required that variants either overlap exonic regions of the gene or overlap nearby candidate enhancers in endothelial cells (as in [32,71]). In particular, for co-regulated genes in each program, we derived an annotation for S-LDSC by including exonic regions (exons from transcripts with Ensembl_canonical, appris_principal, appris_candidate, or appris_candidate_longest tags, as indicated in the GENCODE v38lift37 annotations) as well as endothelial cis-regulatory elements derived from snATAC-seq[72], from which we merged the 9 adult and 8 fetal sets of endothelial peaks into a single annotation, and for each geneset included all peaks within 50 kb of the gene starts and ends. For all peaks, we first converted coordinates from the GRCh38 to the GRCh37 reference assembly using UCSC LiftOver, discarding peaks that could not be converted. We ran S-LDSC using 1000G EUR Phase3 genotype data to estimate LD scores, baseline v2.2 annotations as recommended by the LDSC developers[73], and HapMap 3 SNPs excluding the MHC region as regression SNPs. We ranked programs by their enrichments and reported the p-values of these enrichments.

36 **Polygenic Priority Score (PoPS)**

PoPS is a method to nominate likely causal genes in a GWAS locus, which prioritizes genes based on their being members of many gene sets enriched for heritability genome-wide[3]. We applied PoPS to summary statistics from Aragam et al.[12] using the predefined set of gene sets as previously described[3]. For each GWAS signal, we calculated the PoPS rank among “nearby genes” (2 to either side of the lead SNP, and all within +/-500kb). Previously we have shown that genes with the highest PoP score in the locus are strongly enriched for likely causal genes, as identified by fine-mapped coding variants[3], and that this enrichment increases when further focusing on genes that are both the closest gene and have the highest PoP score. In this analysis, we did not use any features from Perturb-seq and, as such, this method represents an entirely independent method that validates the high likelihood of causality of the set of CAD-associated V2G2P genes.

Analyses supporting the power of combining V2G and G2P inf...

37 In combining V2G and G2P maps, we made several observations that help to explain the ability of V2G2P to identify disease-associated programs and genes:

(i) The intersection of V2G and G2P maps in endothelial cells was important for the identification of disease-associated programs related to endothelial functions. For 195 of 228 non-lipid signals, gene-to-program links identified more than 1 nearby gene (and up to 25), spanning all 50 programs—consistent with the notion that, by chance, a GWAS signal will have multiple nearby genes in various housekeeping and/or EC-specific programs. Statistical tests for enrichment of V2G linked-genes in programs, however, identified only 5 V2G2P CAD-associated programs for CAD, all of which were endothelial cell-specific.

(ii) The intersection of V2G maps with CAD-associated programs was important to identify CAD-associated V2G2P genes and nominate single causal genes associated with GWAS signals. Among the 125 signals that had at least 1 V2G link, 119 signals were linked to more than 1 gene (and up to 5)—in large part due to noncoding variants being predicted to regulate more than one gene, consistent with previous observations[9,33]. By contrast, of the 43 signals associated with CAD-associated V2G2P genes (V2G-linked genes in CAD-associated programs), only 6 had more than one such gene (up to 2). For example, the intersection of V2G links and G2P links to CAD-associated programs reduced the number of likely causal genes at 20p13.1, 10p24.33 & 17q21.3 GWAS signals, where V2G and/or G2P links, individually, predicted multiple possible genes. We conclude that the V2G2P approach substantially refined the list of candidate disease genes compared to using V2G or G2P approaches alone.

(iii) Including epigenetic data from multiple endothelial cell states was important for linking variants to genes. Here, we used ABC maps from various endothelial cell samples, including resting and stimulated conditions for teloHAEC, to catch many possible endothelial cell states where variants might act. Considering the 49 V2G links for the 41 CAD V2G2P genes: 15 were observed only in teloHAEC enhancers (including 8 identified in the resting, unstimulated teloHAEC state, and 7 solely identified in one or more of the 3 stimulated teloHAEC samples); 14 were observed in both teloHAEC and one of the other endothelial cell samples (HUVEC and eahy926, each resting or under various stimulation conditions); 12 were observed only in one of the other endothelial cell samples; and 8 were the result of coding variants.

(iv) The cell-type specificity of V2G links appeared to be important for identifying disease-associated programs. When we used a cell-type agnostic V2G approach in the V2G2P analysis (linking risk variants to the two closest genes, coding variant-containing genes, and two genes with strongest ABC links to enhancers in any cell type, as opposed to just endothelial cells), we found enrichment for 3 additional ubiquitous or non-endothelial cell specific processes: Program 5 (Interferon response), 36 (Steroid hormone response), and 37 (Redox homeostasis) . Similarly, when we used MAGMA, which links variants to genes based on a weighted function of distance, without regard to cell-type-specific information, we also found enrichment for additional programs corresponding to processes not specific to endothelial cells.

(v) Defining programs with Perturb-seq appeared to be important. In one baseline analysis, we applied cNMF to define programs based only on the unperturbed cells in the experiment (5,506 cells carrying negative control guides), and repeated the V2G2P analysis. We used cNMF to discover K=60 components, and defined 60 "control programs" based solely on the 300 co-regulated genes defining each component (because control guides did not target any genes, so there was no regulator information). We found none of the programs derived from unperturbed cells were significantly enriched, and the top program included only 10 genes with V2G links instead of 18 for the top program derived from the full Perturb-seq dataset. This suggests that the scale and/or perturbations present in the full Perturb-seq experiment were important for discovering disease-associated programs and genes. In a second analysis, we computed the V2G2P enrichment using only the co-regulated genes from the Perturb-seq programs (excluding the regulators in each program), and found only Program 8 and 39 to be significant (FDR < 0.05). Furthermore, most of the regulators of these programs, including CCM2, were not identified as CAD-associated V2G2P genes in this co-regulated gene-only analysis, because they were not co-expressed in these programs. This analysis supports that Perturb-seq was important for discovering genes and programs associated with CAD.

Altogether, our results indicate that cell-type specific variant-to-gene and gene-to-program maps can be combined to effectively prioritize disease-associated programs and genes.

Methods to validate Variant-to-Gene regulatory connections

38 **Allelic imbalance analysis for a variant linked to TLNRD1**

We calculated allelic imbalance in ATAC-seq and ChIP-seq signal for the rs1879454 variant, accounting for any mapping bias toward the reference allele following methods previously described79. Specifically, we created two reference genome FASTA files that harbored the reference or alternate alleles at rs1879454; aligned ATAC-seq data to both genome files; selected reads that overlapped the variant coordinate; and used PySuspenders79 and PySAM (https://github.com/pysam-developers/pysam) to assign and count reads that uniquely aligned to one or the other allele. We applied this procedure to ATAC-seq data from TeloHAEC and the ENCODE datasets ENCSR000EVW (GATA2 ChIP-seq on HUVEC) and ENCSR000EOB (DNase-seq and DGF on HMVEC-dLy-Neo).

39 **CRISPRi-FlowFISH for TLNRD1**

We used CRISPRi-FlowFISH to test the effects of 61 candidate enhancers on TLNRD1 expression in teloHAEC, including the enhancer containing rs1879454. We designed gRNAs tiling across all accessible regions (here, defined as the union of the peaks in the chromatin accessibility dataset called by MACS2 with a lenient P-value cut-off of 0.1, and 150-bp regions on either side of the MACS2 summit) in the range chr15:81,267,614-81,427,246 in ATAC-seq data from TeloHAEC. We excluded gRNAs with low specificity scores or low-complexity sequences as previously described 22. We infected teloHAECs with the gRNA lentiviral library with 15µg/mL blasticidin selection for 3 days, and activated CRISPRi with 2µg/mL doxycycline incubation for 5 days. We performed FlowFISH using ThermoFisher PrimeFlow (ThermoFisher 88-18005-210) as previously described 22, using ThermoFisher probesets VA1-3010837-PF for TLNRD1 and VA4-13187-PF for RPL13A. We observed an approximately 2.6-fold signal for TLNRD1 in cells with all probes applied ("stained") versus cells without target gene probes applied ("unstained"). We analyzed these data as previously described 22. In brief, we counted gRNAs in each bin using Bowtie to map reads to a custom index, normalized gRNA counts in each bin by library size, then used a maximum-likelihood estimation approach to compute the effect size for each gRNA. We used the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (implemented in the R stats4 package) to estimate the most likely log-normal distribution that would have produced the observed guide counts, and the effect size for each gRNA is the mean of its log-normal fit divided by the average of the means from all negative-control gRNAs. As previously described, we scaled the effect size of each gRNA in a screen linearly, so that the strongest 20-guide window at the TSS of the target gene has an 85% effect, in order to account for non-specific probe binding in the RNA FISH assay (this is based on our observation that promoter CRISPRi typically shows 80–90% knockdown by qPCR). We averaged the effect sizes of each gRNA across replicates and computed the effect size of an element as the average of all gRNAs targeting that element. We assessed significance using a two-sided t-test comparing the mean effect size of all gRNAs in a candidate element to all negative-control guides. We computed the false-discovery rate (FDR) for elements using the Benjamini–Hochberg procedure and used an FDR threshold of 0.05 to call significant regulatory effects.

Knock down of individual genes

40 **Introduction and rationale.**

Being able to knock down individual genes of interest can be useful at both early and later steps in the V2G2P process, including:

1) Validation of the efficacy of CRISPRi knockdown after BFP selection of your doxycycline-inducible CRISPRi line.

2) Validation of Perturb-seq transcriptional results, and comparison of Perturb-seq transcriptomic effects with those of related genes not tested in the original library.

3) Downstream phenotypic assays to connect transcriptional effects of V2G2P genes with relevant cellular phenotypes.

41 **Generation of single-guide CRISPRi TeloHAEC derivatives**

Paired oligos for individual guides (newly-designed, as described for the Perturb-seq library, or with the best KD efficacy in Perturb-seq) were annealed and cloned into the BsmBI site of a CROP-Opti-Blast vector (plasmid available upon request), which were then used to generate lentivirus (as per 53). CRISPRi TeloHAEC were infected with each virus, in separate wells, and selected for blasticidin (15 µg/ml 4 days), before 5 day dox induction and analysis by bulk RNA-seq, fluorescence imaging or physiological assays. Guides (TargetGene_CloneIndex: ForwardSequence) were: CCM2_C2: GGCAAGAAGGTGAGCGTGCG, CCM2_F6: GAGCCGCTACATGCTCGACCC, CDH5_B8: GCCAGCTGGAAAACCTGAAG, CDH5_D5: GTTGGACTGCCTGTCCGTCCA, ITGB1BP1_C7: GAAGGCCGCGGCACTCCCACG, ITGB1BP1_G8: GAAGTCCGCAACCCGGGGAT, KLF2_C9: GGACCCGGGGAGAAAGGACG, KLF2_G10: GCCGCGGTATATAAGCCGGC, MAP2K5_A11: GCCGAGGCCGCGCGGACTGG, MAP2K5_B5: GTCTGCCCCACCCGGAGACAC, MAP3K3_A4: GTTCCTGAGGTGGAGAACGG, MAP3K3_C3: GCCAATAACAAGAAGGAAGT, MEF2A_C10: GCGGCGCGAAGCGCTGGTGG, MEF2A_H10: GACTGAATTATCCTCTCGGT, Negative_control_B6: GCAACGGTGTACCGCGGATC, Negative_control_D2: GTGGTTCACAACCGGACCCA, Negative_control_D8: GGTGGTTCGGTTTGCGTGGCC, Negative_control_F4: GCTGGGCGGACGTTGGGATA, NFAT5_D4: GGCCTCGCTTCCTGCCGGCG, NFAT5_D7: GGTCCCCGTCCCGCCGGGGG, PDCD10_D11: GACCGAGCAGAAGAGGTCTA, PDCD10_G1: GCCGCTTTACGCCACTCGCGT, TLNRD1_B3: GTGGCTGCGCCGCCGCCCGCA, TLNRD1_D12: GCCTCCGGCAGCCCCTGCGGG.

42 **Ribonucleoprotein-based CRISPR/Cas9 genome editing**

For some experiments, we used Synthego's ribonucleoprotein (RNP) technology as an orthologous method to knock down target genes, as previously described 80. Briefly, TeloHAEC were nucleofected with Synthego's Gene Knockout Kit v2 for non-targeting negative control, CCM2, TLNRD1 or MAP3K3 using the Lonza 4D-Nucleofector system. For each nucleofection reaction, we used 150,000 cells with 20 pmol of Cas9 and 50 pmol of sgRNA. The cells were then nucleofected (program CA-210) using SG cell line nucleofection solution (Lonza; V4XC-3024). The nucleofected cells were seeded in TeloHAEC culture medium, and harvested 48 hrs later for RNA extraction for qRT-PCR analysis and/or RNAseq to measure gene knockdown efficiency and perturbation effects. For MAP3K3 knockdown in single-guide CRISPRi lines, cells were treated with 2 µg/ml doxycycline for 72 hours before nucleofection, and for the 48 hours afterwards.

Validating biochemical predictions from the V2G2P analysis

43 **Computational prediction of the TLNRD1 and CCM protein structure**

AlphaFold2.3 Multimer v3 81 was run using sequences for KRIT1 (UniProt O00522), CCM2 (Uniprot Q9BSQ5, with and without deletion of residues 417-444), PDCD10 (Uniprot Q9BUL8), and TLNRD1 (Uniprot Q9H1K6). Models were visualized using UCSF ChimeraX v1.61. Predicted Alignment Error (PAE) was extracted using AlphaPickle 82 and plotted using combinations of AlphaPickle, Matplotlib v3.7.0, and Seaborn.

Co-immunoprecipitation of CCM2 and TLNRD1

HEK293 cells were transfected with V5-tagged CCM2 full length, V5-tagged CCM2 C-terminal truncation, Flag-tagged TLNRD1 and/or Flag-tagged Akt1, using FuGENE (E2311, Promega) or PEI MAX (Polysciences). Two days after the transfection, cell lysates were extracted with IP lysis buffer (87787, Thermo Scientific) supplemented with 1x Halt Protease Inhibitor Cocktail (1862209, Thermo Scientific).

Protein concentration was determined using the Pierce BCA Assay (ThermoFisher), and equal mass of protein used for each sample. Immunoprecipitation was carried out using magnetic beads (88805, Thermo Scientific) conjugated with 5 µg of either rabbit anti-V5 (13202, Cell Signaling Technology) or mouse anti-Flag (F1804, Millipore Sigma) antibody. Cell lysates were incubated with the antibody-conjugated beads for 20 to 30 mins at room temperature. Beads were then washed three times with IP lysis buffer (1861603, Thermo Scientific), and precipitants were eluted using 2xLDS sample buffer (NP0007, Thermo Fisher Scientific). Precipitants and input lysates were separated by 10% SDS-PAGE and transblotted to nitrocellulose. For the anti-FLAG IP, blots were immunoblotted with 1:1000 rabbit anti-V5 (13202, Cell Signaling Technology), followed by 1: 5000 anti-rabbit HRP secondary (7074, Cell Signaling), then stripped (21059, Thermo Scientific) and re-probed with 1:1000 rabbit anti-TLNRD1 (HPA071766, Sigma). For the anti-V5 IP, blots were immunoblotted with 1:1000 primary mouse anti-Flag (F1804, Millipore Sigma) and 1:5000 secondary anti-mouse HRP (7076, Cell Signaling), and stripped and reprobed with 1:1000 mouse anti-V5 (ab27671, Abcam). The Akt1-FLAG vector is Addgene #9021. CCM2-V5 (ccsbBroad304_04281) and TLNRD1-V5 (ccsbBroad304_03872) vectors were obtained from the Broad Institute Gene Perturbation Platform83. For TLNRD1-FLAG, cDNA sequences were amplified from the TLNRD1-V5 vector using primers that incorporated an in-frame FLAG tag, and cloned into the pcDNA3.1 backbone. The CCM2 C-terminal truncation, was created in the CCM2-V5 vector by site-directed mutagenesis using the QuickChange II Site-Directed Mutagenesis kit (Agilent 200523-5), and the oligos F-CACCCTCAGAGGGGTCAGCATGCCCAAC and R-GTTGGGCATGCTGACCCCTCTGAGGGTG, which resulted in a deletion of amino acids 419-442 at the C-terminus of CCM2, in frame with the V5 tag.

Validating predictions for cellular phenotypes from the V2G2...

44 Trans-endothelial electrical resistance (TEER) measurements

For TEER measurements, we used the ECIS Z-Theta instrument from Applied BioPhysics in the 96-well plate system (Applied BioPhysics; 96W10idf). CRISPRi TeloHAEC expressing individual guides to TLNRD1, CCM2, or non-targeting guides (2 guides each) were treated for 5 days with 2 µg/ml doxycycline. A gold electrode-containing 96-well ECIS plate was incubated at 37oC and 5% CO2 with culture media for 30 min to equilibrate before coating with 2.5 mg/mL fibronectin in 0.1 M bicarbonate buffer at pH 8.0. Then, the coated wells were inoculated with 45,000 cells in 100 mL media. An additional 100 mL of media was added to each well before initiating the measurements at 4000-Hz AC. At 25 hours, after the cells formed a confluent layer, the culture media was replaced with 200 mL of fresh culture media with 1 U/mL thrombin to disrupt cell-cell junctions, and measurements continued until 50 hrs to observe cell junction recovery after thrombin treatment.

45 Measurement of endothelial cell responses to laminar flow

200,000 CRISPRi TeloHAEC cells with individual control, CCM2 or TLNRD1 guides were seeded on flow chamber slides (80176, Ibidi) that had been pre-coated with 0.2% gelatin. After 24 hours, cells were cultured under laminar flow (12 dynes/cm2) for 48 hours (10902, Ibidi pump system). Static culture controls were seeded at the same density. Cells were treated with 2 µg/ml doxycycline for 2 days prior to seeding, and throughout, for a total of 5 days. RNA was harvested with 300 µl of Trizol and extracted with 60 µl of chloroform. After addition of 1 volume 70% ethanol, RNA was loaded onto a Qiagen RNeasy spin column, washed with 350 µl of buffer RW1 and treated for 20 mins at room temperature with 10 µl Purelink DNAse (InVitrogen 12185010) in 80 µl of 1x buffer. Subsequent RNA purification steps were as per the Qiagen RNeasy protocol.

46 Fluorescence imaging and quantitation of TeloHAEC

For quantitation of actin fiber characteristics, CRISPRi teloHAEC expressing individual guideRNAs (targeting CCM2, TLNRD1, or negative control) were treated with 2 µg/ml doxycycline for 5 days. Cells were fixed in situ with by addition of paraformaldehyde to 3.2% for 30 mins at 37oC, washed with PBS, permeabilized by addition of PBS with 0.1% triton X100 for 15 mins at room temperature, washed with PBS and stained with PerkinElmer Cell Painting dyes (Phenovue Fluor 568 - Phalloidin, Phenovue Fluor 488 - Concanavalin A, Phenovue Hoechst 33342 Nuclear Stain & Phenovue 512 Nucleic Acid Stain) according to the manufacturer's instructions. Cells were imaged in four channels as described in 84 on a Perkin Elmer Opera Phenix Imaging System-106513, confocal 63x magnification with 1x binning. The stacks of images for the Phalloidin and Hoechst channels were converted to single images using maximum projection, output ranges standardized, and images exported. Cell boundaries were drawn by hand on a Phalloidin/Hoechst composite image in FIJI and saved as regions of interest (ROI). Phalloidin channel images were loaded into FIJI, converted to 16-bit grayscale, and cell areas and dimensions for each ROI were extracted using the Measure function (reporting Area and Fit Ellipse). Actin fibers were detected and quantified using the LPX FIJI plugin as described in 85, with lineExtract parameters: giwsiter = 5, mdnmsLen = 8, pickup = above (10.0), shaveLen = 3, delLen = 5, and line properties for each ROI measured using LineFeature. Parallelness (a_normAvgRad) ranges from 0 (for randomly-oriented fibers) to 1 (all fibers parallel).

Validating predictions for in vivo phenotypes of *ccm2* and *tlnr…*

47 **Zebrafish husbandry and transgenic lines**

Adult wild type AB, transgenic Tg(flk1:EGFP) (that express EGFP at the surface of blood vessels) and transgenic Tg(cmlc2:EGFP) (that express EGFP in heart muscle) zebrafish lines were maintained at 28.5 °C in circulating system water on a 14-h light/10-h dark cycle under standard conditions. Male and female embryos and larvae (≤ 5dpf) were kept in the dark in an incubator at 28.5 °C for subsequent experiments. At the end point, embryos were euthanized by tricaine overdose (MS-222; Western Chemical Inc.) followed by freezing (for RNA isolation), PFA-fixing (for histological analysis) or bleach treatment. All animal experiments were performed in accordance with relevant guidelines and regulations and with approval from the Mayo Clinic Institutional Animal Care and Use Committee.

tlnrd1 and ccm2 CRISPR knockdown in zebrafish

crRNAs for both ccm2 and tlnrd1 were designed using the Alt-R Predesigned Cas9 crRNA Selection Tool using the Integrated DNA Technologies (IDT) database. All the crRNAs were selected based on published criteria 86. For ccm2, guides were designed to target two distinct exons shared by all transcripts (AA: TTGAACGGAGACACGATACC, AF: ATGGAGCCACAACACCCACC). For tlnrd1, guides either targeted the 5' untranslated region (UTR, AN.1: GGAAACACAAGGGACGTCTC, AF: GCTGAAAGTTACACCCAACG) or the single tlnrd1 exon (AN.2: CTGCCGCTAAGGATGTTGGT, DG: CAAGAGCAAAATGCAGCTGG). For ccm2 and tlnrd1, RNPs were prepared as described; briefly, the crRNA (bearing the guide sequence) was annealed with an equal molar amount of tracrRNA (bearing the gRNA scaffold, IDT, #1072532) in duplex buffer (IDT, #11010301), to form gRNA, by heating at 95 °C for 5 min and subsequently cooling on ice. Guide RNA was assembled with an equal molar amount of Alt-R S.p. Cas9 Nuclease V3 (IDT, #1081058) to form the RNP complex (28.5 µM final concentration), by incubation at 37 °C for 5 min followed by storage at − 20 °C, following the published protocol 86,87. RNP complexes prepared from the tracrRNA/scaffold only were used as a negative control. 3 nl of each RNP complex (28.5 µM final concentration) was injected into the yolk of one-to-two cell stage embryos (wildtype, Tg;Fli:EGFP (for the permeability analysis) or Tg;cmlc2:EGFP (for visualization of the atrioventricular valve, AV)).

tlnrd1 and ccm2 morpholino knockdown in zebrafish

Morpholinos (MOs) to knock down tlnrd1 and ccm2 were designed and injected using standard protocols[88]. The ccm2 morpholino has been validated to cause cardiovascular phenotypes at the 100 µM dose[78]. A custom morpholino for Tlnrd1 (TTCCCCGAGCCACTACTAGCCATAG) was designed to target the translation start site and ordered from Gene Tools, LLC. The control oligo is a single sequence, CCTCTTACCTCAGTTACAATTTATA, that is a validated negative control[88]. Wildtype zebrafish embryos were injected with 3 nl of diluted morpholinos at multiple concentrations (50 µM, 100 µM, 200 µM, 300 µM, of control, tlnrd1 or ccm2 morpholinos) at the one cell stage, using a pico-injector (Harvard Apparatus). For coinjection, tlnrd1 and ccm2 MOs were mixed to give 50 µM of each, and 3nl of the mixture was injected.

Zebrafish imaging and phenotyping

Embryos were observed for mortality and visible phenotypes at 2 days post-fertilization (dpf) and 3 dpf using a light microscope. Images were captured at 2 and 3 dpf on an EVOS microscope (Life technology) and Zeiss Axio-observer Z1. 3 dpf embryos (knock down or control) were scored as having a heart phenotype if they displayed visible atrial chamber enlargement, moderate to severe pericardial edema and slow blood flow in the tail veins. Note that, normal zebrafish undergo cardiac looping between approximately 2dpf and 3dpf (wherein the atrium and ventricle change from a linear posterior-to-anterior arrangement to a right-to-left asymmetric arrangement). Most of the ccm2 or tlnrd1 knockdown embryos that scored positive by the criteria above also showed a looping defect, maintaining the posterior-to-anterior arrangement of atrium and ventricle at 3dpf. However, because looping is a time dependent phenomenon that normally occurs near the 3dpf time when we examined the embryos for heart phenotypes, we did not include this as a scoring criterion. For the additional phenotypic analyses described below (confocal imaging, H&E staining, tail vein morphology, blood flow & vascular permeability), we selected ccm2 or tlnrd1 knockdown embryos that scored as positive for heart phenotype at 2dpf. High resolution images for the vascular permeability and cardiac chamber analyses, were acquired using a confocal microscope LSM 800 (Zeiss).

Histological staining of zebrafish embryos for atrial/ventricular thickness

H&E staining was performed by the Mayo Clinic Comprehensive Cancer Center Histology core lab. Jacksonville, FL. Briefly, zebrafish 3dpf larvae were fixed in 4 % paraformaldehyde overnight at 4 °C. To obtain paraffin sections, fixed larvae were dehydrated stepwise in ethanol/1x PBS dilutions (5, 25, 50, 75 and 100% ethanol). Transverse sections at a thickness of 5 µm using a microtom (MICROME) were produced from the anterior beginning of the otic vesicle and included posterior structures until the cloacal vent. The sectioned region therefore spanned from the glomerulus up to the cloaca and included the complete pronephros. Sections were stained with Gills 1, eosin Y and Harris hematoxylin (Richard Allan Scientific) according to the manufacturer protocol.

FITC-Dextran 2000 kDa & Texas Red-Dextran 70 kDa injections, & imaging for tail vein morphology and vascular permeability

Microangiography was performed as described [89,90]. Briefly, at 3-days post-fertilization (3-dpf), Crispr/Cas9-injected embryos were anesthetized in 0.015% tricaine methanesulfonate (Western Chemical, Inc) and microangiography was performed by inserting a glass microneedle (World precision Instruments, Sarasota, FL) through the pericardium directly into the ventricle. For assessment of vascular morphology, 2000 kDa FITC dextran (Sigma, FD2000S-100MG) was diluted to 2 mg/ml in Zebrafish embryo medium [91], and a total of 4.5 nL was injected. For measurement of vascular permeability, Texas Red-dextran with a molecular weight of 70 kDa was solubilized in embryo medium at a 2 mg/mL concentration and a total of 4.5 nL was injected. Images were acquired after 30 minutes, using a Zeiss LSM 880 confocal microscope, and standard FITC and dsRed filter sets, and 10X objective, at room temperature. For quantitation of permeability, the Raw ".czi" images were preprocessed using the Zeiss software (ZEN2) to generate a maximum intensity projection image. The maximum intensity projection images of controls as well as Crispr mutants were then processed using the MATLAB programming platform, as described in our recent publication [90]. Movies for the blood flow in the heart and tail veins were taken by capturing 60 second bright field-time-lapse images at 60 frames per second, using an EVOS microscope at 20x magnification, as described previously [92].

qRT-PCR assays in zebrafish
klf2b, ccm2 & tlnrd1 expression was measured by qRT-PCR on RNA isolated from 100 µM tlnrd1 morpholino embryos or CRISPR tlnrd1, ccm2 or control embryos at 3 dpf, using primers for klf2b, F: GAAGAGACACCTGTGAGGGC & R: GGACACCGATTCGTAGGACC, for ccm2, F: GGCGGATCAGATGAGGGAAC & R: CAGACAGCAATACGGACCGA, and for tlnrd1, F: ACACGCGAGAGTACCTGTTG & R: TCATCCCGCGACAAATCCAA.

In situ hybridization for tlnrd1 expression in zebrafish. In situ hybridization was performed using previously validated methods[93]. Briefly, a 437 bp fragment of tlnrd1 was amplified from genomic DNA using the PCR primers, F: CATTAACGGAATGGCAGGCG and R: TGCCCGGATAAAGGCAAAGT, subcloned and verified by sequencing. Antisense in situ hybridization probes were generated using an M13 reverse primer with SpeI-linearized plasmid, while sense (negative control) probes were generated using an M13 forward primer with NotI-linearized plasmid. In situ hybridization of embryos was conducted at 24 and 72 hrs post-fertilization using these anti-sense or sense (control) probes against tlrnd1.

Applying the Variant-to-Gene-to-Program Approach to other G…

48 We tested whether the V2G2P method was generally applicable to other traits beyond CAD in endothelial cells, and to other cell types.

We first examined whether the same Perturb-seq dataset in endothelial cells could be applied to interpret variants for other vascular traits related to endothelial cell functions, beyond CAD. We applied V2G2P to 2 additional GWAS traits (Pulse Pressure (PP) and Mean Arterial Pressure (MAP), from the UK Biobank [94], with finemapping information from Hilary Finucane and Jacob Ulirsch: https://www.finucanelab.org/data). We performed V2G analysis by mapping variants associated with these traits onto the same endothelial cell enhancer map we used for CAD, and identified genes linked to PP or MAP variants in endothelial cells. We then performed V2G2P analysis, by testing for enrichment of the PP or MAP V2G gene sets in the 50 endothelial cell programs we identified from Perturb-seq. Note, that we performed the V2G2P enrichment test using only the 300 co-regulated genes in each program, because not all the genes at GWAS loci for these blood pressure traits were targeted for perturbation in our endothelial cell Perturb-seq screen.

We next examined whether the entire analysis framework could be applied to another cell type: K562 erythroid cells, which are a relevant model for red blood cell and platelet traits. Here, we examined 7 GWAS traits for red blood cell and platelet measures: Mean Corpuscular Hemoglobin (MCH), Mean

Corpuscular Volume (MCV), Platelet Count (Plt), Red Blood Cell count (RBC), Mean Corpuscular Hemoglobin Concentration (MCHC), Hemoglobin A1c (HbA1c) and Hemoglobin (Hb), along with 4 traits for which K562 cells are not likely to be an appropriate model: pulse pressure (PP), mean arterial pressure (MAP), systolic blood pressure (SBP) & diastolic blood pressure (DBP), from the UK biobank [94], with finemapping by Hilary Finucane and Jacob Ulirsch: https://www.finucanelab.org/data.

We constructed V2G maps for each trait using ABC data in K562 cells (K562-Roadmap[95]), to identify variant-containing enhancers, and identified the set of V2G genes for each trait (genes with links to variants associated, by GWAS, with each trait). We, then, constructed a gene-to-program map by applying cNMF to the genome-scale Perturb-seq data previously collected in K562 cells [19]. We tested K values over a broad range, and selected K=90 as the number of components that minimized cNMF error and maximized other ranking metrics (see "Choosing the number of components for cNMF analysis" above). Finally, we performed the V2G2P enrichment test (considering both the 300 co-regulated genes for each program and the regulators of each program, identified as the perturbations significantly affecting expression of each program, from Perturb-seq). Of the 90 programs, we found, 32 programs were prioritized for at least one of 6 GWAS traits.

Benchmarking versus other approaches and prior studies

49 We systematically compared the predictions of the V2G2P strategy to other previous studies which prioritized genes in CAD GWAS loci, and to other methods to prioritize gene sets relevant to a given trait. In considering these comparisons, we would note that V2G2P generates mechanistic hypotheses linking candidate variants to target genes to molecular pathways — a level of specificity and detail that goes far beyond other approaches, and that can directly help to guide follow-up mechanistic studies. As such, the evaluations described in this section (i.e., accuracy at identifying genes; accuracy at identifying programs) compare V2G2P to other existing approaches on those specific axes, without considering whether those approaches provide the same level of detail linking specific variants to cell types, genes, and gene expression programs.

Prioritizing genes:

Two prior studies specifically prioritized CAD genes that might act in endothelial cells:
1) Stolze et al.[29] cataloged eQTLs (correlating gene expression with genetic variants) in human aortic endothelial cells (HAECs) isolated from deceased heart donor aortic trimmings and cultured +/- IL-1β (53 individuals), as well as HAECs from another set of 157 donors (cultured +/- oxidized1-palmitoyl2-arachidonoyl-sn-glycero-3-phosphocholine). The authors' colocalization analysis of these eQTLs with CAD GWAS identified only 6 GWAS loci with a single linked eQTL gene.
2) Wunnemann et al.[30] performed a CRISPR perturbation screen of CAD GWAS variant-containing regulatory elements in 83 CAD loci, to identify elements that impacted 6 pre-selected phenotypes (E-selectin, ICAM1, VCAM1, nitric oxide, reactive oxygen species, and intracellular calcium). They identified 21 cases where a single gene was predicted to be regulated by CAD variants in a way that impacted these phenotypes in endothelial cells.
Between these two studies, only 7 of the 41 CAD V2G2P genes were also prioritized in these prior studies. 3 additional genes were known endothelial cell CAD genes. Considering these previous studies, 31 of the 41 V2G2P genes have not previously been nominated as influencing CAD risk through effects in endothelial cells.

We also compared our V2G2P genes to previous studies that prioritized genes in CAD GWAS loci, using

a variety of methods that were not specific to endothelial cells or any other given cell type:
1) Aragam et al.[12] used the Polygenic Priority Score (PoPS) method[3], which prioritizes genes based on their enrichment in gene sets derived from a variety of sources (including Gene Ontology, analysis of gene expression datasets, and others, irrespective of the cell-type-specificity of those gene sets), which were linked to CAD by enrichment of CAD GWAS variants in and around the genes in each set. The authors computed PoPS scores for all protein-coding genes within 500 kb of all GWAS signals, and prioritized the gene with the highest PoPS score in each locus, resulting in 221 unique genes [12].
2) Hodonsky et al.[96] performed eQTL and spliceQTL colocalization using bulk RNA-seq data from 138 human explanted tissue samples from left anterior descending coronary arteries, right coronary arteries, and left circumflex arteries. The authors prioritized 22 genes as being the single eQTL in a locus colocalized with a CAD GWAS signal, and 18 genes based on colocalization with spliceQTLs.
3) OpenTarget L2G[98] used a supervised machine-learning model to learn the weights of multiple evidence sources (distance, molecular QTL colocalization, chromatin interaction, and variant pathogenicity) based on a gold standard of previously identified causal genes. This analysis prioritized 103 genes in CAD GWAS loci.
4) Li et al.[97] performed a transcriptome-wide association study (TWAS) using genotype and expression data from 15 tissues (7 from STARNET and 8 from GTEx). This analysis prioritized 114 genes in CAD GWAS loci.
5) van der Harst and Verweij[10] prioritized CAD GWAS variants using a Probabilistic Annotation Integrator based on several features such as LD information, p-value distribution, coding genes, and H3K4me1 sites. This analysis identified 10 cases where a single gene was prioritized as being regulated by CAD risk variants.
Together, 23 of the 41 V2G2P genes we prioritized for CAD were also prioritized in one or more of these studies that prioritized genes using methods not specific to endothelial cells .

Altogether, our V2G2P analysis prioritizes 17 genes, including CCM2 and TLNRD1, that were not prioritized by any of these previous studies**.**

We next benchmarked our V2G2P approach versus each of these studies using the eight gold standard genes, for which clear evidence exists linking their roles in endothelial cells to atherosclerosis (see table below). The 41 V2G2P genes include 4 of 8 of these gold standard genes (50% recall). The two prior endothelial cell studies Stolze et al. [29] and Wunnemann et al. [30] each prioritized only 1 of 8 of these genes (12.5% recall). Of the non-cell type-specific studies, Hodonsky et al. [96] prioritized none of these genes, van der Harst et al. [10] prioritized 1, Li et al.[97] prioritized 2 (between 0 and 25% recall each). Only two methods had recall comparable to V2G2P, OpenTarget L2G [98], which prioritized 4 (50% recall) and the PoPS analysis by Aragam et al. [12], which prioritized 6 (75% recall). To estimate precision, we considered only the subset of the 8 gold standard gene loci where a call was made by each approach, and calculated the fraction of the prioritized genes corresponding to the gold standards. V2G2P obtained 80% precision, better than both of the endothelial cell-specific approaches. Notably, the one “incorrect” prediction made by V2G2P was for SVIL, located next to JCAD, a known gold standard gene, and it is possible given the known function of SVIL that in fact there are two causal genes in this locus. PoPS was the only method that obtained higher precision, but it prioritizes genes without providing information on likely causal variants, cell types, or gene expression pathways. By contrast, the V2G2P approach achieves good recall and precision while also providing specific molecular hypotheses about the variants, cell types, genes, pathways, and their regulatory relationships that can guide further mechanistic experiments.

	PRDM16	rs7413494, rs2493292	Promoter (rs7413494), missense variant (rs2493292)	https://www.biorxiv.org/content/10.1101/2021.12.05.471275v1 https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.120.318501	
	PLPP3	rs17114036, rs11206803, rs56170783	Closest gene, EC enhancer, CRISPR editing + eQTL (rs11206803)	https://www.biorxiv.org/content/10.1101/2021.05.06.443006v1.full.pdf+html https://pubmed.ncbi.nlm.nih.gov/30429326/ https://www.ahajournals.org/doi/10.1161/atvbaha.113.302335	
	NOS3	rs3918226	Closest gene, promoter variant	https://pubmed.ncbi.nlm.nih.gov/10683374/	
	JCAD	rs9337951	Closest gene, eQTL	https://academic.oup.com/cardiovascres/article/116/11/1863/5581206	JCAD also known as KIAA1462
	FLT1	rs17086617	Closest gene, 3'UTR variants	https://www.nature.com/articles/nm731	
	PECAM1	rs11079536	Closest gene, EC enhancer	https://www.ahajournals.org/doi/10.1161/atvbaha.107.151456 https://www.ahajournals.org/doi/full/10.1161/ATVBAHA.108.164707	
	EDN1	rs9349379	CRISPR editing, pQTL	https://pubmed.ncbi.nlm.nih.gov/28753427/	
	ARHGEF26	rs357494	missense variant	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0055202	

Gold standard genes with strong prior evidence for affecting CAD risk through functions in endothelial cells.

Prioritizing gene sets & programs:
Several previous methods have been developed to identify gene sets relevant to a GWAS trait of interest. Three of the most recent and widely-used approaches are S-LDSC28,71 (which assesses whether variants within 100 Kb of genes in a given gene set are enriched for heritability for disease), MAGMA2 (which assigns a weighted score to each gene in the genome based on GWAS signals for nearby variants, and then correlates this score with a given gene set), and sc-linker42 (which links variants to genes based on a union of enhancer-gene predictions in a given tissue, and then looks for heritability enrichment of variants linked to a given gene set). Of these, only sc-linker prioritizes programs in a way that considers the cell-type specificity of variant-to-gene links.

We compared V2G2P to each of these methods, and found that S-LDSC prioritized 2 out of 5 V2G2P programs (8 and 39), and additionally prioritized one additional endothelial cell program (50) and one housekeeping program (36). MAGMA prioritized 13 programs, including all 5 V2G2P programs plus 8 others. 3 of the MAGMA prioritized programs were not EC-specific, likely because MAGMA does not incorporate any cell-type specific information in linking variants to genes . sc-linker did not identify any

significant programs for CAD, although the V2G2P programs for CAD were highly ranked by sc-linker's heritability enrichment calculation.
Notably, beyond prioritizing programs, V2G2P also nominates specific variants and genes in GWAS loci linked to these programs, whereas these 3 other methods do not.

These results indicate that S-LDSC and MAGMA may have lower specificity for detecting heritability enrichment in relevant programs, likely because these approaches do not consider cell type specific information about likely regulatory variants and their targets (the V2G component of our V2G2P enrichment test). By contrast, sc-linker does incorporate some tissue-specific V2G information, but appears to be less sensitive for the detection of significantly enriched programs. These observations support that the V2G2P approach achieves a higher sensitivity and specificity relative to other gene set prioritization methods by incorporating both variant-to-gene predictions and Perturb-seq data. Interestingly, however, we found that each of these three other approaches still ranked the 5 V2G2P CAD programs highly, consistent with these programs being robustly-associated with CAD heritability.

Detailed information about each prior study and method used in these comparisons is provided below.

Aragam et al.[12], polygenic prioritization score (PoPS): Computed PoPS score for all protein-coding genes within 500 kb of all GWAS signals and prioritized the gene with the highest PoPS score in each locus, resulting in 221 genes. Obtained from their Supplementary Table 25.

Hodonsky et al.[96], eQTL and sQTL colocalization: Bulk RNA-seq was collected from human coronary artery tissue samples from explanted transplant tissue, or collected from rejected transplant donors (138 individuals, from left anterior descending coronary artery, right coronary artery, and left circumflex artery). eQTL colocalization was performed to find eQTL-associated genes (eGenes), or to find splice QTLs (sQTLs). The eQTL list was from Supplementary Table 12 column "vdh_CAD_PPH4" with posterior probability > 0.8 (Methods: "PPH4 >0.8 to support evidence of a shared causal variant"). The slice variant list was from Supplementary Table 22 and subset for variants with posterior probability > 0.8 (column "vdh_CAD_PPH4"). We then identified sGenes linked to the colocalized sQTLs by finding matching genes in Supplementary Table 20 (columns "gene_id" and "spliceid").

Li et al.[97], transcriptome-wide association study (TWAS): Associated genotype and expression data across 15 tissues (7 from STARNET and 8 from GTEx). We used Supplementary Table 4 for significant TWAS genes (114 genes).

OpenTarget L2G[98] : Used a supervised machine-learning model to learn the weights of multiple evidence sources (distance, molecular QTL colocalization, chromatin interaction, and variant pathogenicity) based on a gold standard of previously identified causal genes. The authors applied this model to the van der Harst coronary artery disease GWAS dataset[10]. Prioritized genes had an L2G model score > 0.5 (table downloaded from https://genetics.opentargets.org/Study/GCST005194/associations).

Stolze et al.[29], endothelial cell-specific eQTL colocalization: Human aortic endothelial cells (HAECs) were isolated from deceased heart donor aortic trimmings and cultured +/- IL-1beta (53 individuals, bulk RNA-seq), as well as 157 EC donors' cultured ECs +/- oxPL treatment (microarray). They performed eQTL mapping using Matrix eQTL and used the R package "coloc" for colocalization. We obtained their data from Table S5.

van der Harst and Verweij[10]: Prioritized variants using Probabilistic Annotation Integrator based on several features such as LD information, p-value distribution, coding genes, and H3K4me1 sites. Data were obtained from Table 2 and Online Table XX.

Wunnemann et al.[30], endothelial cell CRISPR screen for 6 phenotypes: The authors used a CRISPR screening approach to identify CAD risk variant-containing regulatory elements in 83 CAD GWAS loci that altered FACS-sortable signals for any of 6 pre-selected phenotypes in endothelial cells (E-selectin, ICAM1, VCAM1, nitric oxide, reactive oxygen species, and intracellular calcium). The identified 26 loci where perturbation of a variant-containing element affected one or more of these phenotypes (prioritizing a single gene in 21 of these loci). Data was obtained from their Fig. 3a and Supplementary Table 4.

50