Convergence of coronary artery disease genes onto endothelial cell programs

Gavin R. Schnitzler[1,2,5*], Helen Kang[3,4,*], Shi Fang[1,5], Ramcharan S. Angom[6], Vivian S. Lee-Kim[1,5], X. Rosa Ma[3,4], Ronghao Zhou[3,4], Tony Zeng[3,4], Katherine Guo[3,4], Martin S. Taylor[15], Shamsudheen K. Vellarikkal[1,5], Aurelie E. Barry[1,5], Oscar Sias-Garcia[1,5], Alex Bloemendal[1,2], Glen Munson[1], Philine Guckelberger[1], Tung H. Nguyen[1], Drew T. Bergman[1,7], Stephen Hinshaw[16], Nathan Cheng[1], Brian Cleary[1,8], Krishna Aragam[1,9], Eric S. Lander[1,10,11], Hilary K. Finucane[1,12,13,14], Debabrata Mukhopadhyay[6], Rajat M. Gupta[1,2,5,†], Jesse M. Engreitz[1,2,3,4,17†]

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 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.

Supplementary Note 1. V2G2P approach and design considerations

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.

Below, we describe key design criteria, features, and performance characteristics of this V2G2P method:

Selection of a cellular model

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 [13,14,50], 2) previous studies have demonstrated strong enrichment of CAD heritability in endothelial cells (*e.g.* [114]), and 3) detailed studies of individual CAD GWAS loci have identified likely causal genes that are clearly related to endothelial cell functions, including NOS3[115], EDN1[14], JCAD[116], ARHGEF26[117], PLPP3[118], and AIDA[119].

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 [21,120]. We confirmed that teloHAEC enhancers were enriched for heritability for CAD (**Extended Data Fig. 1a, Supplementary Tables 2 & 27**). 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 (**Extended Data Fig. 1b-d**). 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.

Building a gene-to-program map using Perturb-seq

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.

Supplementary Note 2. Comparison of V2G2P to other methods and studies for coronary artery disease

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 (**Supplementary Table 16**). Considering these previous studies, 31 of the 41 V2G2P genes have not previously been nominated as influencing CAD risk through effects in endothelial cells (**Supplementary Table 17**).

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 (**Supplementary Table 17**).

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

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 (**Supplementary Table 16**). 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 (**Extended Data Fig. 5g**). 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 (see **Supplementary Note 4**). 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.

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-LDSC[28,71] (which assesses whether variants within 100 Kb of genes in a given gene set are enriched for heritability for disease), MAGMA[2] (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-linker[42] (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) (**Extended Data Fig. 5b**). 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 (**Extended Data Fig. 5a**). 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 (**Extended Data Fig. 5f**).

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.

Supplementary Note 3. Assessing the contributions of each component of the V2G2P approach

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) (**Extended Data Fig. 6b**), 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 (**Fig. 2a,b**).

(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 (**Extended Data Fig. 6a**), 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 (**Extended Data Fig. 7**). We conclude that the V2G2P approach substantially refined the list of candidate disease genes compared to using V2G or G2P approaches alone (**Fig. 2a,c**, **Extended Data Figs. 6h, & 7**).

(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 (**Supplementary Table 26**).

(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) (**Extended Data Fig. 6c**). 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 (**Extended Data Fig. 5a**).

(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 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 (**Extended Data Fig. 6e,** see Methods). 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, **Extended Data Fig. 6f**). 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.

Supplementary Note 4. Loci and functions of additional V2G2P genes

Examining the CAD-associated V2G2P genes downstream of *CCM2* revealed insights into unresolved GWAS loci beyond the *TLNRD1* locus, and highlighted the utility of combining variant-to-gene and gene-to-program maps.

PREX1: At a GWAS signal at *20p13.1*, V2G2P analysis identified 2 genes that were linked by enhancer maps to a noncoding CAD variant (rs2004772) and 2 genes that were members of CAD-associated programs. Only one gene, *PREX1* (one of the 17 novel CAD-associated V2G2P genes that were not prioritized in any other study of CAD GWAS loci, **Supplemental Table 17**), satisfied both criteria (**Extended Data Fig. 7a**). *PREX1* encodes a Rac guanine nucleotide exchange factor known to regulate actin organization[121], similar to other CCM pathway members, and is down-regulated upon *CCM2* knockdown (**Fig. 3c**). Knockout of *PREX1* has been shown to affect endothelial cell migration *in vitro* and increase vascular barrier integrity *in vivo*[122], consistent with a potential role in atherogenesis.

SH3PXD2A and *SLK*: At a GWAS signal at *10p24.3*, noncoding variants located in the intron of *STN1* were predicted to regulate two different CAD-associated V2G2P genes, *SH3PXD2A* and *SLK*, which were both co-regulated genes in Program 8 (**Extended Data Fig. 7b**, **Fig. 2c**). Interestingly, *SH3PXD2A* encodes an adapter protein involved in invadopodia and podosome formation[123], and *SLK* (another of the 17 novel CAD-associated V2G2P genes) encodes a kinase that localizes to podosomes during cell migration[124], suggesting that genetic risk variants at this locus might regulate two genes with related functions.

GOSR2: At *17q21.3*, the noncoding variant rs17608766 has been associated with CAD risk and also with other cardiovascular phenotypes including congenital heart defects[125] and cardiac structure[126–128]. We previously linked this variant via an endothelial cell enhancer to *GOSR2*[129], which encodes a trafficking membrane protein responsible for intra-Golgi transport. Here we observed that *CCM2* knockdown led to up-regulation of *GOSR2* (+122%, $P = 8.9 \times 10^{-7}$, **Supplementary Table 18**), and *GOSR2* knockdown led to up-regulation of Program 30 (ER stress response; +94%, FDR = 2.47×10^{-47}) and down-regulation of the CAD-associated Program 35 (Focal adhesions, JUN; –23%, FDR = 8.78×10^{-4}, **Extended Data Fig. 7c**). This identifies a transcriptional phenotype for *GOSR2* in endothelial cells and suggests that *GOSR2* expression is linked to the CCM complex and other CAD-associated V2G2P genes.

Other novel CAD-associated V2G2P genes included *CALCRL*, *EXOC3L2*, *PRKARIA*, *SCUBE1*, *SPRY4*, *SVIL* and *TFPI*. Below we summarize the literature regarding their functions in endothelial cells and potential roles in atherosclerosis and other vascular diseases:

CALCRL (calcitonin receptor like receptor) encodes a receptor whose ligand (adrenomedullin) ameliorates development of atherosclerosis in Apoe-/- mice [130]. Furthermore, mice w/ endothelial-specific CALCRL knockout have increased atherosclerosis [131]. Interestingly, these phenotypes appear to be related to flow responses, since flow induces adrenomedullin release, and mice with defects in CALCRL or adrenomedullin lose EC flow responses[132].

EXOC3L2 (exocyst complex component 3-like 2) encodes a VEGF-upregulated protein that interacts with the exocyst complex (which controls spatial targeting of exocytic vesicles), and is required for VEGF-mediated directional migration of endothelial cells [133].

PRKARIA (protein kinase, cAMP-dependent, regulatory, type I, alpha) encodes a regulatory subunit for cAMP-dependent protein kinases (PKA). PKA activation has been shown to inhibit the migration of endothelial cells in culture [134]. Activation of PKA by adrenomedulin-induced cAMP is also required for flow-mediated induction of NOS3 (eNOS) which produces nitric oxide to induce vasorelaxation [132]. Interestingly, NOS3 and the adrenomedulin receptor CALCRL were also members of the 41 CAD-associated V2G2P genes found in our study.

SCUBE1 (signal peptide, CUB domain, EGF-like 1) encodes a cell surface glycoprotein expressed in platelets and endothelial cells, that can also be expressed in a soluble form. It promotes tube formation and proliferation, and inhibits apoptosis, in pulmonary artery ECs, and

is important for BMPR2-mediated activation of SMAD1/5/6 [135]. Perhaps relatedly, SMAD3 (a SMAD transcription factor specialized for signaling downstream of TGF-beta receptors, rather than BMP receptors) was also one of the 41 CAD associated V2G2P genes. In vivo, SCUBE1 knockout mice show vascular defects, particularly in neovascularization after ischemic injury [136]. Furthermore, whole body knockout of the soluble form (but not membrane bound form) diminished arterial thrombosis in mice and protected against lethal thromboembolism induced by collagen-epinephrine treatment [137].

SPRY4 (Sprouty 4) is an inhibitor of several MAPK signaling pathways, including EGFR signaling. It also functions to block integrin-mediated cell spreading via inhibition of TESK1-mediated phosphorylation of cofilin [138]. In vivo, SPRY4 knockout causes accelerated neovascularization [139], and enhances (VEGF)-A-induced angiogenesis and vascular permeability [140].

SVIL (supervillin) is a bipartite protein that forms strong attachments to the plasma membrane as well as to actin filaments (interestingly, TLNRD1 is also an actin binding adaptor protein). In vivo, it functions to inhibit platelet adhesion and arterial thrombosis (although it is unclear whether this is mediated by expression in platelets, endothelial cells or both), and SVIL variants are associated with high-shear stress thrombus formation [141]. Note that *SVIL* is in the same locus as the known gold standard gene *JCAD* (and this is the one gold standard locus where V2G2P appeared to have nominated the wrong gene). Given the known functions of *SVIL*, however, it is possible that there are actually 2 causal genes in this locus.

Lastly, TFPI (tissue factor pathway inhibitor), is an inhibitor of activated factor X and VIIa-TF proteases in the blood clotting cascade, that functions to restore hemostasis, which also binds lipoproteins in serum. In vivo, systematic delivery of TFPI improves atherosclerotic plaque stability [142]. Endothelial cell-specific knockout of TFPI also increases vascular permeability [143], and increases $Fe^{++}Cl$-induced thrombosis [144]. TFPI heterozygous knockout animals show increased atherosclerosis [145], but this may be due to TFPI functions in smooth muscle cells, since SMC-specific overexpression of TFPI reduces atherosclerosis in mouse models [146].

In summary, prior studies on members of the 41 CAD-associated V2G2P genes suggests the importance of several endothelial cell pathways in CAD, including 1) CCM and related signaling pathways, 2) flow sensing and response, 3) regulation of thrombosis & 4) regulation of angiogenesis.

Supplementary Note 5. Summary of evidence supporting the robustness, impact and generalizability of the V2G2P approach.

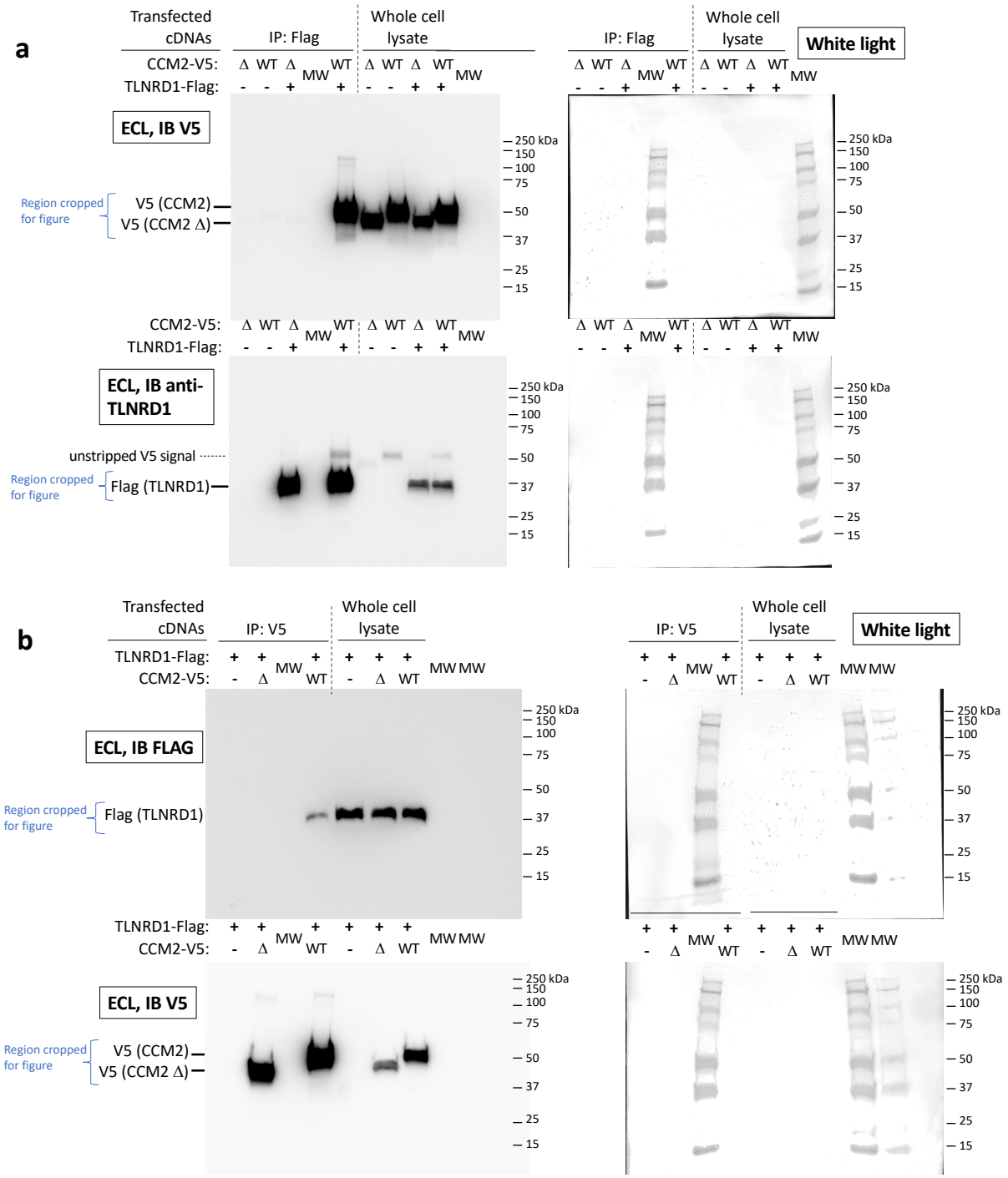
We have extensively validated the V2G2P approach for variant-to-function discovery through benchmarks that are standard in the field (including comparisons to gold standard genes for CAD and to previous studies and methods), by demonstrating its generalizability to other traits and cell types, and by additional validation experiments. These observations are summarized here:

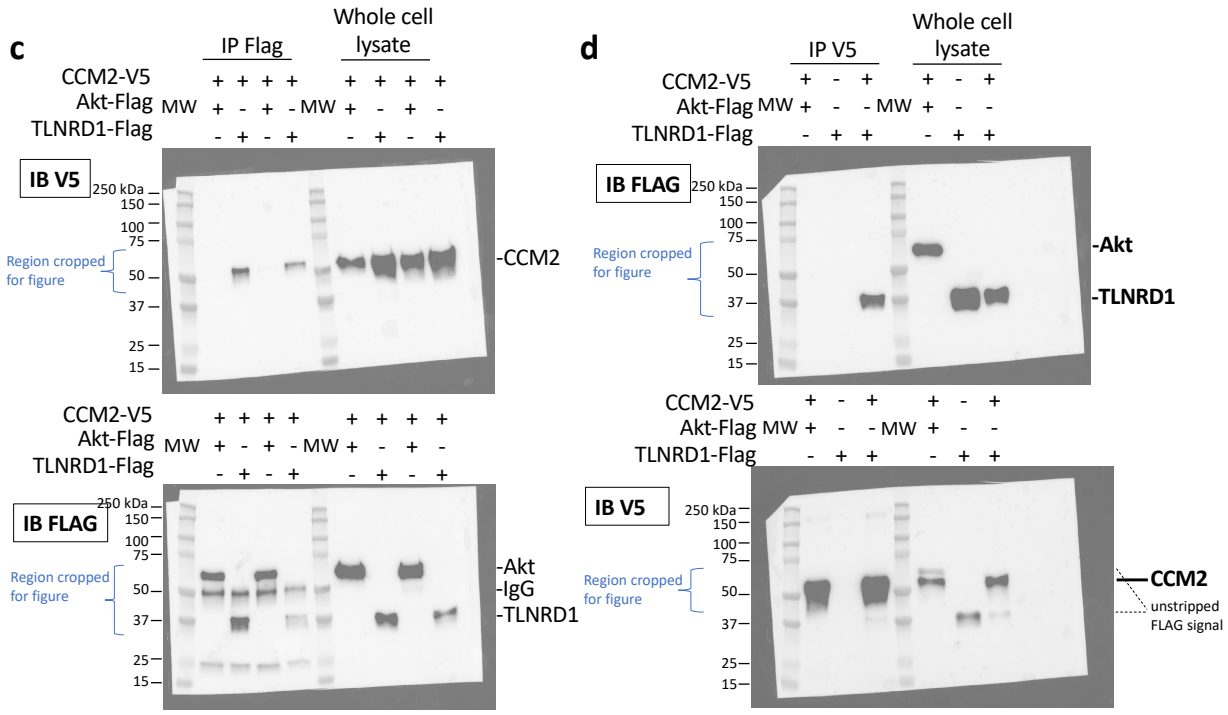
- Our V2G2P approach applied to endothelial cells and CAD prioritizes 4 out of 8 "gold standard" genes which have been shown to be important for CAD through their functions in endothelial cells (50% recall).
- In addition to these 4 genes, 5 more of the 41 V2G2P genes have in vivo evidence, in the literature, for roles in atherosclerosis and/or vascular leakiness (**Fig. 3c,d**).
- The 41 V2G2P genes were highly ranked by an independent gene prioritization method, PoPS, compared to other nearby genes at the same GWAS signals (rank-sum test $P = 2.5 \times 10^{-53}$)
- We compared our approach to 7 other studies that prioritized genes in CAD loci, and found that 31 out of 41 V2G2P genes were not prioritized in the 2 studies that focused on endothelial cells, and that 17 V2G2P genes were not prioritized in any of those studies, including *CCM2* and *TLNRD1* (**Supplementary Note 2**, **Supplementary Table 17**).
- We benchmarked the V2G2P approach, relative to these 7 other studies, for the ability to detect the 8 gold standard genes. We found that V2G2P had higher precision and recall than the two prior studies that focused on endothelial cells, and performed very well relative to the 5 non-cell type-specific prioritization studies (**Supplementary Note 2**, **Extended Data Fig. 5g**).
- Moreover, in addition to high precision and recall, V2G2P provides more detailed molecular information about prioritized genes, relative to other existing approaches. Specifically, by systematically perturbing all relevant genes in all GWAS loci, and by measuring the effects of each perturbation on the transcriptome, V2G2P can identify convergent molecular programs in a way that no other existing approach can. This same analysis also provides detailed information for every prioritized gene, linking variants to genes, and genes to convergent gene expression programs (as co-regulated genes or upstream regulators), in a disease-relevant cell type. This information is expected to establish a solid foundation to guide further studies to understand novel disease mechanisms.
- We benchmarked our approach against 3 state-of-the-art methods to prioritize gene sets/programs (S-LDSC, MAGMA & sc-linker), and found they each had lower sensitivity (detecting no significant programs, for sc-linker) and/or lower specificity (detecting non-endothelial programs, for S-LDSC and MAGMA), highlighting the power of combining cell type-specific variant-to-gene linking (V2G) and Perturb-seq (G2P) approaches (**Supplementary Note 2**). Nevertheless, our 5 V2G2P programs were ranked highly by MAGMA, S-LDSC & sc-linker, consistent with these 5 programs being robustly associated with CAD heritability.
- We performed internal benchmarking studies to show that each component of the V2G2P approach (e.g. cell type-specific epigenomic data, Perturb-seq data versus simple single cell RNA-seq data) and the combinations of these components were necessary for the high sensitivity and specificity of the approach (**Supplementary Note 3**).
- We validated transcriptional phenotypes discovered by Perturb-seq by single guide knockdowns of 9 genes (including genes perturbed in the screen as well as others

predicted to have similar regulatory effects, including CCM signaling components not tested in the original screen, **Fig. 3c**, **Supplementary Table 18**).

- Importantly, we showed that the V2G2P method was generally applicable to other traits in endothelial cells and to a completely different cellular model, K562 cells, identifying programs and genes relevant to each trait and distinct from CAD (**Extended Data Fig. 12**). Moreover, the fact that V2G2P identifies different, relevant programs for blood pressure GWAS loci in TeloHAEC than it does for CAD GWAS loci, confirms that V2G2P is not simply finding programs associated with common, nonspecific endothelial cell functions.
- Regarding biological insights, our discovery that 41 genes linked to 43 CAD GWAS signals converge onto 5 programs all related to CCM signaling strongly supports our founding hypothesis that the large number of GWAS signals will converge onto a smaller number of disease-relevant pathways. Moreover, the observation that 43 CAD GWAS signals converge onto CCM signaling in endothelial cells, considering that 78 CAD loci are associated by GWAS with circulating lipid levels, suggests that this is a key mechanism controlling risk for CAD in the human population.
- The power of this convergence for discovery of molecular mechanisms is highlighted by our finding that the poorly-studied gene *TLNRD1* encodes a novel member of the CCM signaling pathway, that directly binds to CCM2 and has highly similar transcriptional, cell physiological and developmental functions. Furthermore, we demonstrate the likely relevance of both *CCM2* and *TLNRD1* for atherosclerosis, by showing that knockdown of either gene mimics atheroprotective effects of laminar blood flow on transcription and barrier function.

In summary, our extensive benchmarking studies demonstrate that V2G2P identifies genes relevant to CAD, performs well relative to 7 other methods to prioritize CAD locus genes (both in its ability to detect gold standard genes and to detect novel genes), and outperforms 3 state-of-the-art approaches to prioritize gene sets/programs. The validity of our approach is also supported by its generalizability to other traits and cellular models, and by its support for our initial hypothesis: that a systematic approach should reveal a convergence of multiple GWAS loci onto a small number of biological pathways. Finally, our approach provides something that no other method does: identifying molecular connections of all prioritized genes, to risk variants, to each other, and to convergent pathways related to disease.


a
Transfected cDNAs
IP: Flag
Whole cell lysate
IP: Flag
Whole cell lysate
White light
CCM2-V5:
TLNRD1-Flag:
Δ WT Δ MW WT
- - + + MW
ECL, IB V5
Region cropped for figure
V5 (CCM2)
V5 (CCM2 Δ)
250 kDa
150
100
75
50
37
25
15
ECL, IB anti-TLNRD1
unstripped V5 signal
Flag (TLNRD1)
b
Transfected cDNAs
IP: V5
Whole cell lysate
IP: V5
Whole cell lysate
White light
TLNRD1-Flag:
CCM2-V5:
+ + MW + + + + + MW MW
- Δ WT - Δ WT
ECL, IB FLAG
Region cropped for figure
Flag (TLNRD1)
ECL, IB V5
V5 (CCM2)
V5 (CCM2 Δ)



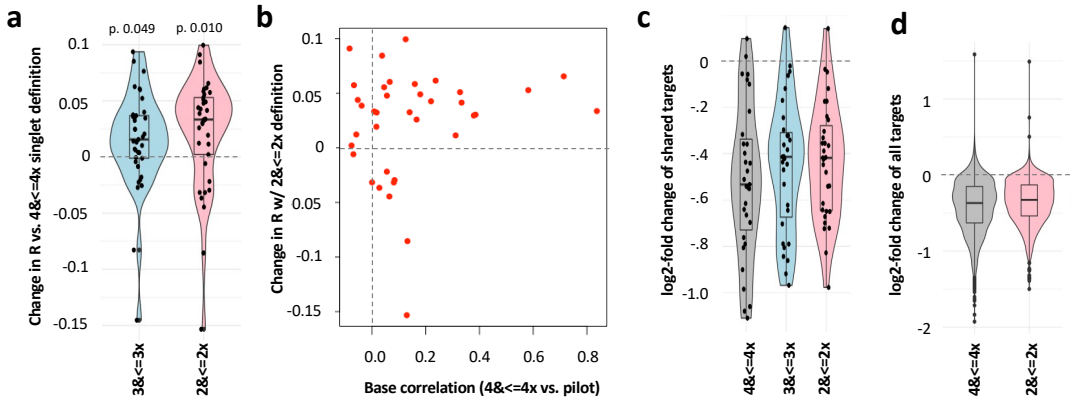
Supplementary Fig. 1. Full Western blot images.

a. Full blots for Fig. 5b. Signal for the non-fluorescent molecular weight ladder (MW, Precision Plus Dual Color Standards, Biorad #1610374) did not show up in the ECL images (left), and molecular weight positions were determined by reference to a white light image (right). Numbers on the side, molecular weights of ladder bands in kilodaltons (kDa). Regions cropped for the final figure are indicated in blue. To control for variability between extracts, protein concentration was determined using the Pierce BCA Assay (ThermoFisher), and an equal mass of protein used for each sample. “Unstripped V5 signal”: Bands remaining from the initial stain after stripping and reprobing.

b. Full blots for Extended Data Fig. 9d. As for panel (a).

c. Full blots for Extended Data Fig. 9e. As for panel (a), except that, in this blot, signal from the marker (Biorad #1610374) was strong enough in the ECL image to assign molecular weight positions without needing a white light image.

d. Full blots for Extended Data Fig. 9f. As for panel (c). “Unstripped FLAG signal”: Bands remaining from the initial stain after stripping and reprobing.



Supplemental Fig. 2. Power and accuracy considerations for singlet thresholds

a. We measured the correlation between differential expression log2 fold changes resulting from 37 perturbed gene targets shared between a pilot library and our full-scale library. Violin plots show the increases in correlation coefficients (*R*) between 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), where singlet thresholds are abbreviated as "[UMIs for the top guide required to assign a guide to a cell] &<= [fold lower number of UMIs for the 2nd to top guide, to assign a singlet]". Boxes: center line, median; limits, upper and lower quartiles. N=37. Significance was assessed by two sided T-test relative to the R values from the base comparison. Boxplot center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range.

b. Plot of the change in *R* for each target using the full library 2&<=2x singlet definition vs. the 4&<=4x singlet definition (Y axis, same as rightmost violin plot in (a)) against the *R* value for the base correlation (between the pilot and the 4&<=4x full library singlet definition, X axis). N=37.

c. Violin plots of log2 fold changes for knock down of the target genes in (a), in the full-scale library, for each singlet definition. Medians: -0.53 for 4&<=4x, -0.41 for 3&<=3x and -0.42 for 2&<=2x. N=37. Boxes as in (a).

d. As in (c), but for all 2885 targets of the full-scale library. Median log2fc for targets with the 4&<=4x singlet definition was -0.368, and with the 2&<=2x singlet definition was -0.327. N=2885. Boxes as in (a).