nature genetics

Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease

Derek Klarin[1–3,8], Qiuyu Martin Zhu[1,2,8], Connor A Emdin[1,2], Mark Chaffin[1,2], Steven Horner[4], Brian J McMillan[4], Alison Leed[4], Michael E Weale[5], Chris C A Spencer[5], François Aguet[6], Ayellet V Segrè[6], Kristin G Ardlie[2], Amit V Khera[1,2], Virendar K Kaushik[4], Pradeep Natarajan[1,2], CARDIoGRAMplusC4D Consortium[7] & Sekar Kathiresan[1,2]

UK Biobank is among the world's largest repositories for phenotypic and genotypic information in individuals of European ancestry[1]. We performed a genome-wide association study in UK Biobank testing ~9 million DNA sequence variants for association with coronary artery disease (4,831 cases and 115,455 controls) and carried out meta-analysis with previously published results. We identified 15 new loci, bringing the total number of loci associated with coronary artery disease to 95 at the time of analysis. Phenome-wide association scanning showed that *CCDC92* likely affects coronary artery disease through insulin resistance pathways, whereas experimental analysis suggests that *ARHGEF26* influences the transendothelial migration of leukocytes.

Coronary artery disease (CAD) is a leading cause of disability and mortality worldwide[2]. Genome-wide association studies (GWAS) have provided new clues to the pathophysiology for this common, complex disease. Largely using a case–control design with cases ascertained on the basis of CAD status, published studies have highlighted at least 80 loci reaching genome-wide significance[3–9].

Population-based biobanks such as UK Biobank offer new potential for genetic analysis of common, complex diseases. New opportunities include increased scale, a diverse range of traits, and the ability to explore a fuller spectrum of phenotypic consequences for identified DNA variants. Leveraging the UK Biobank resource, we sought to (i) perform a genetic discovery analysis; (ii) explore the phenotypic consequences and tissue-specific effects associated with CAD risk alleles; and (iii) characterize the functional consequences of a risk-associated mutation in a promising pathway.

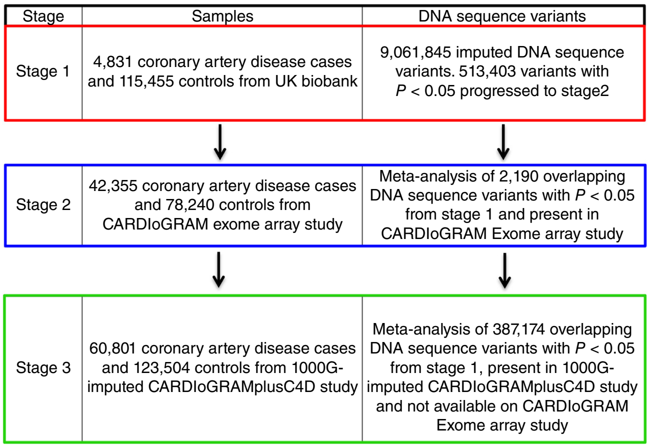
We designed a three-stage GWAS (**Fig. 1**). In stage 1, we tested the association of DNA sequence variants with CAD in UK Biobank. In stage 2, we took forward 2,190 variants that reached nominal significance in stage 1 ($P < 0.05$) for meta-analysis with results from an exome-focused array analysis in 42,355 cases and 78,240 controls[6]. In stage 3, we took forward 387,174 variants that reached nominal significance in stage 1 and were not tested in stage 2 for meta-analysis with results from a genome-wide imputation study in 60,801 cases and 123,504 controls[5]. For each variant, we combined statistical evidence across stages 1 and 2 (or stages 1 and 3) and set a statistical threshold of $P < 5 \times 10^{-8}$ for genome-wide significance.

The characteristics of UK Biobank participants stratified by presence of CAD are presented in **Supplementary Table 1**. CAD cases were more likely to be older, male, on lipid-lowering therapy, to have a history of smoking, and to be affected with type 2 diabetes. After quality control, 9,061,845 DNA sequence variants were tested for association in 4,831 patients with CAD and 115,455 controls in UK Biobank (stage 1). A total of 269 variants at five distinct loci met the genome-wide significance threshold ($P < 5 \times 10^{-8}$) (**Supplementary Figs. 1** and **2**). All five loci have previously been reported[5,10–13]. In UK Biobank, the 9p21–*CDKN2B-AS1* variant rs4977575 (NC_000009.12:g.22124745C>G) was the top association result (49% frequency for the G allele; odds ratio (OR) = 1.24, 95% confidence interval (CI) = 1.19–1.29; $P = 5.40 \times 10^{-23}$); the other four loci were 1p13–*SORT1*, *PHACTR1*, *LPA*, and *KCNE2* (**Supplementary Table 2**). For a set of previously reported CAD-associated loci[5], we compared effect estimates from the published literature with those from the current analysis in UK Biobank and found strong positive correlation in effect sizes ($\beta = 0.92$, 95% CI = 0.77–1.06; $P = 1.8 \times 10^{-17}$; **Supplementary Fig. 3**); these results validate our CAD phenotype definition in UK Biobank. A total of 513,403 variants exceeded nominal significance ($P < 0.05$) and were taken forward to stage 2 or 3.

After meta-analysis, 15 new loci exceeded genome-wide significance (**Tables 1** and **2**), bringing the total number of established CAD loci to 95 at the time of this analysis. Of note, while this manuscript

[1]Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [2]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. [3]Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts, USA. [4]Center for the Development of Therapeutics, Broad Institute, Cambridge, Massachusetts, USA. [5]Genomics plc, Oxford, UK. [6]Cancer Program, Broad Institute, Cambridge, Massachusetts, USA. [7]A list of members and affiliations appears in the **Supplementary Note**. [8]These authors contributed equally to this work. Correspondence should be addressed to S.K. (skathiresan1@mgh.harvard.edu).

Received 13 January; accepted 15 June; published online 17 July 2017; doi:10.1038/ng.3914



Stage	Samples	DNA sequence variants
Stage 1	4,831 coronary artery disease cases and 115,455 controls from UK biobank	9,061,845 imputed DNA sequence variants. 513,403 variants with $P < 0.05$ progressed to stage2
Stage 2	42,355 coronary artery disease cases and 78,240 controls from CARDIoGRAM exome array study	Meta-analysis of 2,190 overlapping DNA sequence variants with $P < 0.05$ from stage 1 and present in CARDIoGRAM Exome array study
Stage 3	60,801 coronary artery disease cases and 123,504 controls from 1000G-imputed CARDIoGRAMplusC4D study	Meta-analysis of 387,174 overlapping DNA sequence variants with $P < 0.05$ from stage 1, present in 1000G-imputed CARDIoGRAMplusC4D study and not available on CARDIoGRAM Exome array study

Figure 1 Study design. Stage 1 consisted of a GWAS for CAD phenotype performed in UK Biobank; variants below a threshold of $P < 0.05$ moved forward to meta-analysis with CARDIoGRAM Exome (stage 2) or CARDIoGRAMplusC4D summary statistics (stage 3). 1000G, 1000 Genomes; CARDIoGRAM, Coronary Artery Disease Genome-Wide Replication and Meta-analysis.

was under review, one of the 15 loci (*HNF1A*) was reported[9]. Effect allele frequencies for the 15 newly identified loci ranged from 13% to 86%, with effect sizes ranging from 1.05 to 1.08. Descriptions of relevant loci appear in **Supplementary Table 3**, and regional association plots for new CAD loci are shown in **Supplementary Figures 4–6**.

To move from these 15 DNA sequence variants to biological insights, we took two approaches: phenome-wide association scanning and functional analysis. Understanding the full spectrum of phenotypic consequences of a given DNA sequence variant may shed light on the mechanism by which a variant or gene leads to disease. Termed a 'phenome-wide association study', or PheWAS, this approach tests the association of a mapped disease-associated variant with a broad range of human phenotypes[14]. In collaboration with Genomics plc, we conducted a PheWAS combining UK Biobank data, mRNA transcript phenotypes in the Genotype-Tissue Expression (GTEx) Project data set[15], and an integrated set of GWAS results from a variety of publically available sources[16–24].

We found that several of the newly identified DNA sequence variants correlated with a range of human traits (**Fig. 2** and **Supplementary Tables 4** and **5**). For example, the intronic variant rs10841443 within *RP11-664H17.1* is in close proximity to *PDE3A*, which encodes a phosphodiesterase previously implicated in an autosomal dominant form of hypertension[25]. PheWAS showed an association for this variant with diastolic blood pressure[26], suggesting that this locus may be acting through hypertension. The variant rs2244608 within *HNF1A* has previously been associated with LDL cholesterol, elevated levels of which represent a causal path to atherosclerosis[16]. The variant rs7500448 within *CDH13* (encoding cadherin 13 or T-cadherin), a vascular adiponectin receptor implicated in hypertensive and insulin resistance biology[27], associates with plasma adiponectin levels. Variant rs2972146 is downstream of *IRS1* (encoding insulin receptor substrate-1; ref. 24) and is a *cis* expression quantitative trait locus (eQTL) for *IRS1* expression in adipose tissue. rs2972146 associates with a range of phenotypes seen in the setting of insulin resistance, including HDL cholesterol, triglycerides, adiponectin, fasting insulin, and type 2 diabetes.

Additional compelling insights from the PheWAS emerged at the *CCDC92* locus. Across 25 distinct traits and disorders, we observed significant association ($P < 0.00013$) for *CCDC92* p.Ser70Cys (rs11057401) with body fat percentage and waist-to-hip circumference ratio, as well as plasma HDL, triglyceride, and adiponectin levels. The directionalities of these associations are hallmarks of insulin resistance and lipodystrophy[17,28], and the association with plasma adiponectin levels localizes these genetic effects to adipose tissue. Recent work has highlighted two candidate genes at this locus, *CCDC92* and *DNAH10* (ref. 29), and further experimental work is necessary to define the causal gene.

However, a few of the CAD-associated loci (*FN1*, *LOX*, *ITGB5*, and *ARHGEF26*) did not significantly associate with any of the studied risk factor traits and, thus, appear to function through pathways beyond known CAD risk factors (**Fig. 2** and **Supplementary Tables 3–5**). A common variant within an intron of *FN1* (ref. 30; encoding fibronectin 1) and a missense variant in *LOX*[31] (encoding lysyl oxidase) suggest potential links to extracellular matrix biology. Of note, rare coding mutations in *LOX* were recently described to cause Mendelian forms of thoracic aortic aneurysm and dissection[32,33], highlighting a potential common link between atherosclerosis and aortic disease, possibly involving altered extracellular matrix biology. A variant downstream of *ITGB5* (ref. 34; encoding integrin subunit β5) suggests a role for pathways underlying cell adhesion and migration.

In aggregate, our analysis brings the total number of known CAD-associated loci to 95 (refs. 3–9), and in **Figure 3** we organize these loci into plausible pathways. Of note, the causal variant, gene, cell type, or mechanism has been definitively identified for only a few of these loci and, as such, additional experimental research will be required, particularly at the >50% of loci without an apparent link to known risk factors.

For one of the new loci not related to known risk factors, *ARHGEF26* (encoding Rho guanine nucleotide exchange factor 26), we performed functional studies. Prior experimental work had connected this gene with atherosclerosis in mice[35]. Earlier studies also established a role for ARHGEF26 in facilitating the transendothelial migration of leukocytes, a key step in the initiation of atherosclerosis[36,37].

Table 1 New loci from analysis of UK Biobank and CARDIoGRAM Exome array study

	UK Biobank							Stage 2 exome study		Combined		
Lead variant	Chr.	Gene	Description	EA	EAF	OR	*P*	OR	*P*	OR	95% CI	*P*
rs2972146	2	(*LOC646736*)	Intergenic	T	0.65	1.07	0.0011	1.05	2.01×10^{-7}	1.06	1.04–1.07	1.46×10^{-9}
rs12493885 (p.Val29Leu)	3	*ARHGEF26*	Missense	C	0.85	1.07	0.039	1.09	8.28×10^{-9}	1.08	1.06–1.11	1.02×10^{-9}
rs1800449 (p.Arg158Gln)	5	*LOX*	Missense	T	0.17	1.09	0.0039	1.07	1.72×10^{-7}	1.07	1.05–1.09	2.99×10^{-9}
rs11057401 (p.Ser70Cys)	12	*CCDC92*	Missense	T	0.69	1.08	0.001	1.05	4.32×10^{-7}	1.06	1.04–1.08	3.88×10^{-9}

Genes for variants that are outside the transcript boundary of the protein-coding gene are shown in parentheses. Chr., chromosome; CI, confidence interval; EA, effect allele; EAF, effect allele frequency; OR, odds ratio.

Table 2 New loci from analysis of UK Biobank and CARDIoGRAMplusC4D 1000 Genomes imputation study

	UK Biobank							Stage 3 1000G-imputed study		Combined		
Lead variant	Chr.	Gene	Description	EA	EAF	OR	P	OR	P	OR	95% CI	P
rs17517928	2	*FN1*	Intronic	C	0.75	1.08	0.0026	1.06	5.14×10^{-7}	1.06	1.04–1.08	1.06×10^{-8}
rs17843797	3	*UMPS–ITGB5*	Intronic	G	0.13	1.11	0.00019	1.07	2.43×10^{-6}	1.07	1.05–1.10	1.52×10^{-8}
rs748431	3	*FGD5*	Intronic	G	0.36	1.04	0.042	1.05	2.14×10^{-7}	1.05	1.03–1.07	2.63×10^{-8}
rs7623687	3	*RHOA*	Intronic	A	0.86	1.09	0.0073	1.07	5.22×10^{-7}	1.08	1.05–1.10	2.00×10^{-8}
rs10857147	4	(*FGF5*)	Regulatory region	T	0.29	1.06	0.014	1.06	5.83×10^{-7}	1.06	1.04–1.08	3.39×10^{-8}
rs7678555	4	(*MAD2L1*)	Intergenic	C	0.29	1.06	0.027	1.06	3.26×10^{-7}	1.06	1.04–1.08	2.91×10^{-8}
rs10841443	12	*RP11-664H17.1*	Intronic	G	0.67	1.06	0.0073	1.05	5.81×10^{-7}	1.05	1.03–1.07	2.23×10^{-8}
rs2244608	12	*HNF1A*	Intronic	G	0.32	1.07	0.003	1.05	1.02×10^{-6}	1.05	1.03–1.07	2.41×10^{-8}
rs3851738	16	*CFDP1*	Intronic	C	0.6	1.07	0.00089	1.05	1.88×10^{-6}	1.05	1.03–1.07	2.43×10^{-8}
rs7500448	16	*CDH13*	Intronic	A	0.75	1.1	0.00016	1.06	2.11×10^{-6}	1.06	1.04–1.09	1.20×10^{-8}
rs8108632	19	*TGFB1*	Intronic	T	0.41	1.06	0.011	1.05	4.76×10^{-7}	1.05	1.03–1.07	2.35×10^{-8}

New loci defined at the time of analysis. Genes for variants that are outside the transcript boundary of the protein-coding gene are shown in parentheses. 1000G, 1000 Genomes; chr., chromosome; CI, confidence interval; EA, effect allele; EAF, effect allele frequency; OR, odds ratio.

ARHGEF26 has been shown to activate RhoG GTPase by promoting the exchange of GDP for GTP and contributing to the formation of ICAM1-induced endothelial docking structures that facilitate leukocyte transendothelial migration[36,37]. In addition, *Arhgef26*-null mice, when crossed with atherosclerosis-prone *Apoe*-null mice, displayed less aortic atherosclerosis[35].

At *ARHGEF26* p.Val29Leu (rs12493885), the allele encoding Leu29, observed in 85% of participants, was associated with increased risk for CAD. We first examined the hypothesis that a haplotype block containing this allele might alter expression of *ARHGEF26* in coronary artery. Although this genomic region demonstrates eQTL effects in a variety of tissues, there was no evidence of alteration of *ARHGEF26* expression in coronary artery from either eQTL or allele-specific expression analysis (**Supplementary Fig. 7**). To further evaluate the possibility that a haplotype containing the allele for Leu29 might alter gene expression, we performed luciferase reporter assays. We cloned a 2.5-kb region immediately upstream of the *ARHGEF26* start codon consisting of the promoter, 5′ UTR, and regions with Encyclopedia of DNA Elements (ENCODE) annotations suggestive of potential *cis*-acting elements. We obtained the reference (in linkage disequilibrium (LD) with the Val29-encoding G allele) and alternative (in LD with the Leu29-encoding C allele) haplotypes of this region from human cells heterozygous at rs12493885. We then coupled each haplotype with a luciferase reporter and measured luciferase activity (**Supplementary Fig. 8**). In HEK293 cells, human aortic endothelial cells (HAECs), and human umbilical vein endothelial cells (HUVECs), there was not a significant difference in luciferase activity between the reference and alternative haplotypes. These data suggest that the *ARHGEF26* allele encoding Leu29 may confer CAD risk via mechanisms other than an effect on *ARHGEF26* transcription or promoter activity in disease-relevant tissue.

Next, we examined the hypothesis that *ARHGEF26* p.Val29Leu might influence disease risk by altering the encoded protein. We knocked down endogenous *ARHGEF26* through small interfering RNA (siRNA) and observed decreased leukocyte transendothelial migration, leukocyte adhesion on endothelial cells, and vascular smooth muscle cell proliferation[38] (**Fig. 4** and **Supplementary Fig. 9**). Overexpression of exogenous wild-type ARHGEF26 rescued these phenotypes. However, overexpression of the exogenous ARHGEF26 Leu29 mutant led to rescued phenotypes that consistently exceeded those observed with overexpression of wild-type protein. These data support the hypothesis that the *ARHGEF26* allele encoding Leu29 associated with increased CAD risk may lead to a gain of protein function.

How could the *ARHGEF26* mutation encoding Leu29 lead to a gain-of-function phenotype? We evaluated the functional impact of this mutation in two ways, addressing ARHGEF26 activity and quantity. First, we assessed whether the mutation introducing Leu29 could alter ARHGEF26 nucleotide-exchange activity on RhoG. To this end, we developed a GTP–GDP nucleotide-exchange assay using recombinant full-length human ARHGEF26 (wild type or Leu29) and

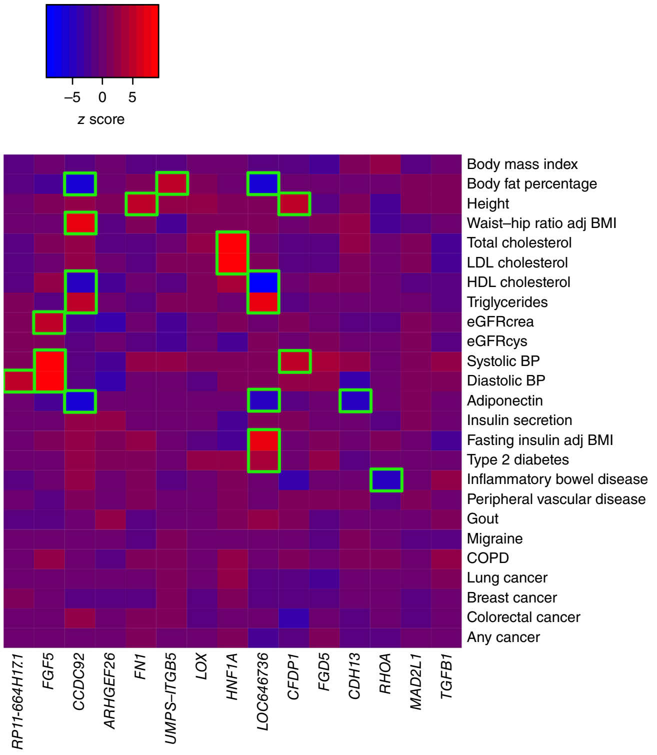


Figure 2 Phenome-wide association results for 15 loci new at the time of analysis. For the 15 new CAD risk variants identified in our study, *z* scores (aligned to the CAD risk allele) were obtained from the Genomics plc Platform and UK Biobank. A positive *z* score (red) indicates a positive association between the CAD risk allele and the disease or trait, whereas a negative *z* score (blue) indicates an inverse association. Boxes are outlined in green if the variant is significantly ($P < 0.00013$) associated with the given trait. Adj, adjusted; BMI, body mass index; BP, blood pressure; crea, creatinine; cys, cystatin-c; COPD, chronic obstructive pulmonary disease; eGFR, estimated glomerular filtration rate.

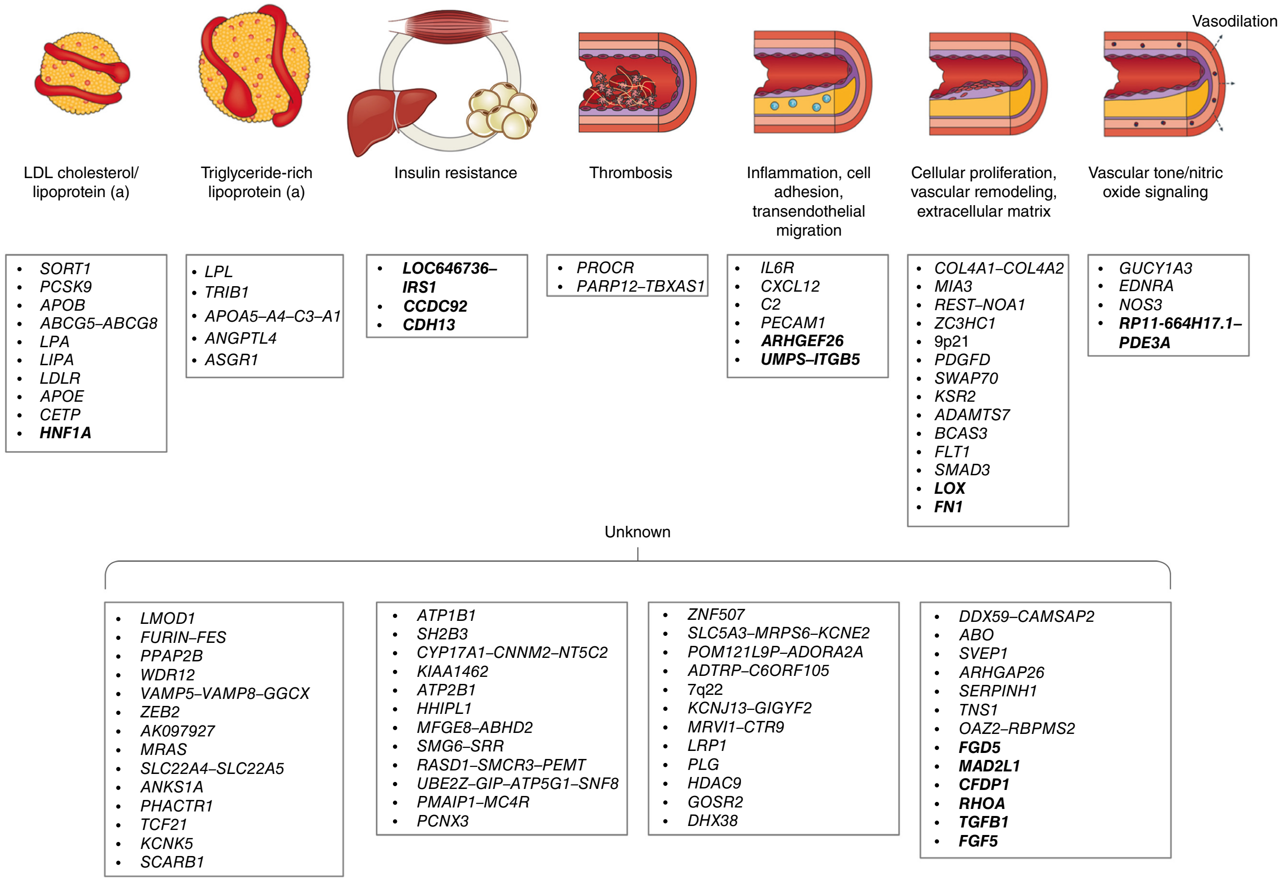


Figure 3 Biological pathways underlying genetic loci associated with coronary artery disease. The CAD GWAS loci identified thus far are depicted along with the plausible relationship to the underlying biological pathway. The 15 new loci described in this paper are shown in bold. Loci names are based on the nearest genes; however, the causal gene(s) remain unclear for most associated loci and, as such, the resultant annotation may prove incorrect in some cases. Adapted from ref. 41.

human RhoG[39]. In a cell-free system, equal amounts of wild-type and Leu29 ARHGEF26 protein were incubated with RhoG preloaded with GDP. After 60 min, we observed no significant difference in nucleotide-exchange activity between the wild-type and mutant ARHGEF26 proteins (**Supplementary Fig. 10**).

Second, we assessed whether the allele encoding Leu29 affects the cellular abundance of ARHGEF26 protein. We examined this possibility by treating cells expressing wild-type or Leu29 mutant ARHGEF26 with cycloheximide, a protein synthesis inhibitor, and compared ARHGEF26 degradation over time by immunoblotting. In comparison to wild-type ARHGEF26, the Leu29 mutant displayed a longer half-life (**Supplementary Fig. 11**). While further work is needed to understand the mechanism *in vivo*, our *in vitro* results suggest that the gain-of-function phenotype observed may be secondary to resistance of the Leu29 mutant protein to degradation.

Our study should be interpreted within the context of its limitations. First, we focused on participants of European ancestry within UK Biobank, and results may therefore not be generalizable to other populations. Second, our CAD phenotype definitions are based largely on interviews and electronic health records, and this may result in misallocation of case status. However, such misclassification should reduce statistical power for discovery and bias results toward the null. Finally, although we observed no evidence of robust changes in *ARHGEF26* expression associated with the haplotype encoding Leu29 in disease-relevant tissue, it is possible that other regulatory mechanisms may potentiate the gain-of-function phenotypes we observed.

In summary, we performed a gene discovery study for CAD using a large population-based biobank, identified 15 new loci, and explored the phenotypic consequences of CAD risk variants through PheWAS and *in vitro* functional analysis. These findings permit several conclusions. First, CAD cases phenotyped via electronic health records and verbal interviews exhibit similar genetic architectures to those derived in epidemiological cohorts and can prove useful in gene discovery efforts. Second, PheWAS with risk variants can provide initial clues to how DNA sequence variants may lead to disease. Lastly, considerable experimental evidence in cells and rodents has suggested that transendothelial migration of leukocytes is a key step in the formation of atherosclerosis[40]; here we provide genetic support in humans for a role of this pathway in CAD.

URLs. UK Biobank, https://www.ukbiobank.ac.uk/; CARDIoGRAM exome and CARDIoGRAMplusC4D data, http://www.cardiogramplusc4d.org/; R statistical software, http://www.R-project.org/; SNPTEST software, http://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html; Genotype-Tissue Expression Project, http://gtexportal.org/.

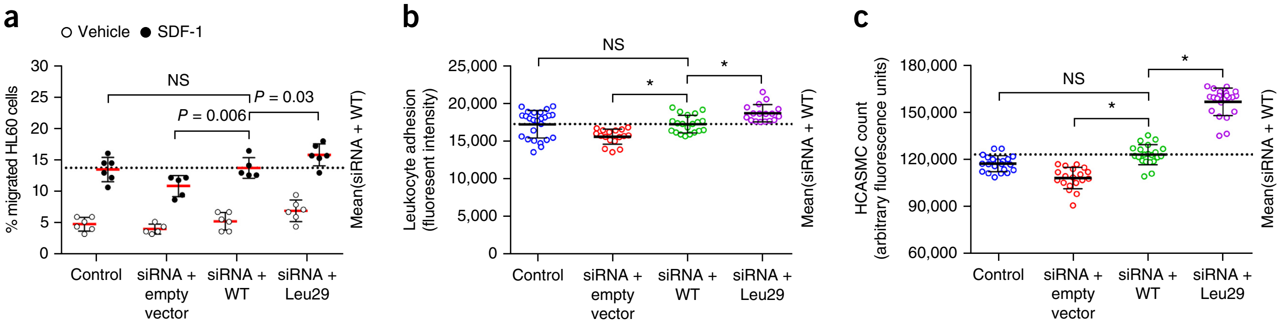


Figure 4 Functional assessment of ARHGEF26 p.Val29Leu *in vitro*. (**a**) ARHGEF26 Leu29 increases leukocyte transendothelial migration. HAECs were transfected with non-targeting siRNA and empty vector (control), siRNA against *ARHGEF26* 3′ UTR and empty vector, siRNA and ARHGEF26 WT (wild type), or siRNA and ARHGEF26 Leu29. Transfected HAECs were plated on Transwell inserts and treated with 10 ng/ml tumor necrosis factor (TNF)-α. Differentiated HL60 cells were loaded on the upper chambers of Transwell plates and allowed to transmigrate across HAECs toward vehicle (unfilled circles) or 50 ng/ml SDF-1 (filled circles). The migrated cells were quantified as the percentage of input cells per well (n = 5 or 6; mean ± s.d.; F = 11.89, degrees of freedom (DF) = 3 by two-way ANOVA within vehicle and SDF-1 subgroups with Fisher's LSD test; variance among vehicle subgroups non-significant; NS, not significant; representative of three independent experiments). (**b**) ARHGEF26 Leu29 increases leukocyte adhesion on endothelial cells. HAECs were transfected as in **a** and cultured on 96-well plates until confluent and were then treated with 10 ng/ml TNF-α. Calcein-AM-labeled THP-1 cells were incubated with the HAECs and washed to remove non-adherent cells. The adherent cells were lysed, quantified by Calcein-AM fluorescence, and compared to siRNA + WT (n = 25, 17, 20, and 17; mean ± s.d.; F = 14.53, DF = 3 by one-way ANOVA; NS, not significant; $^*P$ < 0.0001 compared to siRNA + WT; representative of three independent experiments). (**c**) ARHGEF26 Leu29 increases vascular smooth muscle cell proliferation. HCASMCs were transfected as in **a** and made quiescent by serum starvation for 48 h, followed by 72 h of proliferation in normal serum-containing medium. Cell proliferation was quantified by a luminescence assay and compared to siRNA + WT (n = 20; mean ± s.d.; F = 197.5, DF = 3 by one-way ANOVA; NS, not significant; $^*P < 0.0001$ compared to siRNA + WT; representative of three independent experiments).

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This research has been conducted using the UK Biobank resource, application 7089. The PheWAS carried out by Genomics plc was also conducted using the UK Biobank resource under a separate agreement. The authors thank K. Burridge and E. Wittchen for technical advice on the leukocyte transendothelial migration assay. This analysis was supported by the National Heart, Lung, and Blood Institute of the US National Institutes of Health under award number T32 HL007734 (D.K.), the John S. LaDue Memorial Fellowship in Cardiology at Harvard Medical School (P.N.), the KL2/Catalyst Medical Research Investigator Training award from Harvard Catalyst funded by the US National Institutes of Health (TR001100) (A.V.K.), the Ofer and Shelly Nemirovsky Research Scholar award from Massachusetts General Hospital (MGH), the Donovan Family Foundation, and US National Institutes of Health grant R01 HL127564 (S.K.). The sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

AUTHOR CONTRIBUTIONS

Concept and design: D.K., Q.M.Z., M.E.W., A.V.K., P.N., S.K. Acquisition, analysis, or interpretation of data: D.K., Q.M.Z., C.A.E., M.C., S.H., B.J.M., A.L., M.E.W., C.C.A.S., F.A., A.V.S., K.G.A., A.V.K., V.K.K., P.N., S.K. Drafting of the manuscript: D.K., Q.M.Z., C.A.E., M.E.W., A.V.K., P.N., S.K. Critical revision of the manuscript for important intellectual content: D.K., Q.M.Z., C.A.E., M.C., S.H., B.J.M., A.L., M.E.W, C.C.A.S., F.A., A.V.S., K.G.A., A.V.K., V.K.K., P.N., S.K. Administrative, technical, or material support: D.K., S.K.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).
2. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1459–1544 (2016).
3. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
4. Deloukas, P. *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* **45**, 25–33 (2013).
5. CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
6. Myocardial Infarction Genetics and CARDIoGRAM Exome Consortia Investigators. Coding variation in *ANGPTL4*, *LPL*, and *SVEP1* and the risk of coronary disease. *N. Engl. J. Med.* **374**, 1134–1144 (2016).
7. Nioi, P. *et al.* Variant *ASGR1* associated with a reduced risk of coronary artery disease. *N. Engl. J. Med.* **374**, 2131–2141 (2016).
8. Webb, T.R. *et al.* Systematic evaluation of pleiotropy identifies 6 further loci associated with coronary artery disease. *J. Am. Coll. Cardiol.* **69**, 823–836 (2017).
9. Howson, J.M.M. *et al.* Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat. Genet.* http://dx.doi.org/10.1038/ng.3874 (2017).
10. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
11. Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.* **41**, 334–341 (2009).
12. Trégouët, D.A. *et al.* Genome-wide haplotype association study identifies the *SLC22A3–LPAL2–LPA* gene cluster as a risk locus for coronary artery disease. *Nat. Genet.* **41**, 283–285 (2009).
13. Samani, N.J. *et al.* Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* **357**, 443–453 (2007).
14. Denny, J.C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
15. Aguet, F. *et al.* Local genetic effects on gene expression across 44 human tissues. Prepinrt at *bioRxiv* http://dx.doi.org/10.1101/074450 (2016).
16. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
17. Manning, A.K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
18. Prokopenko, I. *et al.* A central role for *GRB10* in regulation of islet function in man. *PLoS Genet.* **10**, e1004235 (2014).
19. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).

20. Berndt, S.I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* **45**, 501–512 (2013).
21. Pattaro, C. *et al.* Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat. Commun.* **7**, 10023 (2016).
22. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
23. Dastani, Z. *et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8**, e1002607 (2012).
24. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
25. Maass, P.G. *et al.* *PDE3A* mutations cause autosomal dominant hypertension with brachydactyly. *Nat. Genet.* **47**, 647–653 (2015).
26. Kato, N. *et al.* Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.* **47**, 1282–1293 (2015).
27. Chung, C.M. *et al.* A genome-wide association study reveals a quantitative trait locus of adiponectin on *CDH13* that predicts cardiometabolic outcomes. *Diabetes* **60**, 2417–2423 (2011).
28. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
29. Lotta, L.A. *et al.* Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat. Genet.* **49**, 17–26 (2017).
30. Sakai, T., Larsen, M. & Yamada, K.M. Fibronectin requirement in branching morphogenesis. *Nature* **423**, 876–881 (2003).
31. Erler, J.T. *et al.* Lysyl oxidase is essential for hypoxia-induced metastasis. *Nature* **440**, 1222–1226 (2006).
32. Lee, V.S. *et al.* Loss of function mutation in *LOX* causes thoracic aortic aneurysm and dissection in humans. *Proc. Natl. Acad. Sci. USA* **113**, 8759–8764 (2016).
33. Guo, D.C. *et al.* *LOX* mutations predispose to thoracic aortic aneurysms and dissections. *Circ. Res.* **118**, 928–934 (2016).
34. Hood, J.D. & Cheresh, D.A. Role of integrins in cell invasion and migration. *Nat. Rev. Cancer* **2**, 91–100 (2002).
35. Samson, T. *et al.* The guanine-nucleotide exchange factor SGEF plays a crucial role in the formation of atherosclerosis. *PLoS One* **8**, e55202 (2013).
36. van Rijssel, J. *et al.* The Rho-guanine nucleotide exchange factor Trio controls leukocyte transendothelial migration by promoting docking structure formation. *Mol. Biol. Cell* **23**, 2831–2844 (2012).
37. van Buul, J.D. *et al.* RhoG regulates endothelial apical cup assembly downstream from ICAM1 engagement and is involved in leukocyte trans-endothelial migration. *J. Cell Biol.* **178**, 1279–1293 (2007).
38. Zahedi, F. *et al.* Dicer generates a regulatory microRNA network in smooth muscle cells that limits neointima formation during vascular repair. *Cell. Mol. Life Sci.* **74**, 359–372 (2017).
39. Ellerbroek, S.M. *et al.* SGEF, a RhoG guanine nucleotide exchange factor that stimulates macropinocytosis. *Mol. Biol. Cell* **15**, 3309–3319 (2004).
40. Gerhardt, T. & Ley, K. Monocyte trafficking across the vessel wall. *Cardiovasc. Res.* **107**, 321–330 (2015).
41. Khera, A.V. & Kathiresan, S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.* **18**, 331–344 (2017).

ONLINE METHODS

Study design and samples. We performed a three-stage sequential analysis to identify new genetic loci associated with CAD. In stage 1, we first tested the association of DNA sequence variants with CAD in UK Biobank. Beginning in 2006, individuals aged 45 to 69 years were recruited from across the UK for participation in the UK Biobank Study[1]. At enrollment, a trained healthcare provider ascertained participants' medical histories through verbal interview. In addition, participants' electronic health records (EHRs), including inpatient International Classification of Disease (ICD-10) diagnosis codes and Office of Population and Censuses Surveys (OPCS-4) procedure codes, were integrated into UK Biobank. Individuals were defined as having CAD on the basis of at least one of the following criteria:

(1) Myocardial infarction (MI), coronary artery bypass grafting, or coronary artery angioplasty documented in their medical history at the time of enrollment by a trained nurse;
(2) Hospitalization for an ICD-10 code for acute myocardial infarction (I21.0, I21.1, I21.2, I21.4, I21.9);
(3) Hospitalization for an OPCS-4 coded procedure: coronary artery bypass grafting (K40.1–40.4, K41.1–41.4, K45.1–45.5);
(4) Hospitalization for an OPCS-4 coded procedure: coronary angioplasty with or without stenting (K49.1–49.2, K49.8–49.9, K50.2, K75.1–75.4, K75.8–75.9).

All other individuals were defined as controls. In total, genotypes were available for 120,286 participants of European ancestry.

In stage 2, we took forward 2,190 variants that reached nominal significance in stage 1 for meta-analysis in the CARDIoGRAM Exome Consortia exome array analysis that incorporated 42,355 cases and 78,240 controls[6] (**Supplementary Table 6**). In stage 3, we took forward 387,174 variants that reached nominal significance in stage 1 (and were not available in stage 2) for meta-analysis in the CARDIoGRAMplusC4D 1000 Genomes–imputed study containing 60,801 cases and 123,504 controls[5]. Informed consent was obtained for all participants, and UK Biobank received ethical approval from the Research Ethics Committee (reference number 11/NW/0382). Our study was approved by a local institutional review board at Partners Healthcare (protocol 2013P001840).

Genotyping and quality control. UK Biobank samples were genotyped using either the UK BiLEVE[42] or UK Biobank Axiom array, with array analysis performed in 33 separate batches of samples by Affymetrix. A total of 806,466 directly genotyped DNA sequence variants were available after variant quality control. The UK Biobank team then performed imputation from a combined 1000 Genomes and UK10K reference panel; phasing was performed using SHAPEIT3 and imputation was carried out via IMPUTE3. The variant-level quality control exclusion metrics applied to imputed data for GWAS included the following: call rate < 95%, Hardy–Weinberg equilibrium $P < 1 \times 10^{-6}$, posterior call probability < 0.9, imputation quality < 0.4, and MAF < 0.005. Sex chromosome and mitochondrial genetic data were excluded from this analysis. In total, 9,061,845 imputed DNA sequence variants were included in our analysis. For sample quality control, the UK Biobank analysis team removed individuals with relatedness corresponding to third-degree relatives or closer, and an additional 480 samples with an excess of missing genotype calls or more heterozygosity than expected were excluded. In total, genotypes were available for 120,286 participants of European ancestry.

Statistical analysis. *Stage 1 association analysis.* BOLT-LMM software[43] was used to generate linear mixed models (LMMs) for association testing. CAD case status was analyzed while adjusting for age, sex, and chip array at runtime. This analysis was used to derive statistical significance. As effect estimates from BOLT-LMM software are unreliable because of the treatment of binary-phenotype data as quantitative data, we performed logistic regression to derive effect estimates for each variant that exceeded genome-wide significance. Effect estimates for top variants were derived from logistic regression using allelic dosages adjusting for age, sex, chip at runtime, and ten principal components under the assumption of additive effects using the R v3.2.0 and SNPTEST statistical software programs.

Stage 2 and 3 meta-analysis. In stage 2, top variants ($P < 0.05$) from UK Biobank were then meta-analyzed with exome chip data from the CARDIoGRAM Exome Consortia[6]. Tested variants in the CARDIoGRAM exome array study were analyzed through logistic regression with an additive model adjusting for study-specific covariates and principal components of ancestry as appropriate. Top variants from UK Biobank that were not available for analysis in the CARDIoGRAM exome array study were then meta-analyzed with data from the 1000 Genomes–imputed CARDIoGRAMplusC4D GWAS[5] in stage 3.

Given differences in effect size units between the UK Biobank stage 1 data and the CARDIoGRAM exome or 1000 Genomes–imputed CARDIoGRAMplusC4D data, both stage 2 and stage 3 meta-analyses were performed via a weighted z-score method, adjusting for an unbalanced ratio of cases to controls. To derive effect size estimates for variants exceeding genome-wide significance, we meta-analyzed logistic regression results using inverse-variance weighting with fixed effects (METAL software)[44]. We set a combined statistical threshold of $P < 5 \times 10^{-8}$ for genome-wide significance. P values reported in analysis stages 1–3 are all two-sided.

Phenome-wide association study. For all 15 new DNA sequence variants associated with CAD in our study, we collaborated with Genomics plc to conduct a PheWAS. This PheWAS used Genomics plc Platform, UK Biobank, and GTEx Consortium eQTL data. The Genomics plc Platform includes PheWAS data across 545 distinct molecular and disease phenotypes, at an integrated set of over 14 million common variants, from 677 GWAS. UK Biobank analyses within the Genomics plc Platform were conducted under a separate research agreement. We selected 25 phenotypes across a range of relevant diseases and metabolic and anthropometric traits from either previously published GWAS data sets or UK Biobank. Complete details of phenotype definitions, sample sizes, and GWAS data sources are provided in **Supplementary Tables 7** and **8**. In the PheWAS, quantitative traits were standardized to have unit variance, imputation was performed to generate results for all variants within the 1000 Genomes reference panel, and P values were recalculated based on a Wald test statistic for uniformity.

Phenotypes were declared to be significantly associated with a risk variant if they achieved a Bonferroni-corrected P value of <0.00013 (0.05/(25 traits × 15 DNA sequence variants)). Phenome scan results were then depicted in a heat map based on the z scores for all variant–disease/trait associations aligned to the CAD risk allele as implemented in the gplots package in R v3.2.0. To identify loci that might influence gene expression, we used previously published *cis*-eQTL mapping data from the Genotype-Tissue Expression (GTEx) Project across 44 tissues[15]. We queried the 15 new variants identified in our study for overlap with genome-wide significant variant–gene pairs from the GTEx portal.

Allele-specific expression analysis. Allele-specific expression (ASE) data from the GTEx Project were obtained from dbGaP (accession phs000424.v6.p1). The generation of these data is summarized in ref. 15 and relied on methods described earlier[45]. In brief, only uniquely mapping reads with base quality ≥10 at a SNP were counted, and only SNPs covered by at least eight reads are reported. For *ARHGEF26* p.Val29Leu, ASE counts were available for 20 heterozygous individuals. A two-sided binomial test was used to identify SNPs with significant allelic imbalance in each individual, and Benjamini–Hochberg-adjusted P values were calculated across all sites measured in an individual.

Luciferase reporter assays. HUVECs heterozygous at rs12493885 from European-ancestry donors were identified by SNP genotyping. A 2.9-kb genomic fragment spanning the region from upstream of *ARHGEF26* to exon 2 (rs12493885) was cloned into the pMiniT 2.0 vector (New England BioLabs) using genomic DNA from heterozygous HUVECs as a template and sequenced for the reference and alternative alleles. The reference and alternative haplotypes located from −2,516 to +2 with respect to *ARHGEF26* (NC_000003.12:154,119,477–154,121,994) were amplified from the 2.9-kb region by PCR with primers designed to create 5′ NheI and 3′ HindIII restriction sites in the PCR products. The amplified fragments were subcloned between the NheI and HindIII sites of the promoterless firefly luciferase (*luc2*) expression vector pGL4.10 (Promega) to create two plasmids: pGL4.10-Ref and pGL4.10-Alt. Promoterless pGL4.10-control and pGL4.73[*hRluc*/SV40] vector containing the *Renilla* luciferase *hRluc* reporter gene and an SV40 early enhancer/promoter were used as the negative control and co-reporter, respectively. HEK293 cells, HAECs, and HUVECs were cotransfected with equal amounts of *luc2* expression plasmid (pGL4.10-control, pGL4.10-Ref, and pGL4.10-Alt) and pGL4.73 vector by Lipofectamine 2000. Cells were collected at 48 h after transfection followed by a Dual-Glo Luciferase Assay (Promega) to measure firefly and *Renilla* luciferase activities. Firefly luciferase activity was normalized to *Renilla* luciferase activity in the same sample and is expressed as the fold change relative to the pGL4.10-control group.

Nucleotide-exchange assays. Full-length human ARHGEF26 (wild type or Leu29) and human RhoG (residues 1–188) proteins, both with N-terminal His-SUMO tags, were expressed in *Escherichia coli* BL21 (DE3) cells in TB medium. Nucleotide-exchange assay samples were prepared in buffer containing 10 mM HEPES pH 7.4, 150 mM NaCl, 1 mM $MgCl_2$, 0.5 µM MANT-GTP, and 2 mM TCEP with 1 µM ARHGEF26. Just before reading, RhoG protein, preloaded with GDP, was added to a final concentration of 0.4 µM. MANT-GTP fluorescence was monitored for 60 min on a SpectraMax M2 at 37 °C using an excitation wavelength of 280 nm and an emission wavelength of 440 nm with a cutoff of 435 nm. Fluorescence data were imported into Prism GraphPad for analysis.

Functional characterization of *ARHGEF26* p.Val29Leu in arterial tissue. To investigate the functional effects of *ARHGEF26* p.Val29Leu (rs12493885), we knocked down the expression of endogenous ARHGEF26 in cultured HAECs and HCASMCs by RNA interference. We then overexpressed wild-type or mutant (Leu29) ARHGEF26 resistant to siRNA and measured leukocyte transendothelial migration, leukocyte adhesion on endothelial cells, and HCASMC proliferation *in vitro*. We also evaluated the degradation of wild-type and Leu29 ARHGEF26 with a cycloheximide chase assay and immunoblotting. Additional details on experimental techniques are provided in the **Supplementary Note**.

A **Life Sciences Reproducibility Summary** for this paper is available.

Data availability. Stage 2 and stage 3 data contributed by CARDIoGRAM Exome and CARDIoGRAMplusC4D investigators are available online (http://www.CARDIOGRAMPLUSC4D.org/). The genetic and phenotypic UK Biobank data are available upon application to the UK Biobank (https://www.ukbiobank.ac.uk/). Genotype-Tissue Expression Project data are available online (http://gtexportal.org/).

42. Wain, L.V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir. Med.* **3**, 769–781 (2015).
43. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
44. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
45. Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).

doi:10.1038/ng.3914