A large-scale multi-ethnic genome-wide association study of coronary artery disease

Themistocles Assimes (tassimes@stanford.edu)
Stanford University School of Medicine https://orcid.org/0000-0003-2349-0009

Catherine Tcheandjieu
Stanford University

Xiang Zhu
The Pennsylvania State University https://orcid.org/0000-0003-1134-6413

Austin Hilliard
VA Palo Alto Health Care System

Shoa Clarke
Stanford University School of Medicine https://orcid.org/0000-0002-6592-1172

Valerio Napolioni
University of Camerino

Shining Ma
Stanford University School of Medicine

Huaying Fang
Stanford University School of Medicine

Bryan R Gorman
Massachusetts Veterans Epidemiology and Research Information Center, Veterans Affairs Boston Healthcare System

Kyung Min Lee
Edith Nourse Rogers Memorial Veterans Hospital https://orcid.org/0000-0001-8995-0448

Fei Chen
University of Southern California

Saiju Pyarajan
Massachusetts Veterans Epidemiology and Research Information Center, Veterans Affairs Boston Healthcare System

Rebecca Song
VA Boston Healthcare System

Mary Plomondon
Rocky Mountain Regional VA Medical Center

Thomas Maddox
Washington University School of Medicine

Stephen Waldo

Rocky Mountain Regional VA Medical Center

Nasa Sinnott-Armstrong

Stanford University https://orcid.org/0000-0003-4490-0601

Yuk-Lam Ho

Massachusetts Veterans Epidemiology and Research Information Center, Veterans Affairs Boston Healthcare System

Genevieve Wojcik

Johns Hopkins University

Steven Buyske

Rutgers University

Charles Kooperberg

Fred Hutchinson Cancer Research Center

Jeffrey Haessler

Fred Hutchinson Cancer Research Center

Ruth Loos

Ichan School of Medicine at Mount Sinai https://orcid.org/0000-0002-8532-5087

Ron Do

Icahn School of Medicine at Mount Sinai https://orcid.org/0000-0002-3144-3627

Marie Verbanck

Université de Paris

Kumardeep Chaudhary

Icahn School of Medicine, Mount Sinai https://orcid.org/0000-0002-4117-6403

Kari North

University of North Carolina at Chapel Hill

Christy Avery

University of North Carolina

Christopher Haiman

University of Southern California

Loic Le Marchand

University of Hawaii Cancer Center

Lynne Wilkens

University of Hawaii System

Joshua Bis

University of Washington

Hampton Leonard

National Institute on Aging

Botong Shen

National Institute on Aging

Leslie Lange

University of Colorado Anschutz Medical Campus

Ayush Giri

Vanderbilt University Medical Center

Yingchang Lu

Vanderbilt University Medical Center

Ozan Dikilitas

Mayo Clinic https://orcid.org/0000-0002-9906-8608

Iftikhar Kullo

Mayo Clinic https://orcid.org/0000-0002-6524-3471

Ian Stanaway

University of Washington https://orcid.org/0000-0002-0783-0918

Gail Jarvik

University of Washington School of Medicine

Adam Gordon

Department of Pharmacology, Northwestern University Feinberg School of Medicine, Chicago, IL

Scott Hebbring

Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI

Bahram Namjou

Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center
https://orcid.org/0000-0003-4452-7878

Kenneth Kaufman

Cincinnati Children's Hospital Medical Center

Satoshi Koyama

Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University

Kaoru Ito

RIKEN Center for Integrative Medical Sciences https://orcid.org/0000-0003-1843-773X

Kazuyoshi Ishigaki

RIKEN https://orcid.org/0000-0003-2881-0657

Yoichiro Kamatani

Kyoto University Graduate School of Medicine

Noah Tsao

University of Pennsylvania Perelman School of Medicine

Shefali Verma

University of Pennsylvania

Marylyn Ritchie

University of Pennsylvania

Rachel Kember

University of Pennsylvania Perelman School of Medicine

Aris Baras

Regeneron Pharmaceuticals https://orcid.org/0000-0002-6830-3396

Luca Lotta

Regeneron Genetics Center

Pradeep Natarajan

Massachusetts General Hospital https://orcid.org/0000-0001-8402-7435

Michael Levin

University of Pennsylvania

Elizabeth Hauser

Durham VA Health Care System

Donald Miller

Edith Nourse Rogers Memorial Veterans Hospital

Marijana Vujkovic

University of Pennsylvania School of Medicine https://orcid.org/0000-0003-4924-5714

Jennifer Huffman

VA Boston Healthcare System https://orcid.org/0000-0002-9672-2491

Sridharan Raghavan

VA Eastern Colorado Healthcare System

Derek Klarin

Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge https://orcid.org/0000-0002-4636-5780

Jennifer Lee

VA Palo Alto Health Care System

Benjamin Voight

University of Pennsylvania https://orcid.org/0000-0002-6205-9994

Daniel Rader

Perelman School of Medicine at the University of Pennsylvania https://orcid.org/0000-0002-9245-9876

Kyong-Mi Chang

The Cpl Michael J Crescenz VAMC & University of Pennsylvania https://orcid.org/0000-0001-6811-9364

Scott Damrauer

Philadelphia VA Medical Center https://orcid.org/0000-0001-8009-1632

Julie Lynch

VA Salt Lake City Healthcare System https://orcid.org/0000-0003-0108-2127

Peter Wilson

Emory

Hua Tang

Stanford University

Yan Sun

Emory University https://orcid.org/0000-0002-2838-1824
Phil Tsao
Stanford University School of Medicine https://orcid.org/0000-0001-7274-9318
Christopher O'Donnell
VA Boston Healthcare System https://orcid.org/0000-0002-2667-8624
Sekar Kathiresan
Broad Institute https://orcid.org/0000-0002-6724-032X
Danish Saleheen
Columbia University Medical Center https://orcid.org/0000-0001-6193-020X
J. Michael Gaziano
Harvard Medical School at VA Boston Heatlhcare System
Peter Reaven
Phoenix Veterans Affairs Health Care System https://orcid.org/0000-0001-8923-6690
Kelly Cho
VA Boston Healthcare System
Alexander Bick
Vanderbilt University Medical Center
Mariaelisa Graff
University of North Carolina at Chapel Hill
Jie Huang
Peking University School of Public Health https://orcid.org/0000-0002-9036-4304

Article

Keywords: Million Veteran Program, X-chromosome, Common Haplotypes, Risk Stratification, Polygenic Risk Scores, Insulin Resistance

Posted Date: March 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-275591/v1

Version of Record: A version of this preprint was published at Nature Medicine on August 1st, 2022. See the published version at https://doi.org/10.1038/s41591-022-01891-3.

A large-scale multi-ethnic genome-wide association study of coronary artery disease

Catherine Tcheandjieu[1,2], Xiang Zhu[1,3,4,5], Austin T Hilliard[1], Shoa L. Clarke[1,2], Valerio Napolioni[6,7], Shining Ma[3], Kyung Min Lee[8], Huaying Fang[9], Fei Chen[10], Yingchang Lu[11], Noah L. Tsao[12], Sridharan Raghavan[13,14], Satoshi Koyama[15], Bryan R. Gorman[16,17], Marijana Vujkovic[18,19], Derek Klarin[16,20,21,22], Michael G. Levin[18,19], Nasa Sinnott-Armstrong[9,1], Genevieve L. Wojcik[23], Mary E. Plomondon[24,25], Thomas M. Maddox[26,27], Stephen W. Waldo[24,25,28], Alexander G. Bick[29], Saiju Pyarajan[16,30], Jie Huang[16,31], Rebecca Song[16,32], Yuk-Lam Ho[16], Steven Buyske[33], Charles Kooperberg[34], Jeffrey Haessler[34], Ruth J.F. Loos[35], Ron Do[35,36], Marie Verbanck[35,36,37], Kumardeep Chaudhary[36,35], Kari E. North[38], Christy L. Avery[38], Mariaelisa Graff[38], Christopher A. Haiman[10], Loïc Le Marchand[39], Lynne R. Wilkens[39], Joshua C. Bis[40], Hampton Leonard[41,42], Botong Shen[43], Leslie A. Lange[44,45,46], Ayush Giri[47,48], Ozan Dikilitas[49], Iftikhar J. Kullo[49], Ian B. Stanaway[50], Gail P. Jarvik[51,52], Adam S. Gordon[53], Scott Hebbring[54], Bahram Namjou[55,56], Kenneth M. Kaufman[55], Kaoru Ito[15], Kazuyoshi Ishigaki[57], Yoichiro Kamatani[57,58], Shefali S. Verma[59,60], Marylyn D. Ritchie[59,60], Rachel L. Kember[61,18], Aris Baras[62], Luca A. Lotta[62], Regeneron Genetics Center[63], CARDIoGRAMplusC4D Consortium[63], Biobank Japan[63], Million Veteran Program[63], Sekar Kathiresan[64,21,65,66], Elizabeth R. Hauser[67,68], Donald R. Miller[69,70], Jennifer S Lee[1,71], Danish Saleheen[72,18], Peter D. Reaven[73,74], Kelly Cho[16,30], J. Michael Gaziano[16,30], Pradeep Natarajan[21,75,64], Jennifer E. Huffman[16], Benjamin F. Voight[18,76,59,77], Daniel J Rader[19], Kyong-Mi Chang[18,19], Julie A. Lynch[78,79], Scott M. Damrauer[18,12], Peter W. F. Wilson[80,81], Hua Tang[9], Yan V. Sun[82,83], Philip S. Tsao[1,71], Christopher J. O'Donnell[16,30], Themistocles L. Assimes[1,2]

[1]VA Palo Alto Health Care System, CA, USA, [2]Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, CA, USA, [3]Department of Statistics, Stanford University School of Medicine, CA, USA, [4]Department of Statistics, The Pennsylvania State University, PA, USA, [5]Huck Institutes of the Life Sciences, The Pennsylvania State University, PA, USA, [6]School of Biosciences and Veterinary Medicine, University of Camerino, Italy, [7]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, CA, USA, [8]VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System, UT, USA, [9]Department of Genetics, Stanford University School of Medicine, CA, USA, [10]Department of Preventive Medicine, Center for Genetic Epidemiology, University of Southern California, California, USA, [11]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, TN, USA, [12]Department of Surgery, University of Pennsylvania Perelman School of Medicine, PA, USA, [13]Medicine Service, VA Eastern Colorado Healthcare System, CO, USA, [14]Department of Medicine, University of Colorado Anschutz Medical Campus, CO, USA, [15]Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Kanagawa, Japan, [16]VA Boston Healthcare System, MA, USA, [17]Booz Allen Hamilton, VA, USA, [18]Corporal Michael J. Crescenz VA Medical Center, PA, USA, [19]Department of Medicine, University of Pennsylvania Perelman School of Medicine, PA, USA, [20]Center for Genomic Medicine, Massachusetts General Hospital, MA, USA, [21]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, MA, USA, [22]Division of Vascular Surgery and Endovascular Therapy, University of Florida School of Medicine, FL, USA, [23]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, MD, USA, [24]Department of Medicine, Rocky Mountain Regional VA Medical Center, CO, USA, [25]CART Program, VHA Office of Quality and

Patient Safety, DC, USA, [26]Healthcare Innovation Lab, JC HealthCare/Washington University School of Medicine, MO, USA, [27]Division of Cardiology, Washington University School of Medicine, MO, USA, [28]Division of Cardiology, University of Colorado School of Medicine, CO, USA, [29]Department of Biomedical Informatics, Division of Genetic Medicine, Vanderbilt University Medical Center, [30]Department of Medicine, Brigham Women's Hospital, Harvard Medical School, MA, USA, [31]Department of Global Health, Peking University School of Public Health, China, [32]Department of Epidemiology, Boston University School of Public Health, MA, USA, [33]Department of Statistics, Rutgers University, NJ, USA, [34]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, WA, USA, [35]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, NY, USA, [36]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, NY, USA, [37]EA 7537 BioSTM, Université de Paris, France, [38]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, NC, USA, [39]Cancer Epidemiology Program, University of Hawaii Cancer Center, University of Hawaii, Hawaii, USA, [40]Department of Medicine, Cardiovascular Health Research Unit, University of Washington, WA, USA, [41]Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, MD, USA, [42]Data Tecnica Int'l, LLC, MD, USA, [43]Health Disparities Research Section, National Institute on Aging, National Institutes of Health, MD, USA, [44]Department of Medicine, Division of Biomedical Informatics and Personalized Medicine, CO, USA, [45]Lifecourse Epidemiology of Adiposity and Diabetes (LEAD) Center, CO, USA, [46]Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, CO, USA, [47]Department of Medicine, Division of Epidemiology, Vanderbilt University Medical Center, TN, USA, [48]Department of Obstetrics and Gynecology, Division of Quantitative Sciences, Vanderbilt University Medical Center, TN, USA, [49]Department of Cardiovascular Medicine, Mayo Clinic, MN, USA, [50]Department of Medicine, Division of Nephrology, University of Washington, WA, USA, [51]Department of Medicine, Medical Genetics, University of Washington School of Medicine, WA, USA, [52]Department of Genome Sciences, University of Washington School of Medicine, WA, USA, [53]Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, IL, USA, [54]Center for Precision Medicine Research, Marshfield Clinic Research Institute, WI, USA, [55]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, OH, USA, [56]Department of Pediatrics, University of Cincinnati College of Medicine, OH, USA, [57]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Kanagawa, Japan, [58]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences - the University of Tokyo, Japan, [59]Department of Genetics, University of Pennsylvania Perelman School of Medicine, PA, USA, [60]Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine, PA, USA, [61]Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, PA, USA, [62]Regeneron Genetics Center, NY, USA, [63]consortium name (affiliation does not apply), [64]Department of Medicine, Harvard Medical School, MA, USA, [65]Department of Medicine, Cardiology Division, Massachusetts General Hospital, MA, USA, [66]Verve Therapeutics, MA, USA, [67]Cooperative Studies Program Epidemiology Center-Durham, Durham VA Health Care System, NC, USA, [68]Department of Biostatistics and Bioinformatics, Duke University Medical Center, NC, USA, [69]Center for Healthcare Organization and Implementation Research, Bedford VA Healthcare System, MA, USA, [70]Center for Population Health, Department of Biomedical and Nutritional Sciences, University of Massachusetts, MA, USA, [71]Department of Medicine, Stanford University School of Medicine, CA, USA, [72]Department of Medicine, Division of Cardiology, Columbia University, NY, USA, [73]Phoenix VA Health Care System, AZ, USA, [74]College of Medicine, University of Arizona, AZ, USA, [75]Cardiovascular Research Center, Massachusetts General Hospital, MA,

USA, [76]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, PA, USA, [77]Institute of Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine, PA, USA, [78]VA Salt Lake City Health Care System, UT, USA, [79]College of Nursing and Health Sciences, University of Massachusetts, MA, USA, [80]Atlanta VA Medical Center, GA, USA, [81]Division of Cardiology, Emory University School of Medicine, GA, USA, [82]Atlanta VA Health Care System, GA, USA, [83]Department of Epidemiology, Emory University Rollins School of Public Health, GA, USA

Yan V. Sun, Philip S. Tsao, Christopher J. O'Donnell, & Themistocles L. Assimes jointly supervised this work

Corresponding Author:
Themistocles L. Assimes, MD PhD FAHA
1070 Arastradero Rd; Suite 300
Palo Alto, CA 94304
Phone: 650-498-4154
tassimes@stanford.edu

Abstract

Coronary artery disease (CAD) is a leading cause of death, yet its genetic determinants are not fully elucidated. We report a multi-ethnic genome-wide association study of CAD involving nearly a quarter of a million cases, incorporating the largest cohorts to date of Whites, Blacks, and Hispanics from the Million Veteran Program with existing studies including CARDIoGRAMplusC4D, UK Biobank, and Biobank Japan. We verify substantial and nearly equivalent heritability of CAD across multiple ancestral groups, discover 107 novel loci including the first nine on the X-chromosome, identify the first eight genome-wide significant loci among Blacks and Hispanics, and demonstrate that two common haplotypes are largely responsible for the risk stratification at the well-known 9p21 locus in most populations except those of African origin where both haplotypes are virtually absent. We identify 15 loci for angiographically derived burden of coronary atherosclerosis, which robustly overlap with the strongest and earliest loci reported to date for clinical CAD. Phenome-wide association analyses of novel loci and externally validated polygenic risk scores (PRS) augment signals from the insulin resistance cluster of risk factors and consequences, extend previously established pleiotropic associations of loci with traditional risk factors to include smoking and family history, and confirm a substantially reduced transferability of existing PRS to Blacks. Downstream integrative genomic analyses reinforce the critical role of endothelial, fibroblast, and smooth muscle cells within the coronary vessel wall in CAD susceptibility. Our study highlights the value of a multi-ethnic design in efficiently characterizing the genetic architecture of CAD across all human populations.

Introduction

Remarkable progress in the prevention and treatment of coronary artery disease (CAD) has been made over the last half century. Yet, the rate of decrease in the age-adjusted prevalence of CAD has slowed substantially in the last decade, and CAD remains the leading cause of death worldwide[1]. Sizeable differences in the age-adjusted fatality rates of CAD persist between men and women and among the major racial/ethnic groups in the US with non-Hispanic Black men persistently demonstrating the highest risk of fatal CAD[2]. Although health care disparities play an important role in explaining these differences[3], the degree to which genetics contribute remains unclear, in part due to limited genome-wide studies in non-White populations[4,5]. A persistent need exists to further understand both the between-population and the population-specific genetic causes of CAD as an avenue towards improved risk prediction and the development of novel therapies.

Large-scale population genetic studies provide an opportunity to improve our understanding of the inherited basis of complex traits. Twin studies report a heritability of 40-60% for fatal CAD[6,7] and genome-wide association studies (GWAS) to date have identified 208 susceptibility loci[8,9]. These loci explain a modest fraction (~15%) of this heritability, have largely been identified in European populations, and are exclusively autosomal[8,9]. Approximately one half of established loci appear to confer risk through effects on traditional risk factors[8-10]. A preponderance of these loci implicates lipids and blood pressure with fewer links to other risk factors[8,10]. Several loci discovered in Europeans have also reached genome-wide significance (GWS) in South and East Asian populations suggesting an overlap in the genetic architecture of CAD across these three racial/ethnic groups[9,11,12]. Yet, 14 years after the discovery of the first susceptibility locus at 9p21, no region has reached GWS in Black or Hispanic populations, which represent a sizable and growing proportion of the US population[13,14].

New multi-ancestry DNA biobanks are poised to fulfill this knowledge gap. Here we describe results from analyses of the Million Veteran Program (MVP)[15], a nationwide cohort drawn from an integrated health care system serving a diverse population including a large number of Blacks and Hispanics. Using these large-scale, multi-ethnic GWAS data meta-analyzed with extant GWAS of CAD from public resources, we extend discovery of CAD loci within and across racial/ethnic groups for both the autosomes and the X-chromosome (X-chr). In addition, we incorporated data from a national registry of cardiac catheterization procedures[16,17] in the discovery of novel CAD loci, to better interpret of downstream mechanism of action of established loci, and the study of polygenic risk scores.

Results

Racial/ethnic diversity in the MVP population

Fig. 1a summarizes new and existing cohorts included in our analyses stratified by racial/ethnic groups and the analytic approach for the clinical CAD phenotype. A majority (90.8%) of veteran participants are male with 95,151 cases and 197,287 controls being classified as non-Hispanic White, hereinafter referred to as White, (73.1%), 17,202 cases and 59,507 controls as non-Hispanic Blacks, hereinafter referred to as Black, (19.2%), and 6,378 and 24,270 as Hispanic (7.7%) **(Supplemental Table 1)**. A majority of cases (85.6%) showed evidence of CAD at the time of enrollment in the MVP (i.e., "prevalent"). The mean age at first evidence of CAD in the electronic health record (EHR) was 63 years with a mean combined EHR follow-up either prior to and/or after enrollment of 10 years.

Equivalent heritability across multiple racial/ethnic groups

We first estimated the SNP-based heritability using GREML-LDMS-I[18] in equally sized subsets of Whites, Blacks with the least European admixture, and Hispanics with the least

African admixture from MVP, as well as Japanese participants from Biobank Japan after matching on the age of onset and severity of disease of cases and the age of controls observed among Hispanics (**Methods, Fig. 1b, Supplementary Table 2**). Assuming a prevalence of CAD of 8.2%, 6.5%, 4.9%, and 6.0% in the same populations[19,20], we derived roughly equivalent heritability on the liability scale of 36.3% (±7.0%), 30.0% (±8.1%), 32.6% (±3.9%), and 36.0% (±5.4%), respectively (**Fig. 1c-d**).

GWAS followed by meta-analysis in Whites and trans-ethnic meta-analysis identifies 107 novel loci

We conducted a GWAS of autosomes and X-chr stratified by race/ethnicity of all MVP participants. The genomic control inflation (λ) for these GWAS was 1.360 (Whites), 0.988 (Blacks), and 0.986 (Hispanics). The LD score regression intercept[21] for Whites was 1.077 (±0.014), indicating a majority of the inflation was polygenic in nature. We found a high rate of replication of established loci as of 2019[8] among Whites with 100% of 163 lead SNPs being directionally concordant, 85.9% reaching nominal significance ($p<0.05$), and 36 (22.1%) reaching genome-wide significance (GWS). Effect sizes were also highly correlated (Pearson rho=0.94) (**Supplementary Tables 3**).

The GWAS of MVP Whites was followed by a meta-analysis with existing predominantly European-ancestry GWAS from CARDIoGRAMplusC4D[22] and the UK Biobank[8] yielding 37 novel loci at GWS (lead SNP $P<5\times10^{-8}$), including five on the X-chr (**Fig. 2, Supplementary Table 4**). Our trans-ethnic meta-analysis further incorporated the GWAS data in MVP Blacks and MVP Hispanics as well as GWAS in the Biobank Japan[23], yielding an additional 66 novel autosomal loci and four more novel loci on the X-chr (**Fig. 2, Supplementary Table 5**).

Trans-ethnic mapping and two-stage joint meta-analysis identifies the first CAD loci in Black and Hispanic populations

Our GWAS of MVP Blacks and Hispanics did not yield any GWS loci passing quality control within either population in isolation. However, XPEB[24], an empirical Bayes mapping approach that adaptively incorporates trans-ethnic evidence with an 'auxiliary base GWAS' (CAD meta-analysis in Whites), identified 37 SNPs at 16 loci in MVP Blacks and 157 SNPs at 38 loci in MVP Hispanics with a local False Discovery Rate (FDR) < 0.05 (**Supplementary Table 6**). All but one of the loci identified by XPEB were GWS in the base GWAS (meta-analysis in Whites).

We then extended our GWAS analysis of MVP Blacks and MVP Hispanics to include additional data from multiple external cohorts for the most promising variants from our GWAS ($P<1\times10^{-5}$) as well as all SNPs identified by XPEB **(Supplementary Text, Tables 6-12**). A two-stage joint meta-analysis of all promising SNPs yielded the first five GWS loci in Blacks and the first three in Hispanics (**Fig. 2a, Supplementary Table 11**), all of which have been previously established in Whites[13]. Three out of five loci in Blacks (*LPA*, *FGD5*, and *LPL*) included GWS signals generated by low-frequency African specific genetic variation (**Supplementary Fig. 1**). The SNPs identified through XPEB and trans-ethnic evidence include loci with more moderate allelic effects; therefore, *a priori,* we do not expect all of them to reach GWS in the much smaller two-stage meta-analysis of Blacks and Hispanics. However, this group of SNPs exhibits significantly higher proportion of directional consistency and correlation of effect sizes between the MVP discovery cohort and the external cohorts, for both Blacks (13 out 15 loci with available data in external cohorts were directionally consistent, binomial $P=0.0032$, Pearson's rho=0.82) and Hispanics (33 out of 36 loci directionally consistent, $P=1.1\times10^{-7}$, rho=0.80).

GWAS of angiographically determined degree of CAD identifies 15 genome-wide significant loci

We conducted the largest GWAS to date of angiographically determined burden of coronary atherosclerosis, defined by number of significantly obstructed coronary arteries, in an analysis of 41,507 MVP participants (**Methods**, **Supplementary Tables 13-14)**. A total of 15 loci reached GWS in the trans-ethnic meta-analysis of which 12 also reached GWS in Whites alone and 1 (*LPL*) in Blacks alone (**Fig. 2b, Supplementary Table 15**). All 15 loci have been previously reported for clinical CAD, and eight (*CDKN2-AS, SORT1, CXCL12, WDR12, PHACTR1, LDLR, KCNE2, ADAMTS7*) were among the 12 earliest loci associated with clinical CAD by GWAS and all but *TGFB1* were identified prior to 2013[13].

Local ancestry and haplotype analysis reveals a protective haplotype at the 9p21 susceptibility locus that is virtually absent among chromosomes of African descent

The well-established susceptibility locus at 9p21 did not reach GWS among Blacks nor among Hispanics even after two-stage meta-analysis involving >27,000 and >12,100 CAD cases, respectively. We explored whether the ancestral origin of the high-risk haplotype block at 9p21 among Blacks influences the observed magnitude of association with CAD (**Methods)**. Using RFMix[25], we stratified MVP Blacks into three subgroups based on whether they had inherited two (Black_AFR = +/+, 66.8%), one (Black_AFR = +/-, 29.6%), or zero (Black_AFR = -/-, 3.6%) chromosomal 9p21 segments from African (AFR) ancestry when compared to European (EUR) ancestry through admixture (**Supplementary Fig. 2a**). Only the first two of these three subgroups had adequate power to detect an association at 9p21. Between these two, we found notably stronger associations with CAD among Blacks with one AFR segment (Black_AFR = +/-, lowest P=6.4×10^{-7}) despite a sample size of less than one half of Blacks with two AFR segments (Black_AFR = +/+, lowest P=1×10^{-3}) (**Supplementary Fig. 2b, Supplementary Table 16**). Haplotype analysis at 9p21 (**Methods**) revealed a largely non-overlapping set of haplotypes when comparing Whites to Blacks with zero 9p21 segments derived from EUR (**Fig. 3a, Supplementary Table 17**). Two haplotypes (AACATT, GGTTCA) account for a large majority (87%) of observed haplotypes among Whites but these same two haplotypes are virtually absent (<0.5%) among the majority of Blacks with no EUR admixture at 9p21. Our haplotype trend regression analysis suggests the second most common haplotype (GGTTCA) is associated with an increased risk for CAD when compared to the most common haplotype among Whites (AACATT) and these two haplotypes are largely responsible for the risk-stratifying potential of this locus within this group (**Fig. 3b-c, Supplementary Table 17**). Analyses of the frequency of the same haplotypes in the 1000G populations suggest that these 2 haplotypes likely provide a majority of the risk-stratifying potential in all but West African populations where both haplotypes are virtually absent (**Supplementary Table 18**). Supporting these observations, we found that a single SNP (rs1333050) reaches GWS among Hispanics when GWAS analysis is restricted to the subgroup of Hispanics with no AFR admixture at 9p21 despite a very substantial reduction in sample size (**Supplementary Table 19**).

Pleiotropy assessment of novel loci strengthens and extends links to traditional risk factors

We explored the potential mechanisms of action of our novel loci by performing an extended phenome-wide association study (PheWAS)[26,27] in MVP of all 107 lead novel SNPs (**Methods**). All but five (95%) of these SNPs were associated with one or more non-CAD phenotypes at an FDR<0.05. A total of 62 (57%) were associated with ≥ 1 traditional risk factor (TRF) for CAD, defined by blood lipid levels/hyperlipidemia (38 loci), blood pressure/hypertension (26 loci), diabetes mellitus (16 loci), body mass index/obesity (14 loci), and/or smoking/tobacco use disorder (eight loci) (**Fig. 4, Supplementary Table 20-22**). Of these 62 loci, 33 (53% of TRF loci, 31% overall) were also associated with one or more TRFs even after excluding CAD cases. The five most pleiotropic loci (*TCF7L2*, *FTO*, *PNPLA3*,

CDK12, and *TDGF1P3*) were linked to a range of 135 to 353 phenotypes while six additional loci (*BPTF*, *DSTYK*, *NPC1*, *IL1F10*, *SETDB1* and *WWP2*) were associated with >50 phenotypes. Of these 11 highly pleiotropic loci, five (*FTO*, *IL1F10*, *PNPLA3*, *TCF7L2*, *TDGF1P3*) were linked to a family history of the same dominant TRF even among MVP participants without CAD. Other phenotypes found to associate frequently with our novel loci included white blood cell related counts (26 loci), cancer (21 loci), renal function (17 loci), platelets (16 loci) and height (14 loci).

Gene and pathway-based association analyses highlight importance of cell cycle, replication, and growth gene-sets as well as endothelial, fibroblast, and smooth muscle cells within the coronary artery in the pathogenesis of CAD

Almost all genes implicated by four gene-based analyses (**Methods**) fell within or very near previously or our newly implicated loci that have reached GWS (**Supplementary Tables 23-25**). Comparing the DEPICT[28] analyses before and after the addition of MVP GWAS of Whites, we found a large majority (95.6%) of the 19,460 genes tested were not implicated in either analyses. Among the 437 genes at FDR<0.05 in the previously published analysis[8], 73% had a similar or lower FDR after the addition of MVP data while the remainder had a higher FDR or were no longer implicated. Adding MVP data also implicated 189 new genes at FDR<0.05. While the probability of a gene being implicated within a tissue relevant to CAD in our predicted gene expression analyses (MetaXcan[29]) increased in tandem with the fraction of the remaining three algorithms that implicated the gene, the proportion was still very low with only 9.3% of the 321 genes implicated by DEPICT, MAGMA[30-32], and RSS-E[33] also being implicated by MetaXcan.

Gene-set enrichment analyses using MAGMA, RSS-E and DEPICT highlight the involvement of many of the same pathways identified through similar analyses in previous large-scale GWAS of CAD (**Supplementary tables 26-28**). A sizable fraction of the most significant curated gene-sets tested by MAGMA, RSS-E, as well as the protein-protein interaction subnetworks tested by DEPICT involve basic cellular processes/gene networks responsible for cell cycle, division/replication, and growth. For at least some of these gene-sets/networks, the 'hub gene' includes a gene mapped to either one of our novel loci (e.g., *CDKN1A*) or within previously established loci (e.g., *TCF21*).

We implemented MAGMA and DEPICT to prioritize cells and systems/tissues based on our GWAS meta-analysis of Whites (**Methods**, **Fig. 5, Supplementary Tables 29-32**). MAGMA identified 15 of 54 (27%) GTEx tissues, 95 of 729 (13%) Mouse Atlas cell types, 27 of 119 (22%) Tubula Muris FACS, and 19 out of 75 (25%) Tubula Muris Droplet cells as enriched in the expression of genes associated with CAD. A total of 35 out of 209 tissues/cell types reached an FDR<0.05 in DEPICT. MAGMA gene property analyses of a wide range of single-cell RNA datasets from mice as well as a more restricted set of cell types in humans highlight the relevance of the endothelial, stromal/fibroblast, and smooth muscle cells in the pathogenesis of CAD (**Fig. 5a-b**) with DEPICT reinforcing these findings and further delivering strong signals for hepatocytes and adipocytes (**Fig 5d**). The most significant system/tissue for both algorithms involved arteries, with MAGMA producing a top signal specifically for the 'coronary artery', a tissue almost exclusively made up of endothelial, stromal/fibroblast, and smooth muscle cells (**Fig. 5c, f**). In DEPICT, these findings were supported by significant associations in related vasculature (e.g., veins, portal system). Additional tissues prioritized across both algorithms included: i. components of the female reproductive system rich in smooth muscles (e.g., uterus, cervix, and the fallopian tube) with DEPICT implicating the myometrium specifically, ii. the esophagus and the sigmoid colon (MAGMA) as well as other components of the upper GI track including the liver and the pancreas (DEPICT), iii. the steroidogenic endocrine tissues of the ovary (MAGMA) and the adrenal cortex (DEPICT), iv. the lung, v. the bladder, and vi. multiple sources and types of adipose tissue (DEPICT). Findings unique to DEPICT include a signal involving the 'aortic

valve' second only to 'arteries' in strength, the spleen, and a cluster of four signals involving joint related tissues.

Externally validated polygenic risk scores associate with CAD and burden of coronary atherosclerosis, but show variable degradation of performance across racial/ethnic groups

Four externally derived polygenic risk scores (PRS) of CAD (**Methods**) predicted clinical CAD status in all racial/ethnic groups (**Fig. 6a, Supplementary Tables 33-34**). The LDPred[34] and MetaGRS[35] PRSs generated the highest odds ratios (ORs) per standard deviation (SD) increase of PRS with differences between the four scores least evident among Blacks. ORs were higher among the subset of cases with EHR evidence of myocardial infarction and/or a revascularization procedure and subjects with an age of onset of CAD below the median. The former subgroup also allowed for a direct comparison of the performance of the LDPred and the MetaGRS PRS to that observed in the validation cohorts in the UK Biobank Whites. Based on the ratio of the log ORs, this comparison demonstrated a relative efficiency of the PRS of 75% to 80% when transferred to MVP Whites and as low as ~30-35% when transferred to Blacks consistent with prior studies[35,36]. ORs were notably lower among the subset of cases with first evidence of CAD after enrollment in MVP (i.e., incident cases) as compared to cases with first even prior to enrollment (i.e., prevalent), a finding that is also consistent with prior studies[35,36]. The four PRSs were also near linearly associated with burden of CAD among Whites with a similar ranking of performance to that observed for clinical events (**Fig. 6b**). Overall, we found the MetaGRS slightly but consistently outperformed LDPred PRS on the basis of the point estimate of the OR with the most notable difference between the two observed among Hispanics.

Phenome-wide association study of PRS among controls in MVP extends links between polygenic risk scores and risk factors of CAD to all risk factors including smoking and family history

We explored factors through which a PRS mediates susceptibility to CAD by conducting a PheWAS of the MetaGRS among MVP participants. To minimize ascertainment bias of risk factors, the PheWAS was restricted to MVP White controls with further exclusion of subjects with evidence of peripheral arterial disease (PAD) or ischemic stroke (IS). After excluding only subjects with CAD, we found evidence that a higher PRS of CAD was associated with a higher risk of non-coronary related atherosclerosis complications (stroke, PAD, abdominal aneurysm, erectile dysfunction) and all TRFs **(Supplementary Tables 35)**. When further excluding subjects with PAD or IS, associations with all TRFs were sustained (**Fig. 6c, Supplementary Tables 36**). In addition to 'tobacco use disorder', we found evidence of a more widespread predisposition to substance abuse. Extending the PheWAS to self-reported family history revealed not only an association with a family history of CAD but also with a family history of high cholesterol, hypertension, and diabetes. Extending the PheWAS to physical exam measures and laboratory measurements not only reinforce our Phecode findings through robust associations with analogous quantitative traits but also linked the PRS to renal function. Additional non-TRF associations included three lab indices derived from a complete blood count and several other commonly measured chemistries as well as hypothyroidism, viral hepatitis C, multiple common disorders of the eyes (cataract, glaucoma, blindness/low vision), and shorter height.

Discussion

We report the largest multi-ethnic GWAS for CAD to date incorporating nearly a quarter of a million cases from four racial/ethnic groups. We increase the total number of GWS loci for CAD by ~50% to 315 through the identification of 107 novel loci including nine X-chromosome loci. Our analysis in large numbers of participants from multiple racial/ethnic groups provides several important insights on the genetic architecture of CAD.

First, we document a largely equivalent degree of heritability of CAD across multiple ancestries using a uniform and unbiased approach of estimation among unrelated individuals. Our results suggest the balance between genetic and environmental determinants of CAD is equivalent across all major racial/ethnic groups in developed countries such as the US and Japan. Our absolute estimates of heritability are somewhat lower than the range previously reported in twin studies for fatal CAD[6,7], but the remaining heritability may be captured through future large-scale whole genome sequencing association studies of more severe disease[37].

Second, the CAD susceptibility loci of populations with a high proportion of either African and/or Native American ancestry are likely to overlap substantially with those identified to date in other populations, as the first eight loci reaching GWS in our African and Hispanic American populations have all been previously identified among the initial GWAS involving White, South Asian, and/or East Asian populations. Further supporting the presence of such overlap is the number of established loci implicated by XPEB and the degree of replication/correlation observed for these loci in our external Stage-2 Black and Hispanic cohorts. As these cohorts expand in size, many of the XPEB loci may reach GWS.

Third, GWAS in admixed populations may be leveraged to better understand the source of heterogeneity of effects across racial/ethnic groups at some CAD loci. We show this for the widely replicated susceptibility locus at 9p21[38]. The high-risk region harbors a 50kb haplotype block containing many common SNPs with allele frequencies near 50% in Whites[39]. Common SNPs in the same haplotype block are GWS in South and East Asians[11,12] but not in Blacks or Hispanics. Taking advantage of the admixture among our Black and Hispanic populations, we provide compelling evidence for the presence of a protective haplotype at this locus which is common in all but African descent chromosomes where it is virtually absent. Further, the presence of an association signal among Blacks and Hispanics at 9p21 is dependent on the inheritance of non-African haplotypes in the region. Thus, the 9p21 locus is unlikely to ever serve as a key risk stratifying locus among populations with a high proportion of African ancestry at this locus, in stark contrast to its prominent risk-stratifying role in all other ancestral populations. As the number of CAD loci reaching GWS grows over time in admixed populations, similar approaches may be useful to gain insight on causal haplotypes and heterogeneity of effects across major racial/ethnic groups.

The degree to which genetic variation underlies sex differences in the incidence of CAD remains unclear. Initial GWAS of CAD did not detect sex differences in the magnitude of effects of autosomal susceptibility loci between men and women[40] but more recent GWAS of adiposity-related traits such as waist-to-hip ratio as well as a study examining a PRS of CAD in the UK biobank have identified compelling sex differences[41,42]. While gonadal hormones undoubtedly serve as a major determinant of sex-differences in obesity and related traits, the X-chr may further contribute to sex differences in the rates of CAD through dosage effects on adiposity, lipid level and inflammation-related traits[43]. Determining the contribution, if any, of the novel and X-chr loci to sex-differences in the rates of CAD will require the study of additional very large populations of females with CAD.

Our GWAS of angiographically derived burden of coronary atherosclerosis did not identify novel CAD loci. Larger sample sizes may prove more fruitful, and our current results suggest that a large fraction of the initial loci uncovered for CAD increase risk of clinical disease by promoting coronary plaque rather than predisposing to plaque rupture or thrombosis. That hypothesis is consistent with prior reports examining the relationship between early GWAS loci for CAD and subclinical coronary atherosclerosis[44].

PheWAS for our 107 lead novel SNPs continue to suggest that about one half of CAD loci influence risk through risk factors[8-10]. We note a more prominent role of highly pleiotropic loci operating through the obesity, insulin resistance, and diabetes risk axis among our novel loci including the top GWAS signals for obesity (*FTO)*[45], diabetes (*TCF7L2)*[46], and

non-alcoholic fatty liver disease *(PNPLA3)*[47], as well as the previously known lipid loci *TDGF1P3* and *NPC1* which are also associated with metabolic indices[48,49]. Furthermore, we note the appearance of loci associated with smoking status. These findings for single novel SNPs were consistent with our PheWAS of the externally derived MetaGRS[35] which now provides evidence that a genome-wide PRS for CAD incorporates a strong readout for predisposition to every well-established TRF including a family history of not only CAD but also risk factors for CAD.

Our gene-based association analyses expand on prior efforts to identify the most likely causal gene within a susceptibility locus. Despite substantially larger sample sizes and an improvement in analytic methods, it remains a challenge to unambiguously identify a causal gene within susceptibility loci. Our results highlight the need for integrative and orthogonal genomic methods to reliably identify the most likely causal gene and its putative mechanism within specific tissues[50].

Our gene-set enrichment analyses continue to highlight well-established relevant biology in CAD such as lipoprotein metabolism/transport, vessel wall development/structure/remodeling, cellular migration/interactions with the extra-cellular matrix, and bleeding/coagulation. The results also point to an enrichment of pathways related to basic cellular processes/gene networks responsible for cell cycle, division/replication, and growth. This observation is buttressed by our PheWAS findings which link nearly ~1/3 of the 107 novel loci to either a cancer or to height, traits directly relevant to these basic cellular processes. Intriguingly, others have recently documented the genetic basis of longstanding epidemiologic correlations between height, CAD, and cancer[51,52]. We suspect that these links reflect the prominence of these processes in tissues and cell types most relevant to CAD such as the de-differentiation, proliferation, and migration of vascular smooth muscle cells, fibroblasts, and fibromyocytes within the vascular wall in response to the development of coronary atherosclerosis[50,53,54].

Cell types prioritized for CAD include endothelial cells, fibroblasts, smooth muscle cells, hepatocytes, and adipocytes using two independent analytic algorithms. The first three comprise the vast majority of the cells in the normal vasculature[55] consistent with top tissue signals observed for these tissues as well as the vessel rich lung. Strong signals involving the aortic valve, joints, joint capsule, synovial membrane, and cartilage may reflect shared gene networks expressed in these subtypes of connective tissue[55]. The aortic valve is not only primarily made up of fibroblast-like interstitial cells, but also enveloped by a single layer of endothelial cells[55]. Signals involving the female reproductive tract, the GI tract, and the bladder may reflect the smooth muscle cell make up in these tissues[55] with signals in the pancreas and the small intestine possibly further amplified by the key role these tissues play in the digestion and absorption of dietary lipids and cholesterol[56]. Lastly, strong signals in the liver, adrenal gland, and serum likely reflect the dominance of cholesterol-related gene networks within these tissues.

Our testing of externally derived PRSs of CAD in multi-ethnic MVP participants confirms previously observed patterns with greater precision and some additional insights. First, genome-wide PRSs of CAD substantially outperform genetic risk scores restricted to genome-wide significant loci. Second, higher ORs are observed for prevalent vs. incident, younger vs. older onset, and more severe (e.g., acute myocardial infarction and/or revascularization procedure) vs. less severe manifestations of CAD. These patterns likely reflect a higher average burden of CAD in one subgroup of cases when compared to the other with a proportional increase in the mean PRS for that subgroup. This hypothesis is supported by the strong linear relationship we observed between the same PRSs and the number of obstructed coronary arteries, a proxy for burden. Third, we observe a reduction in predictive performance of PRSs derived and validated externally among largely European participants when these scores are transferred to MVP. The reduction in performance is most evident

among MVP Blacks but is also observed to a smaller degree among MVP Whites and Hispanics, consistent with previous validation reports in smaller multi-ethnic EHR cohorts[36,57]. Overall, our results underscore the need to develop data and/or methods that eradicate such differences in performance to minimize the potential for exacerbating existing health disparities as PRSs are implemented into clinical practice[5].

In conclusion, our large-scale multi-ethnic GWAS provides important new insights into the genetic basis of CAD. We confirm similar heritability across multiple racial/ethnic groups, substantially extend discovery particularly through the addition of non-White populations, compare and contrast the genetic determinants of disease between admixed and non-admixed groups, and strengthen genetic links between established risk factors and CAD. This progress brings us closer to precision medicine approaches for CAD across the diversity spectrum, but follow-up studies are needed to improve the transferability of PRS for CAD, to identify and understand mechanisms of causal genes, and to develop trans-ethnic and racial/ethnic-specific novel therapies based on this understanding.

Online Methods

Design

The design of the MVP has been previously described[15]. Briefly, active users of the Veterans Health Administration (VA) of any age have been recruited from more than 60 VA Medical Centers nationwide since 2011 with current enrollment at >800,000. Informed consent is obtained from all participants to provide blood for genomic analysis and access to their full EHR within the VA prior to and after enrollment including inpatient International Classification of Diseases (ICD9/10) diagnosis codes, Current Procedural Terminology (CPT) codes, clinical laboratory measurements, and reports of diagnostic imaging modalities. The EHR is continuously being integrated with MVP genomic data and access to these linked coded data is provided to approved investigators. All participants are also asked to optionally complete two short surveys, the Baseline and Lifestyle questionnaires, designed to augment data contained in the EHR. The study received ethical and study protocol approval from the VA Central Institutional Review Board.

Genetic Data and Quality Control

We genotyped 468,961 multi-ethnic participants who enrolled in MVP between 2011 and 2017 with a customized Affymetrix Axiom array in two batches. The first batch including 359,964 unique samples and the second batch including 108,997 unique samples. Quality control (QC) is extensively described elsewhere[58]. We initially imputed to the 1000 Genomes phase 3 version 5 reference panel (1000G)[59] in each batch of genotyped data separately using EAGLE v2.3[60] and Minimac3[61] before joint imputation was performed in the two batches combined using EAGLE v2.4 and Minimac4. Prior to imputation, variants that were poorly called (genotype missingness > 5%) or that deviated from their expected allele frequency observed in the reference data (1000G) were excluded. Genotyped SNPs were interpolated into the imputation file.

Assignment of Racial/Ethnic Groups

We assigned racial/ethnic group to participants using HARE[62], an algorithm that integrates genetically inferred ancestry with self-identified race/ethnicity. HARE assigned >98% of participants with genotype data to one of four non-overlapping groups: non-Hispanic Whites (Europeans), non-Hispanic Blacks (Africans), Hispanics, and non-Hispanic Asians. The sample size of Non-Hispanic Asians was too small for discovery and was excluded from further analyses[62].

Additional Quality Control for X-chromosome

We implemented additional QC steps for analyses involving the X-chr to minimize risk of false positive associations due to sex-specific genotype calling errors. First, we excluded subjects with suspected XXY (n = 350) and XYY (n = 850) karyotypes based on an analysis of the median logR ratios of nonPAR X and Y chromosome SNP intensities. Second, we quarantined 6,707 out of 17,809 genotyped X-chr SNPs that met one or more of the following criteria: i. out of Hardy-Weinberg equilibrium among females ($P<1\times10^{-6}$); ii. demonstrated differential missingness between cases and controls and/or between males and females ($P<1\times10^{-6}$); iii. demonstrated differential minor allele frequencies between males and females ($P<1\times10^{-6}$); iv. high homology to another chromosome (mostly for the Y-chr within the pseudo-autosomal 3 region). Lastly, we phase and re-imputed the X-chr across all genotyped subjects combined using only the remaining 11,102 SNPs before proceeding with association analyses.

Phenotype

Clinical CAD

We used inpatient and outpatient ICD diagnostic and CPT procedure codes to identify subjects with clinical CAD in MVP. EHR data was available retrospectively before enrollment going back to October 1999 and prospectively after enrollment until mid-August 2018. An individual was classified as a case if he or she had: 1) any admission to a VA hospital with a discharge diagnosis of acute myocardial infarction (AMI) or 2) any procedure code for revascularization of the coronary arteries, or 3) two or more ICD codes for CAD (410 to 414) in at least two different encounters. Individuals with only one ICD code for CAD in a single encounter and no discharge diagnoses for AMI or revascularization procedures were excluded from the analyses. The remaining subjects were classified as controls.

We accessed individual level genetic and phenotypic data for the UK Biobank and implemented the same case-control definitions for clinical CAD used by others[8] to conduct association analyses involving the X-chr.

Angiographic burden of CAD based on number of obstructed vessels

We linked MVP participants to the Veterans Affairs Clinical Assessment, Reporting, and Tracking (CART) Program, a national quality and safety organization for invasive cardiac procedures, to reliably estimate the burden of atherosclerosis among participants who had undergone at least 1 coronary angiogram by October 2018[16]. Data were available retrospectively starting in 2004 in select sites and from all sites by 2010[17]. A total of 31,658 non-Hispanic White, 7,313 non-Hispanic Black, and 2,536 Hispanic participants, a majority of which were subjects with clinical CAD, were found to have at least one evaluation of the degree of angiographically defined coronary atherosclerosis. For each angiogram, we classified an individual's extent of disease to one of the following categories of disease of the native vessels: normal, non-obstructive, 1 vessel, 2 vessel, 3 vessel and/or left main coronary artery disease. Obstructive disease of a native vessel was defined as the presence of at least one lesion >50% or a prior revascularization procedure involving that vessel. Non-obstructive disease of a native vessel was defined as a vessel with at least one stenosis >20% of luminal diameter but no lesion >50%. We modified a previously validated algorithm to derive these classifications by decreasing the threshold of significant disease in a vessel from at least one lesion >70% to one lesion >50%[63]. Entries were filtered to remove those where disease severity was missing or listed as "other", then subjects were removed if they were missing a HARE assignment, date of birth, sex, or had previously received a cardiac transplant. For subjects with multiple angiograms over follow up where at least one reported disease, we assigned severity based on the procedure reporting the most advanced disease. If more than one angiogram reported the same advanced disease, we used the earliest one. Age was

calculated on the date of the cardiac catheterization with the most severe disease for cases and the last normal angiogram for controls.

Statistical Analysis

Genetic Relatedness

We used KING, version 2.0, to identify 20,881 related participants at a 3rd degree or closer[58]. Among these individuals, we preferentially retained 5,289 unrelated cases and 4,909 unrelated non-cases in analyses and excluded the remaining individuals (1,023 cases and 9,601 non-cases).

Analyses of heritability across racial/ethnic groups

We used GREML-LDMS-I as implemented in Genome-wide Complex Trait Analysis (GCTA) 1.93.0beta to estimate the multicomponent narrow sense heritability of CAD in our three HARE-define MVP groups and in the Biobank Japan dataset[18]. GREML-LDMS-I method was adopted because it has been shown to be one of most accurate methods of estimation of heritability when considering common factors that may bias such estimates[64]. To minimize the confounding effects of admixture, we ran heritability analyses in minimally admixed subsets of individuals in each of the HARE groups identified through PCA of MVP data with the 1000G data and selection of White, Black, and Hispanic subjects clustering most closely with the 1000G European, African and Peruvian populations, respectively. Restricted by computing memory requirements, we next selected a random subset of 19,400 subjects from our smallest group of least-admixed Hispanics to run through GREML-LDMS after first implementing additional stringent QC of SNPs for binary traits[65,66]. To minimize the influence of differences in the severity of the cases and the age of controls between racial/ethnic groups on the estimates of heritability, we then selected an approximately equal number of MVP Blacks, MVP Whites, and Japanese from Biobank Japan matched to the Hispanic group on the case-control status, EHR-based estimated age of onset of CAD, the type of case (MI/revascularization versus other), and the age of controls. These sample sizes provided us with >80% power to detect a heritability of at least 7% on the liability scale and 100% of at least 11% assuming a prevalence of disease of 8%[67]. We then ran heritability analyses in each group after applying identical QC procedures. First, SNP dosages were converted to hard-call genotypes, and SNPs that were multi-allelic, had MAC < 3, or genotyping call-rate < 95% were removed. Since CAD case status is a binary trait, SNPs with $p < 0.05$ for Hardy-Weinberg equilibrium or differential missingness in cases vs controls were also removed. LD scores were computed on each autosome using GCTA default settings with an r^2 cutoff of 0.01, and the genome-wide LD score distribution was used to assign SNPs to 1 of 4 LD quartile groups, where groups 1-4 represented SNPs with progressively higher LD scores. Within each LD group, SNPs were further stratified into 6 MAF bins ([0.001, 0.01], [0.01, 0.1], [0.1, 0.2], [0.2, 0.3], [0.3, 0.4], [0.4, 0.5]) and a genetic relatedness matrix (GRM) was constructed from each bin, ultimately creating 24 GRMs. Finally, GCTA --reml was used to fit a model of CAD case status based on the 24 GRMs, with age and sex as covariates. Total observed heritability estimates were transformed to estimate disease liability[49] across a range of presumed CAD prevalence estimates in the general population.

Genome-wide association study in MVP

We performed a GWAS of autosomes for clinical CAD and for coronary angiographic burden of disease within each of the three ethnic groups using logistic and linear regression, respectively, implemented in PLINK 2.0 alpha. Models assumed an additive genetic effect adjusted for sex and the respective first 10 ancestry-specific principal components (PCs). For burden of disease, we further adjusted models for age at the time of angiography. Association tests were performed within each HARE group and across 2 tranches of MVP genotyped data. Thus, six GWAS were performed for each phenotype. Each set of results was filtered

separately using PLINK and EasyQC which removed SNPs with i. racial/ethnic specific imputation $r^2 < 0.4$, ii. invalid OR, p-value and/or SE; iii. multi-allelic SNPs, and iv. SNPs with minor allele count (MAC) <6. METAL[68] (classical standard error approach) was then used to apply a genomic control to each input dataset and meta-analyze GWAS results across genotype releases within each HARE group. For Whites, we also ran METAL with genomic control turned off to create a dataset suitable for LD score regression[21].

X-chr association testing in MVP for both phenotypes was conducted stratified by sex in addition to HARE group. We implemented a standard logistic regression model assuming presence of X-chr inactivation (males coded as 0/2, females as 0/1/2). In the UK Biobank, X-chr analyses were restricted to unrelated subjects of White/European descent (34,541 CAD cases and 261,984 controls).

Meta-analysis with external datasets

We used METAL to conduct 2 fixed-effect meta-analyses for the clinical CAD phenotype[12]. The first involved the MVP Whites with the CARDIoGRAMplusC4D 1000G study[2] and the UK Biobank CAD study[3] and the second further incorporated the MVP Blacks, MVP Hispanics, and Biobank Japan[23]. Genomic control was applied to each input dataset by METAL. This second trans-ethnic meta-analysis was also performed using MR-MEGA[13]. METAL and MR-MEGA were also used to conduct a trans-ethnic meta-analysis of the CART derived burden of CAD. For the X-chr, we used GWAMA[69] to perform a meta-analysis of males and female strata within each HARE group followed by METAL to conduct a meta-analysis of the X-chr data in MVP Whites with the UK Biobank and the X-chr study by CARDIoGRAMplusC4D[70] . Lastly, we used MR-MEGA to conduct a trans-ethnic meta-analysis of the X-Chr through further inclusion of the MVP Blacks, MVP Hispanics, and Biobank Japan.

Definition of a locus including parameters for lead and candidate genetic variants

We used FUMA[30] to define genomic risk loci including independent, lead, and candidate variants. First, independent genetic variants were identified as variants with a *P* below a specific threshold and not in substantial linkage disequilibrium (LD) with each other ($r^2 < 0.6$). Second, variants in LD ($r^2 \geq 0.6$) with an independent variant and with $p < 0.05$ were retained as candidate variants to form an LD block. Third, LD blocks within 500kb of each other were merged into one locus. Lastly, a second clumping of the independent variants was performed to identify the subset of lead SNPs with LD $r^2 < 0.1$ within each locus. For our meta-analyses of Whites alone and our trans-racial/ethnic meta-analyses, we used a UK Biobank release 2b EUR reference panel of genotype data imputed to the UK10K/1000G SNPs created by FUMA including ~17 million SNPs. This panel includes a random subset of 10,000 unrelated subjects among all subjects with genotype data mapped to the 1000G populations based on the minimum Mahalanobis distance. We used the 1000G AFR reference panel of 661 subjects with ~43.7 million SNPs for our Blacks, and the AMR reference panel of 347 subjects with ~29.5 million SNPs for our Hispanics.

Two-stage joint analysis of most promising findings in non-Europeans

We sought replication of all promising genomic risk loci in our MVP Black and MVP Hispanic GWAS for clinical CAD in multiple external datasets. Replication was attempted not only for all lead SNP(s) with $P<1\times10^{-5}$ but also for all other independent and candidate genetic variant members of these loci. In the same external datasets, we also sought replication for all SNP with local FDR < 0.05 from our XPEB analyses as described below.

Definition of a significant and novel locus

A locus was considered GWS if at least one lead genetic variant within it reached a $P<5\times10^{-8}$ in any of the terminal meta-analyses. For meta-analyses involving METAL, the variant also had to lack any significant heterogeneity ($P>5\times10^{-8}$ for test of heterogeneity). A GWS locus

was considered novel if none of its lead, independent, or candidate SNPs overlapped with a SNP that has previously reached GWS for CAD. Novel GWS loci were identified at three stages: i. after the meta-analysis of all GWAS available among Whites, ii. after combining genome-wide summary statistics in Blacks and Hispanics, respectively, with external replication data limited to promising loci, and iii. after trans-ethnic meta-analyses of all summery statistics of GWAS (i.e., not including 2nd-stage data in Blacks and Hispanics). For the trans-ethnic meta-analysis, we first identified novel loci with lead SNPs with no significant heterogeneity using METAL and supplemented these with any additional non-overlapping genome-wide findings identified with MR-MEGA.

Cross-population empirical Bayes method

We implemented the trans-ethnic empirical Bayes method, XPEB[24], for the clinical CAD phenotype. XPEB takes as input *P*-value summary statistics from two GWAS, a target-GWAS that is typically a smaller non-European population of primary interest and a base-GWAS that is typically a much larger GWAS of Europeans and adaptively reprioritizes variants in the target population to compute local false discovery rates. We ran XPEB with the MVP Blacks as the target GWAS and the meta-analysis of MVP Whites, CARDIoGRAMplusC4D, and the UK Biobank as the base-GWAS. We then ran it a second time with the MVP Hispanics as the target GWAS. For both runs, analyses were restricted to genotyped SNPs in the target populations.

Calculation of Polygenic Risk Scores

We derived four PRS for CAD of increasing complexity: i. a weighted PRS restricted to a curated list of up to 163 independent SNPs having reached GWS among predominantly populations of European ancestry as of 2019, ii. the best performing weighted PRS in the UK Biobank calculated from a standard pruning & thresholding method of the CARDIoGRAMPplusC4D 1000G summary statistics involving 1.5 million SNPs[34], iii. the metaGRS, a 1.7 million-SNP PRS consisting of a weighted average of three standardized risk scores followed by LD pruning[35]; and iv. the best performing PRS in the UK Biobank derived from applying the LDPred algorithm onto the CARDIoGRAMPplusC4D 1000G summary statistics involving 6.6 million SNPs but assuming 0.1% of SNPs are causal[34]. All scores were standardized to a mean of zero and standard deviation (SD) of one within each HARE group.

Risk prediction

We estimated the increase in risk of clinical CAD associated with a 1 SD increase in PRS for each of the four PRSs within each of the three HARE groups using logistic regression adjusting for imputation release batch, age, sex and the first 10 HARE specific PCs where age was defined as the age at the time of first ICD code for cases and age at the time of last visit to the VA for controls. We also estimated Similarly, we estimated the increase in the burden of disease per 1 SD increase in PRS using linear regression where age was defined as age at time of coronary angiography.

Phenome-wide association study

We conducted a PheWAS for each of the lead SNPs at all novel loci, for the 163 SNP PRS, and for the genome-wide PRS with the highest OR for CAD in MVP. We adopted the standard PheWAS protocol[26,27] and augmented this basic approach by including phenotypes derived from the physical exam (e.g., measured weight, height, blood pressure, and heart rate), laboratory results (e.g., blood cell counts and biochemistries), and select variables derived from the MVP questionnaires (family history, smoking status, and alcohol use). For individual novel SNPs, we ran the PheWAS in each HARE group separately in both cases and controls combined and controls alone, with associations considered significant if their FDR was < 0.05 by the Benjamini-Hochberg method. For the PheWAS PRS, we restricted association analyses to Whites and ran analyses in i. all subjects; ii. after excluding CAD

cases; and iii. after further excluding subjects with other manifestations of atherosclerosis including peripheral arterial disease and ischemic stroke.

We generated a network plot with the Yifan Yu proportional multi-level layout and Atlas 2 layout algorithms implemented in Gephi Software using the subset of significant individual novel SNP PheWAS associations. The node size was defined using the weighted in-degree network statistic with the directionality from SNP to phenotype. The edge size was defined by the number of connections between two nodes (SNPs and phenotypes) and only include associations between SNP and phenotype represent by the z-score statistic of the SNP-phenotype association. The size of the label of the node was proportional to the weighted degree statistic. The color of the edges was define using the modularity matrix, a network statistic for unfolding communities in large network.

Local ancestry inference and haplotype analysis at susceptibility loci of interest

We used RFMix[25] to derive the most likely ancestral origin of the chromosomal segment encompassing loci of interest in MVP Blacks and Hispanics. The YRI, MEL and IBR populations from the 1000G project as the African reference, and the GBR, CEU and TSI populations as the European reference to infer the most likely sequence of ancestry within the locus. The results allowed us to subdivide the MVP Blacks into three groups: i. subjects with a high probability of African ancestry on both chromosomes (homozygote Africans), ii. subjects with high probability of one African and one European ancestry chromosome (heterozygotes), and iii. subjects with a high probability of European ancestry on both chromosomes. For haplotype analyses within loci of interest, we identified all common (MAF>10%) SNPs in linkage equilibrium ($r^2<0.05$) in our homozygote Africans Blacks among all SNPs reaching GWS ($P<5\times10^{-8}$) in our meta-analysis of Whites and used these SNPs to construct haplotypes and perform a haplotype trend regression of this region using the R package haplo.stats.

Downstream analyses to prioritize genes, pathways, cells, and tissues/systems relevant to CAD

We conducted downstream analyses to prioritize genes, pathways, and tissues involved in the pathogenesis of CAD based on the results of our meta-analyses. We applied four analytic algorithms to the summary statistics including Multi-marker Analysis of GenoMic Annotation (MAGMA) v1.09 for gene, gene-set, and gene-property analysis, as implemented in FUMA[30-32], a model-based enrichment method for GWAS summary data using biological pathways to define gene-sets, Regression with Summary Statistics exploiting Enrichments (RSS-E)[33], Data-driven Expression Prioritized Integration for Complex Traits (DEPICT)[28], and MetaXcan[29]. Gene and cell/tissue/system specificity/prioritization analyses incorporating gene-expression data into their algorithms were restricted to Whites given a majority of the gene-expression data incorporated into these analyses are derived from Whites. We harmonized gene level results by MAGMA, RSS-E, DEPICT, and MetaXcan, and compared to the DEPICT analyses performed on the CARDIoGRAMplusC4D and UK Biobank meta-a alone[8]. MAGMA gene-set analyses were run on 10,678 gene sets (curated gene sets: 4,761, GO terms: 5,917) from MSigDB v6.2 while gene-property analyses were conducted on GTEx V8 and multiple single cell RNA-seq databases incorporated into the FUMA bioinformatic pipeline including the Mouse Cell Atlas, the Tabula Muris dataset (FACS and droplet) and several datasets of human brain, pancreas, and blood. For RSS-E, gene-sets were derived from nine databases (BioCarta, BioCyc, HumanCyc, KEGG, miRTarBase, PANTHER, PID, Reactome, WikiPathways) that are archived by four repositories: Pathway Commons v7, NCBI Biosystems, PANTHER (v3.3), and BioCarta. We downloaded preprocessed pathway and gene data from http://doi.org/10.5281/zenodo.1473807 on October 29, 2018 and used a list of 3,803 pathways that contains between 2 to 400 autosomal protein-coding genes per pathway in the present study.

URLs

CARDIoGRAMplusC4D http://www.cardiogramplusc4d.org;
Japanese ENcyclopedia of GEnetic associations by Riken: http://jenger.riken.jp/en/result
R statistical software, www.R-project.org;
EasyQC, https://www.uni-regensburg.de/medizin/epidemiologie-praeventivmedizin/genetische-epidemiologie/software/;
PLINK: https://www.cog-genomics.org/plink/;
LDSC: https://github.com/bulik/ldsc;
Gephi: https://gephi.org/;
FUMA, http://fuma.ctglab.nl/;
PheWAS, https://github.com/PheWAS/PheWAS;
RFMixv2: https://github.com/slowkoni/rfmix;
GCTA, http://cnsgenomics.com/software/gcta/#Overview;
METAL: https://genome.sph.umich.edu/wiki/METAL;
GWAMA: https://genomics.ut.ee/en/tools/gwama;
MAGMA: https://ctg.cncr.nl/software/magma;
DEPICT: https://data.broadinstitute.org/mpg/depict/;
RSS-E: https://github.com/stephenslab/rss;
MetaXcan: https://github.com/hakyimlab/MetaXcan;

Data availability

The full summary level association data from the individual racial/ethnic association analyses in MVP as well as the trans-ethnic meta-analysis from this report will be available through dbGaP, with accession number phs001672.v4.p1 at the time of publication in a peer reviewed journal. This research has been conducted using the UK Biobank Resource under Application Numbers 13721 & 19416.

Funding/support

This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration, and was supported by Veterans Administration awards I01-01BX03362, I01-BX004821, I01-BX003340 and VA HSR RES 13-457. The content of this manuscript does not represent the views of the Department of Veterans Affairs or the United States Government.

The eMERGE Network was initiated and funded by NHGRI through the following grants: **Phase III**: U01HG8657 (Kaiser Permanente Washington/University of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine) **Phase II:** U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative/University of Washington); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center).

If the project includes data from the eMERGE imputed merged Phase I and Phase II dataset, please also add U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers. And/or The PGRNSeq dataset (eMERGE PGx), please also add U01HG004438 (CIDR) serving as a Sequencing Center. **Phase I**: U01-HG-004610 (Group Health Cooperative/University of Washington); U01-HG-004608 (Marshfield Clinic Research Foundation and Vanderbilt University Medical Center); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-04603 (Vanderbilt University Medical Center, also serving as the Administrative Coordinating Center); U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers.

The **Population Architecture Using Genomics and Epidemiology** (PAGE) program is funded by the National Human Genome Research Institute (NHGRI) with co-funding from the National Institute on Minority Health and Health Disparities (NIMHD), supported by U01HG007416 (CALiCo), U01HG007417 (ISMMS), U01HG007397 (MEC), U01HG007376 (WHI), and U01HG007419 (Coordinating Center). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The PAGE consortium thanks the staff and participants of all PAGE studies for their important contributions. The complete list of PAGE members can be found at http://www.pagestudy.org. The **MultiEthnic Study** (MEC) was supported by U01 CA164973.

The **Women's Health Initiative (**WHI) program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant S10OD028685. Funding support for the "Exonic variants and their relation to complex traits in minorities of the **(**WHI) study is provided through the NHGRI PAGE program (U01HG004790).

The **Atherosclerosis Risk in Communities** (ARIC) Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C).

Cardiovascular Health Study: This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006; and NHLBI grants U01HL080295, R01HL085251, R01HL087652, R01HL105756, R01HL103612, R01HL120393, and U01HL130114 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

BioBank Japan (BBJ) was supported by the Tailor-Made Medical Treatment Program of the Ministry of Education, Culture, Sports, Science, and Technology and Japan Agency for Medical Research (AMED) under grant numbers JP17km0305002 and JP17km0305001.

Healthy Aging in Neighborhoods of Diversity across the Life Span (HANDLS): was funded by the Interlaboratory Proposal Funding of the Intramural Research Program of the National Institute on Aging (NIA), the National Institutes of Health (NIH), Baltimore, Maryland. Funding number: [AG000989].

X. Z. was supported by the Stein Fellowship. J.A.L. and K.M.L. are supported by the U.S. Department of Veterans Affairs (IK2-CX001780). Y.L. is supported by NIH R56HL150186. S.Koyama and K.Ito were supported by AMED under Grant Numbers JP20km0405209 and JP20ek0109487. K.E.N. is supported by NIH R01HL142302. R.D. is supported by NIH R35GM124836 & R01HL139865. F.C. is supported by NCI T32CA229110. B.F.V. was supported by the NIH DK101478 and a Linda Pechenik Montague Investigator Award.

Acknowledgements

Support for title page creation and format was provided by AuthorArranger, a tool developed at the National Cancer Institute.

Consortia

Regeneron Genetics Center, Biobank Japan, CARDIoGRAMplusC4D, The VA Million Veteran Program

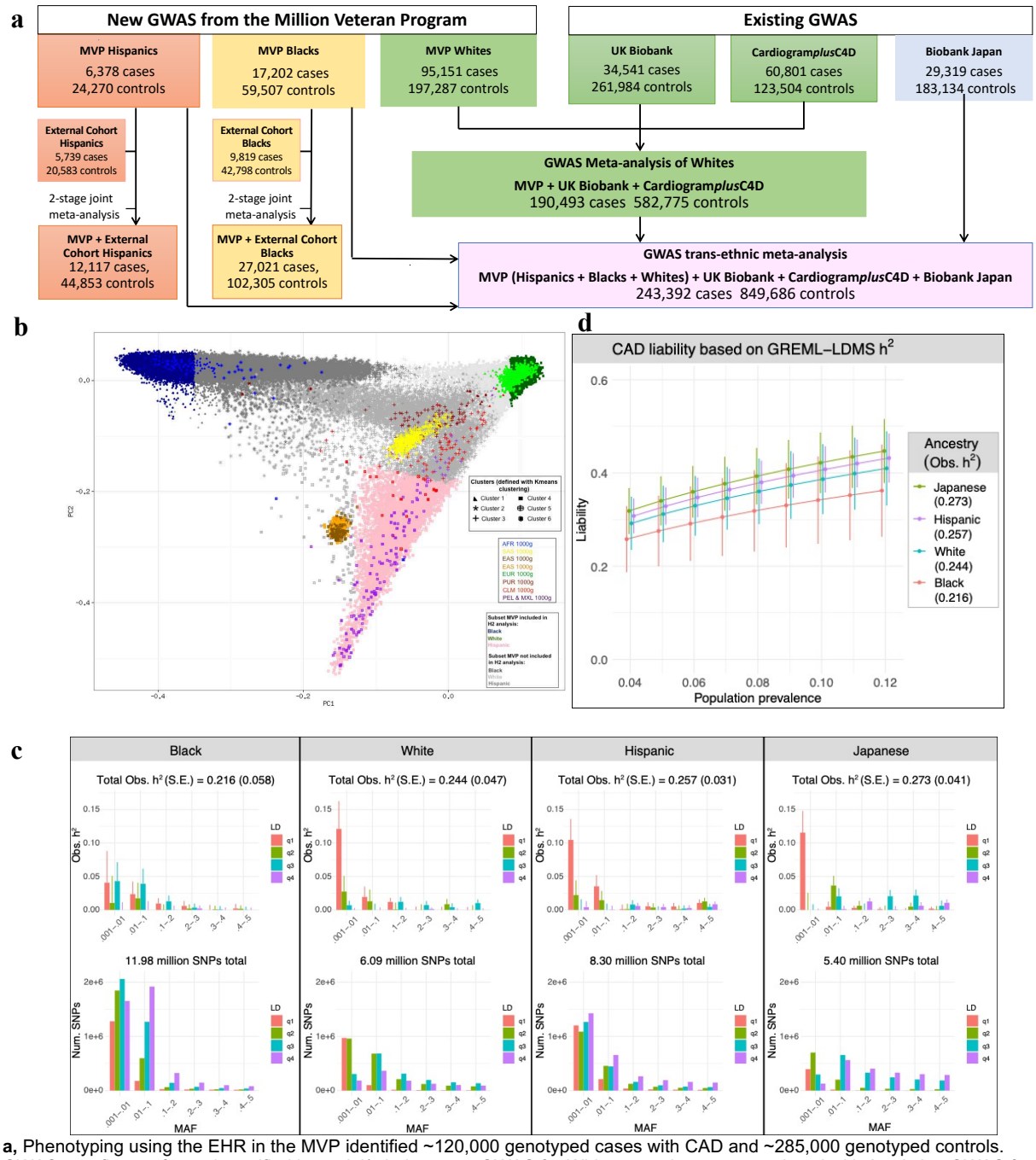
Author Contributions

Concept and design: C.T., X.Z., A.T.H., S.L.C., V.N., D.J.R., K-M. C., J.A.L., S.M.D, P.W.F.W., H.T, Y.V.S, P.S.T, C.J.O, T.L.A *Acquisition, analysis, or interpretation of data:* C.T., X.Z., A.T.H., S.L.C., V.N., S.M., B.R.G., K.L., H.F, F.C., Y.L, S.K, N.L.T, M.V., S.R., M.E.P, T.M.M., S.W.W., A.G.B, M.G.L., S.P, J.H., N.S-A., Y.H., G.L.W., S.B., C.K., J.H., R.J.F.L, R.D., M.V., K.C., K.E.N, C.L.A., M.G., C.A.H, L.L., L.R.W., J.C.B., H.L., B.S., L.A.L., A.G., O.D., I.J.K, I.B.S., G.P.J, A.S.G., S.H., B.N., J.B.H, K.M.M., K.I., K.I, Y.K., S.S.V., M.D.R., R.L.K, A.B., L.A.L., S.K., E.R.H., D.R.M., J.S.L, D.S, P.D.R., K.C., J.M.G., J.E.H., B.F.V., D.J.R., K-M.C., J.A.L., S.M.D., P.W.F.W, H.T., Y.V.S., P.S.T., C.J.D., T.L.A. *Drafting of the manuscript:* C.T., T.L.A. *Critical revision of the manuscript for important intellectual content:* X.Z., A.T.H., S.L.C, V.N., M.V, D.K., S.R., M.G.L., R.D., K.E.N., C.K., J.C.B., I.J.K, M.R., P.N., B.F.V., J.A.L., S.M.D., P.W.F.W, H.T., Y.V.S., P.S.T, C.J.O

Ethics Declarations

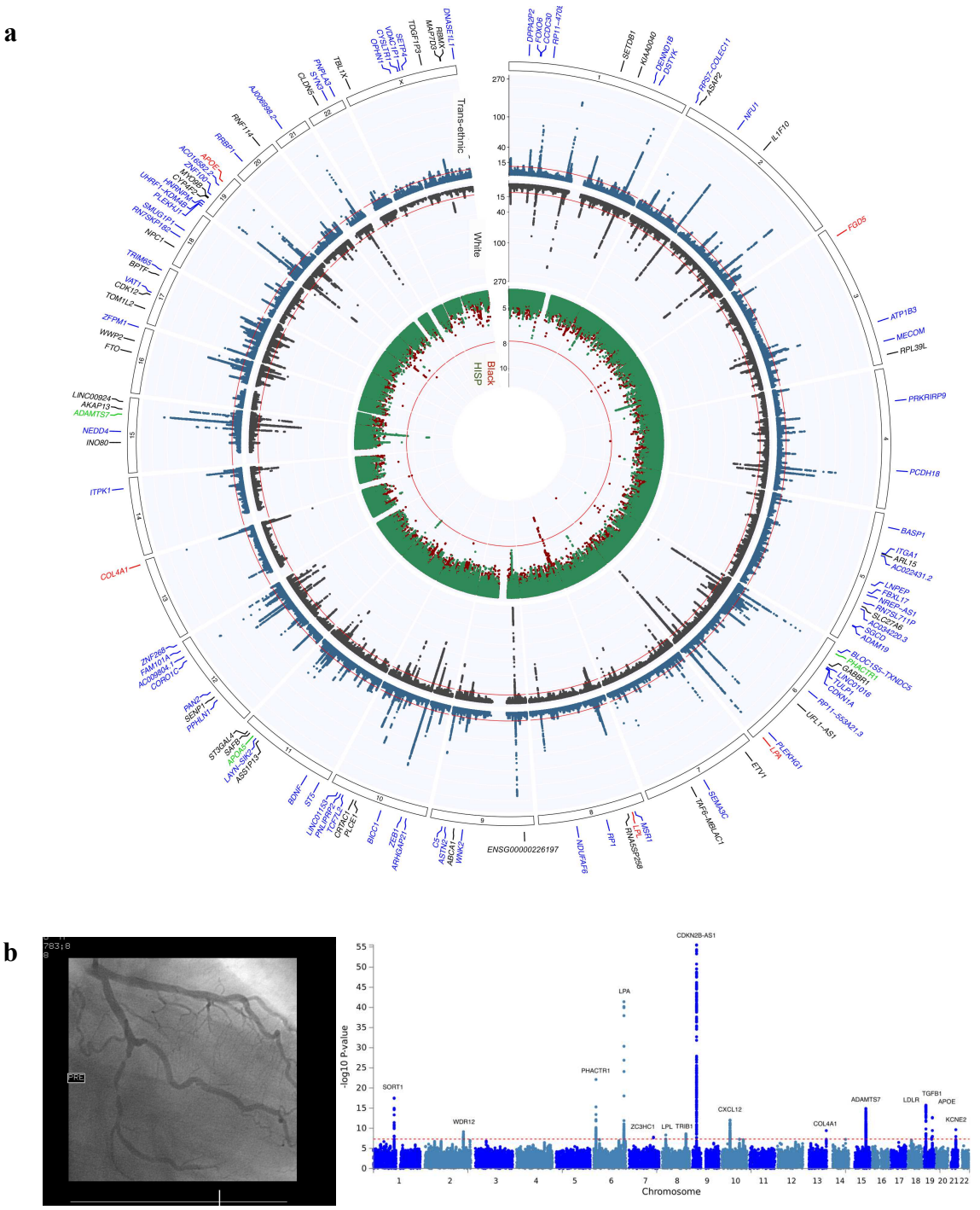
A.B. and L.A.L. are employees of Regeneron Pharmaceuticals. R.D. has received grants from AstraZeneca, grants and nonfinancial support from Goldfinch Bio, being a scientific co-founder, consultant and equity holder for Pensieve Health and being a consultant for Variant Bio. T. M. M. is an employee of the Healthcare Innovation Lab at BJC HealthCare / Washington University School of Medicine, an advisor of Myia Labs, and a compensated director the J.F Maddox Foundation in New Mexico. S.K. is an is an employee of Verve Therapeutics, holds equity in Verve Therapeutics and Maze Therapeutics, and has served as a consultant for Acceleron, Eli Lilly, Novartis, Merck, Novo Nordisk, Novo Ventures, Ionis, Alnylam, Aegerion, Haug Partners, Noble Insights, Leerink Partners, Bayer Healthcare, Illumina, Color Genomics, MedGenome, Quest, and Medscape. D.J.R. is on the Scientific Advisory Board of Alnylam, Novartis, and Verve Therapeutics. M.D.R – is on the scientific advisory board for Goldfinch Bio and Cipherome.

Fig. 1: Design of multi-ethnic genome wide association study (GWAS) of coronary artery disease (CAD) and estimates of heritability of CAD using GREML-LDMS-I for 4 racial/ethnic groups



a, Phenotyping using the EHR in the MVP identified ~120,000 genotyped cases with CAD and ~285,000 genotyped controls. GWAS was first performed stratified by racial/ethnic group. GWAS for Whites was then meta-analyzed with 2 existing GWAS for initial discovery among Whites. The GWAS for MVP Hispanics and MVP Blacks as well as the Biobank Japan GWAS of CAD was further incorporated into a single trans-ethnic meta-analysis. Two-stage joint meta-analysis of the most promising SNPs was performed for the Hispanics and Blacks with multiple external cohorts for racial/ethnic specific discovery. **b-d,** Heritability (h^2) analyses for CAD in four major racial groups using GREML-LDMS-I. **b.** Principal component analysis of MVP participants combined with 1000 genomes was first performed to identify a random subset of 19,400 Hispanics with the highest proportion of Native American ancestry (pink). A random subset of the least admixed Whites (dark green) and the least admixed Blacks (dark blue), respectively, were then matched 1:1 on case-control status, age of first EHR evidence of CAD, type of CAD presentation, and age of controls to the Hispanics. Similar matching was performed for participants from Biobank Japan study. **c,** h^2 on the observed scale stratified by linkage disequilibrium (LD) score quartile blocks minor allele frequency bins (top panel) with corresponding number of SNPs in the same block passing stringent quality control for binary trait GREML h^2 (bottom panel). **d,** h^2 on the liability scale for each racial/ethnic group as a function of the presumed population prevalence of CAD.

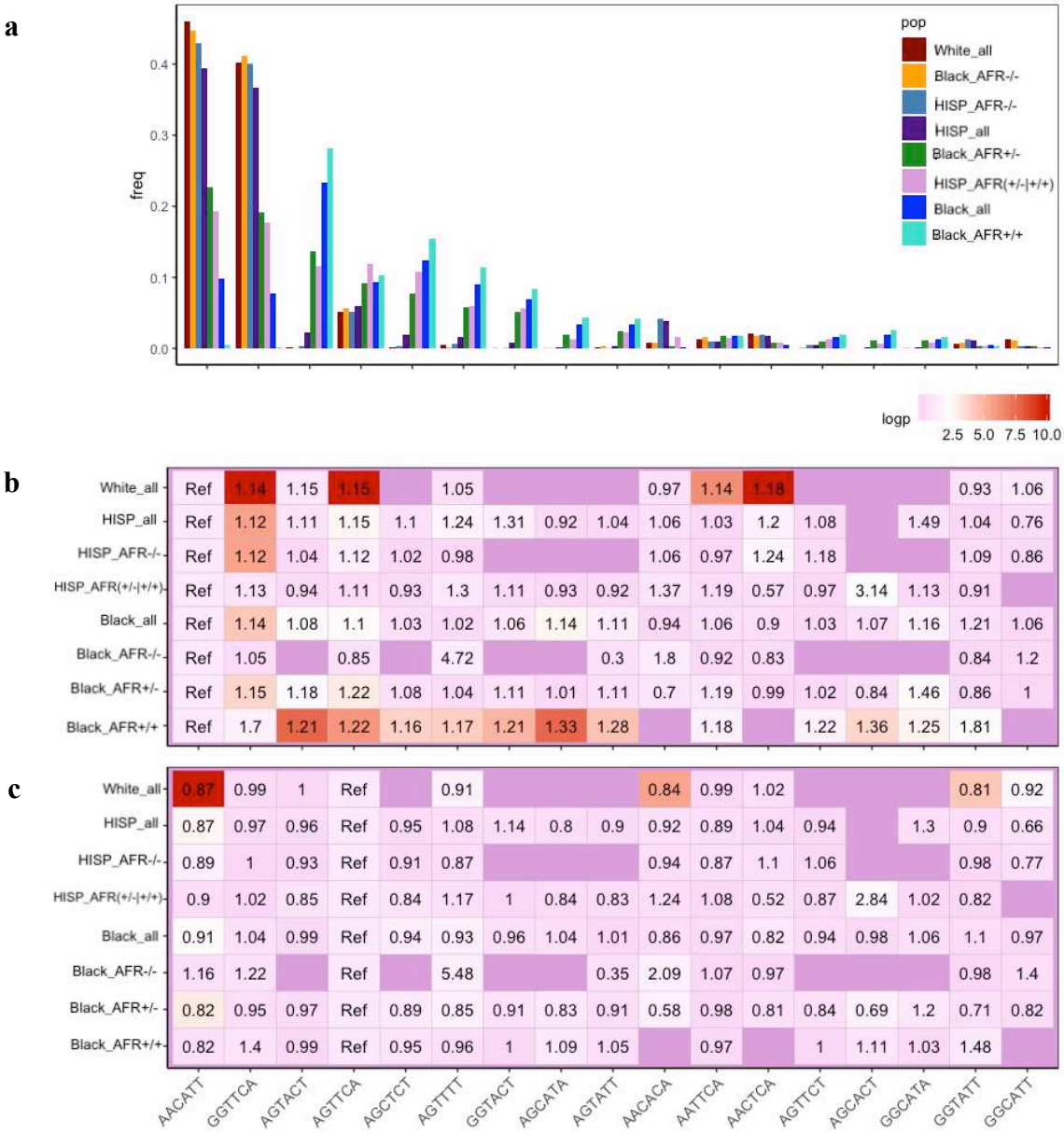
Fig. 2: Racial/ethnic specific GWAS and trans-ethnic meta-analysis identifies 107 novel loci for clinical CAD including nine on the X-chromosome, eight previously known loci among Blacks and Hispanics, and 15 previously known loci for angiographically derived burden of coronary atherosclerosis.



a, Circos plot indicating the –log$_{10}$(*P*) for association with CAD for racial/ethnic specific and trans-ethnic GWAS meta-analyses. The inner track plots the 2-stage meta-analysis association results for MVP Black/African Americans (AFR) in red and the MVP Hispanic Americans (HISP) in green while the middle track plots the results for the grand meta-analysis of White/European in Black and the trans-ethnic metanalysis further incorporating the MVP AFR, MVP HISP, and Biobank Japan in blue. The red line indicates genome-wide significance (GWS) ($P = 5.0 \times 10^{-8}$). The outer track lists the nearest mapped gene to the lead SNPs reaching GWS in each of these four meta-analyses including five loci in Blacks (red font), three loci in Hispanics (green font), 31 novel loci among Whites (black font), and 71 additional novel loci after the trans-ethnic meta-analysis (blue font). **b,** Manhattan plot (right) of trans-ethnic meta-analysis of GWAS for burden of coronary atherosclerosis as estimated by the number of coronary obstructions >50% on coronary angiogram (example left, "angiogram one" by jlt is licensed under CC BY-NC-SA 2.0) where 15 loci reach GWS. All but *LPL*, *COL4A1*, and *TGFB1* reach GWS in Whites and LPL was the only locus to reach GWS in AFR (details not shown). No locus reached GWS in Hispanics alone.

926

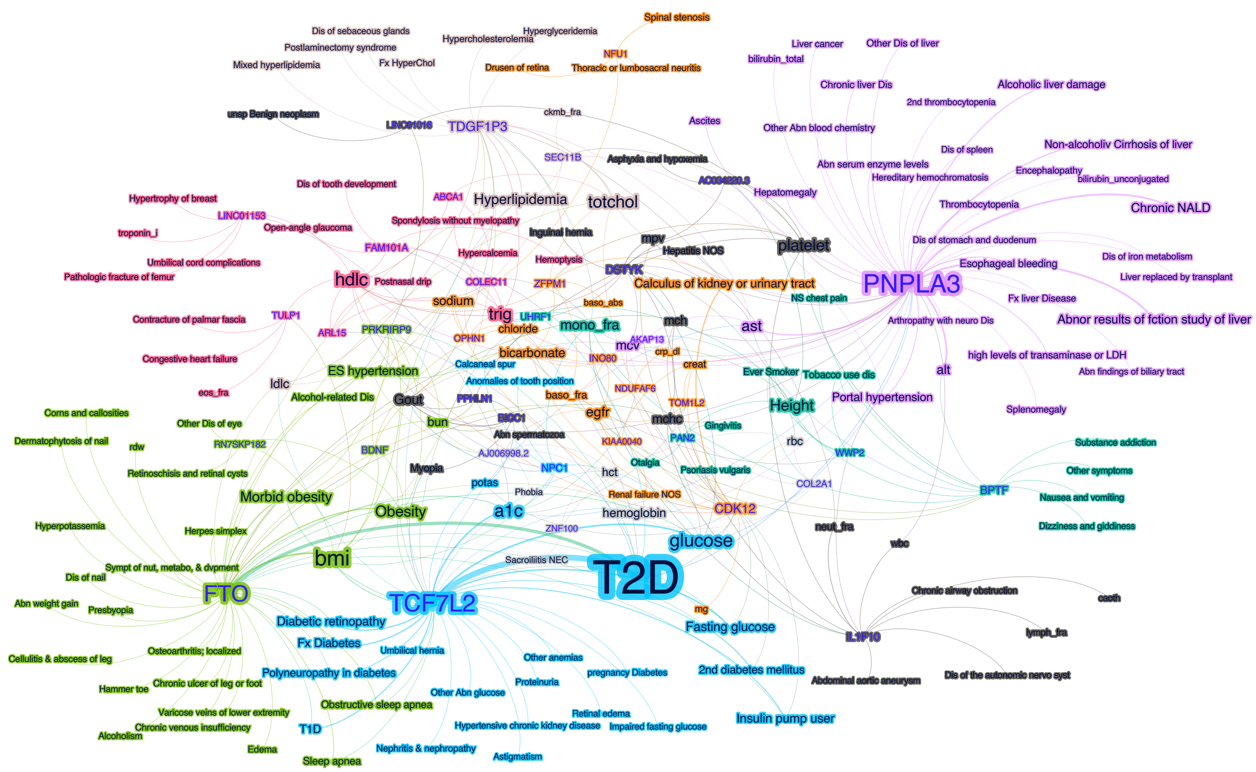
Fig. 3: Local ancestry and haplotype analyses reveals a protective haplotype at the 9p21 susceptibility locus for CAD that is virtually absent among African chromosomes



a-c, Black and Hispanic MVP participants were stratified into groups based on the degree of African ancestry at the 9p21 locus for CAD as determined by RFMix. Whites were analyzed as a single non-admixed group. The three subgroups among Blacks formed includes subjects with a high probability of having inherited two African (Black_AFR+/+) derived chromosomes in the 9p21 region, one African and one European (Black_AFR+/-), or two European chromosomes (Black_AFR-/-). The two subgroups among HA generated included those with high probability of having either 1 or 2 African chromosomes (Hisp_AFR+/-|+/+) vs. those without any African ancestry in this region but rather only Native and/or European American ancestry (Hisp_AFR-/-). Among SNPs in the high-risk region of 9p21 that reached genome wide significance among Whites, six with a minor allele frequency >10% in Black_AFR+/+ were used to infer haplotypes in the region. Each column along the x-axis represents a haplotype, named by the alleles of the six defining SNPs. **a,** only 17 out of a possible 32 haplotypes were observed to any appreciable frequency (y-axis). The first two haplotypes (AACATT and GGTTCA) dominate in Whites as well as subgroups of Blacks and Hispanics with high proportion of European ancestry at 9p21. However, these same 2 haplotypes are virtually absent among chromosomes of African descent. Most of the remaining haplotypes are present to an appreciable frequency in Black_AFR+/+ (teal) but are virtually absent in Whites. Only one haplotype (AGTTCA) has appreciable frequency in both Whites (~5%) and Black_AFR+/+ (~10%). **b-c,** summarizes the odds ratio (OR) of CAD and -log10(p) value obtained through a haplotype trend regression analyses where AACATT is the reference haplotype in **b** and AGTTCA is the reference haplotype in **c**. The most common haplotype in Whites (AACATT, 47%) is associated with a lower risk of CAD in relation to several other haplotypes but this same haplotype is unable to risk stratify among Blacks given it is virtually absent among Black_AFR+/+. Any signal among Blacks is dependent on the presence of this haplotype through local admixture with Whites, although analyses among the small subgroup Black_AFR-/- do not generate a reliable signal likely because of inadequate power.

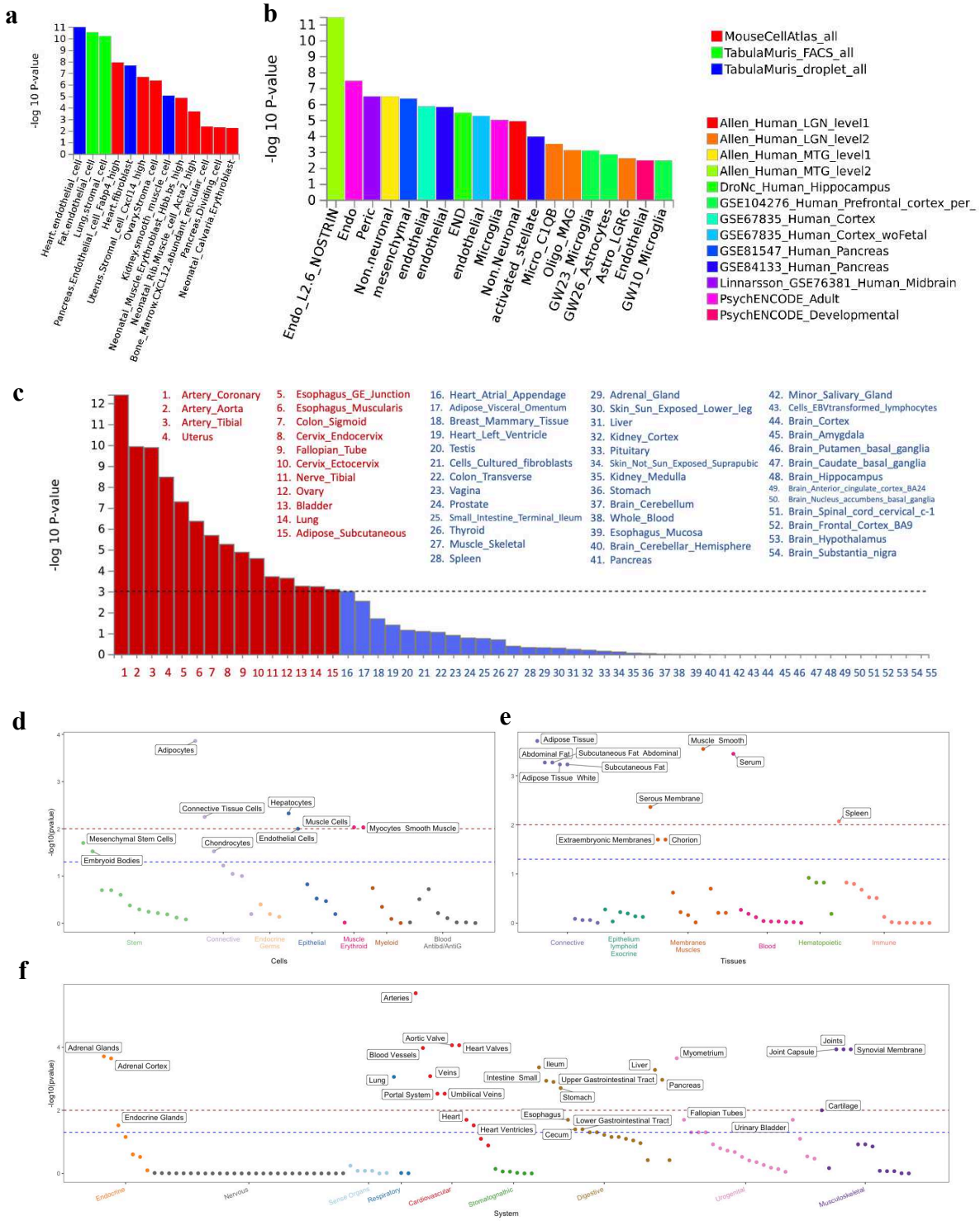
927

Fig. 4: Pleiotropic assessment of 107 novel loci through extended phenome wide association of lead SNPs among White/European controls in the Million Veteran Program highlights and strengthens links between CAD and obesity-insulin resistance-diabetes axis of risk



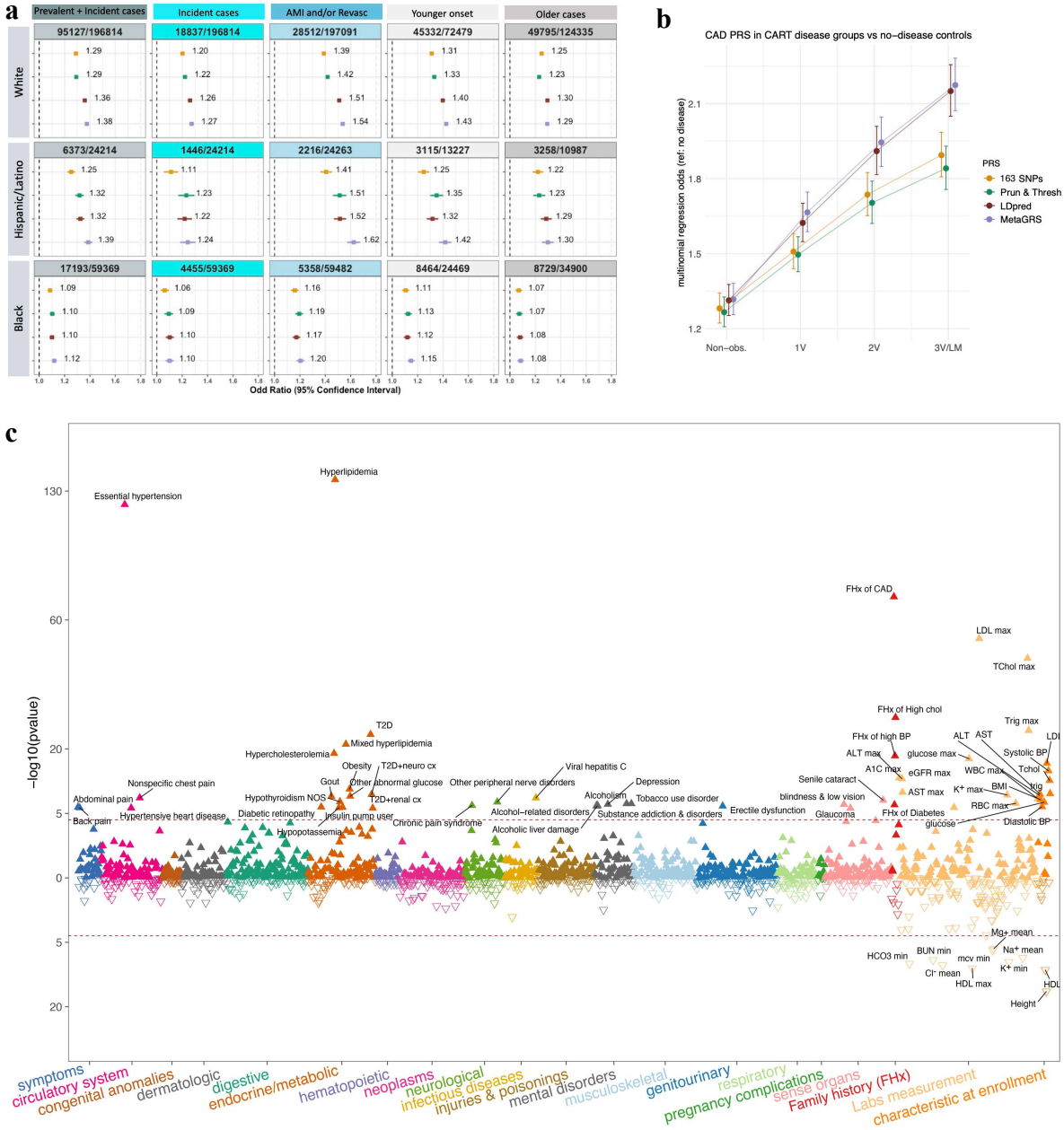
Network plot of genotype-phenotype associations reaching significance at FDR<0.05 among White/Europeans MVP control participants for the lead SNPs in the 107 novel loci. Nodes are labelled either with the mapped gene for a lead SNP (purple font) or a phenotype tested in the PheWAS (black font). To highlight most pleiotropic SNPs and facilitate interpretation, the plot is restricted to lead SNPs associated with at least three distinct phenotypes. Distinct colors of nodes and edges represent a group of genotypes and phenotypes in the same dominant network. The thickness of the edges is correlated with the strength of the SNP-phenotype association (z-score). The size of the labels is dictated by the number of connections to phenotypes or genes and the strength of association. Network plot was created using Yifan Yu proportional and Atlas 2 layout algorithms as implemented in Gephi software.

Fig. 5: Downstream analyses to prioritize systems, pathways, tissues, and cells relevant to CAD



a-c, MAGMA gene-property analyses to test relationship between expressed genes in specific cells or tissues and genetic associations (meta-analysis of Whites) as implemented in FUMA. Data in **a** is restricted to three mouse single-cell RNA-seq (sc-RNA) datasets involving a broad range of cell types/organs while data in **b** is restricted to human datasets mostly involving the brain but also the pancreas and blood. Results show only independent cell-type associations based on within-dataset conditional analyses ordered by p value across datasets. Data in **c** shows results for 54 specific tissue from the GTEx RNA-seq dataset v8 in order of p-value significance with red bars and font highlighting statistically significant tissues after adjusting for multiple testing (horizontal black dashed line) while remaining tissues are in blue. **d-f**, DEPICT following standard algorithm on the same GWAS used for MAGMA analyses in **a-c**. DEPICT results are separated into **d,** cells **e,** tissues, and **f,** systems. -log10(pvalue) for a false discovery rate (FDR) of <0.05 is demarcated by red dashed line while the FDR <0.2 threshold is shown in blue. Only cells/tissues reaching an FDR<0.2 are labelled. Endothelial, stromal/fibroblast, smooth muscle cells as well as adipocytes and hepatocytes are prioritized as well as multiple tissues rich in these cell types or their derivatives. Please see text for more details on methods, summary and interpretation of results.

Fig. 6: Testing externally validated polygenic risk scores (PRS) for association with clinical CAD, burden of coronary atherosclerosis, and other phenotypes in the Million Veteran Program



a, b Four progressively complex weighted polygenic risk scores (PRS) of CAD are constructed, standardized to mean 0 and unit-variance, and tested for association with clinical CAD and burden of atherosclerosis in MVP using logistic and multinomial regression, respectively and reporting the odds ratio of risk associated with 1 standard deviation increase in PRS. The simplest score, '163 SNPs', is restricted to lead SNPs of genome wide significant as of 2019 from CARDIoGRAMplusC4D and the UK Biobank. The remaining genome-wide PRSs were derived in external datasets using either a standard pruning and thresholding strategy, 'Prun & Thresh', modeling linkage disequilibrium, 'LDPred', or through the meta-analysis of the weights of 3 separate scores, 'metaGRS'. **a,** PRS were tested in MVP Whites, Blacks, and Hispanics, separately. In addition to all cases combined, subgroups of incident only cases (after enrollment), severe cases with evidence of either a myocardial infarction and/or a revascularization procedure, and early onset vs older onset cases (divided by median age of onset) were tested. **b,** PRS are tested for burden of coronary atherosclerosis only among Whites. The reference group is subjects with normal coronaries on angiography. For progressively higher burdens of disease are tested against the reference group including non-obstructive disease ('Non-obs.'), 1-vessel disease (1V), 2-vessel disease (2V), and 3-vessel or left main disease (3V/LM). **c,** The best performing score in **a** and **b**, the meta-GRS, is tested for association with Phecodes, clinical labs and anthropomorphic measures, as well as selected components of the baseline questionnaires among whites with no EHR evidence of atherosclerosis related complications at the end of EHR follow up.

References

1. Roth, G.A. *et al*. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J Am Coll Cardiol* **76**, 2982-3021 (2020).
2. Statistics;, N.C.f.H. Health, United States Spotlight: Racial and Ethnic Disparities in Heart Disease. (Centers for Disease Control and Prevention, 2019).
3. Churchwell, K. *et al*. Call to Action: Structural Racism as a Fundamental Driver of Health Disparities: A Presidential Advisory From the American Heart Association. *Circulation* **142**(2020).
4. Popejoy, A.B. & Fullerton, S.M. Genomics is failing on diversity. *Nature* **538**, 161-164 (2016).
5. Martin, A.R. *et al*. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584-591 (2019).
6. Zdravkovic, S. *et al*. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *J Intern Med* **252**, 247-54 (2002).
7. Wienke, A., Holm, N.V., Skytthe, A. & Yashin, A.I. The heritability of mortality due to heart diseases: a correlated frailty model applied to Danish twins. *Twin Res* **4**, 266-74 (2001).
8. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ Res* **122**, 433-443 (2018).
9. Koyama, S. *et al*. Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat Genet* **52**, 1169-1177 (2020).
10. 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).
11. Peden, J.F. *et al*. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat Genet* (2011).
12. Lu, X. *et al*. Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. *Nat Genet* **44**, 890-894 (2012).
13. Assimes, T.L. & Roberts, R. Genetics: Implications for Prevention and Management of Coronary Artery Disease. *J Am Coll Cardiol* **68**, 2797-2818 (2016).
14. Buniello, A. *et al*. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
15. Gaziano, J.M. *et al*. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* (2015).
16. Byrd, J.B. *et al*. Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: the VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program. *Am Heart J* **165**, 434-40 (2013).
17. Maddox, T.M. *et al*. A national clinical quality program for Veterans Affairs catheterization laboratories (from the Veterans Affairs clinical assessment, reporting, and tracking program). *Am J Cardiol* **114**, 1750-7 (2014).
18. Yang, J. *et al*. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**, 1114-20 (2015).
19. (U.S.), N.C.f.H.S. Crude percentages of coronary heart disease for adults aged 18 and over, United States, 2015-2018. National Health Interview Survey. in *National Center for Health Statistics, National Health Interview Survey, 2015-2018* (2020).
20. Evaluation, I.f.H.M.a. GBD Compare Data Visualization. (University of Washington, Seattle, WA, 2018).
21. Bulik-Sullivan, B.K. *et al*. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
22. Nikpay, M. *et al*. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121-1130 (2015).
23. Ishigaki, K. *et al*. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat Genet* **52**, 669-679 (2020).
24. Coram, M.A. *et al*. Leveraging Multi-ethnic Evidence for Mapping Complex Traits in Minority Populations: An Empirical Bayes Approach. *Am J Hum Genet* (2015).
25. Maples, B.K., Gravel, S., Kenny, E.E. & Bustamante, C.D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* **93**, 278-88 (2013).
26. Denny, J.C. *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205-10 (2010).
27. Wu, P. *et al*. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325 (2019).
28. Pers, T.H. *et al*. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6**, 5890 (2015).
29. Barbeira, A.N. *et al*. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* **9**, 1825 (2018).

30. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
31. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015).
32. de Leeuw, C.A., Stringer, S., Dekkers, I.A., Heskes, T. & Posthuma, D. Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure. *Nat Commun* **9**, 3768 (2018).
33. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat Commun* **9**, 4361 (2018).
34. Khera, A.V. *et al*. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224 (2018).
35. Inouye, M. *et al*. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol* **72**, 1883-1893 (2018).
36. Dikilitas, O. *et al*. Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups. *Am J Hum Genet* **106**, 707-716 (2020).
37. Wainschtein, P. *et al*. Recovery of trait heritability from whole genome sequence data. *bioRxiv*, 588020 (2019).
38. McPherson, R. *et al*. A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488-91 (2007).
39. Schunkert, H. *et al*. Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. *Circulation* **117**, 1675-84 (2008).
40. Consortium, C.A.D. *et al*. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* **45**, 25-33 (2013).
41. Shungin, D. *et al*. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-196 (2015).
42. Huang, Y. *et al*. Sexual Differences in Genetic Predisposition of Coronary Artery Disease. *Circ Genom Precis Med* (2020).
43. Zore, T., Palafox, M. & Reue, K. Sex differences in obesity, lipid metabolism, and inflammation-A role for the sex chromosomes? *Mol Metab* **15**, 35-44 (2018).
44. Salfati, E. *et al*. Susceptibility Loci for Clinical Coronary Artery Disease and Subclinical Coronary Atherosclerosis Throughout the Life-Course. *Circ Cardiovasc Genet* **8**, 803-11 (2015).
45. Speliotes, E.K. *et al*. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**, 937-948 (2010).
46. Vujkovic, M. *et al*. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet* **52**, 680-691 (2020).
47. Speliotes, E.K. *et al*. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet* **7**, e1001324 (2011).
48. Klimentidis, Y.C. *et al*. Phenotypic and Genetic Characterization of Lower LDL Cholesterol and Increased Type 2 Diabetes Risk in the UK Biobank. *Diabetes* **69**, 2194-2205 (2020).
49. Fletcher, R. *et al*. The role of the Niemann-Pick disease, type C1 protein in adipocyte insulin action. *PLoS One* **9**, e95598 (2014).
50. Wirka, R.C. *et al*. Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. *Nat Med* **25**, 1280-1289 (2019).
51. Nelson, C.P. *et al*. Genetically determined height and coronary artery disease. *N Engl J Med* **372**, 1608-18 (2015).
52. Ong, J.S. *et al*. Height and overall cancer risk and mortality: evidence from a Mendelian randomisation study on 310,000 UK Biobank participants. *Br J Cancer* **118**, 1262-1267 (2018).
53. Tabas, I., Garcia-Cardena, G. & Owens, G.K. Recent insights into the cellular biology of atherosclerosis. *J Cell Biol* **209**, 13-22 (2015).
54. Nagao, M. *et al*. Coronary Disease-Associated Gene TCF21 Inhibits Smooth Muscle Cell Differentiation by Blocking the Myocardin-Serum Response Factor Pathway. *Circ Res* **126**, 517-529 (2020).
55. Lowrie Jr., D.J. Histology : an essential textbook. (Thieme Publishers, New York, 2020).
56. Ko, C.W., Qu, J., Black, D.D. & Tso, P. Regulation of intestinal lipid metabolism: current concepts and relevance to disease. *Nat Rev Gastroenterol Hepatol* **17**, 169-183 (2020).
57. Fahed, A.C. *et al*. Transethnic Transferability of a Genome-wide Polygenic Score for Coronary Artery Disease. *Circ Genom Precis Med* (2020).
58. Hunter-Zinck, H. *et al*. Genotyping Array Design and Data Quality Control in the Million Veteran Program. *Am J Hum Genet* **106**, 535-548 (2020).
59. Genomes Project, C. *et al*. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
60. Loh, P.R. *et al*. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448 (2016).
61. Das, S. *et al*. Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).

62. Fang, H. *et al*. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am J Hum Genet* **105**, 763-772 (2019).
63. Maddox, T.M. *et al*. Nonobstructive coronary artery disease and risk of myocardial infarction. *JAMA* **312**, 1754-63 (2014).
64. Evans, L.M. *et al*. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat Genet* **50**, 737-745 (2018).
65. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011).
66. Lee, S.H. *et al*. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* **44**, 247-50 (2012).
67. Visscher, P.M. *et al*. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet* **10**, e1004269 (2014).
68. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
69. Magi, R. & Morris, A.P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
70. Loley, C. *et al*. No Association of Coronary Artery Disease with X-Chromosomal Variants in Comprehensive International Meta-Analysis. *Sci Rep* **6**, 35278 (2016).

Figures

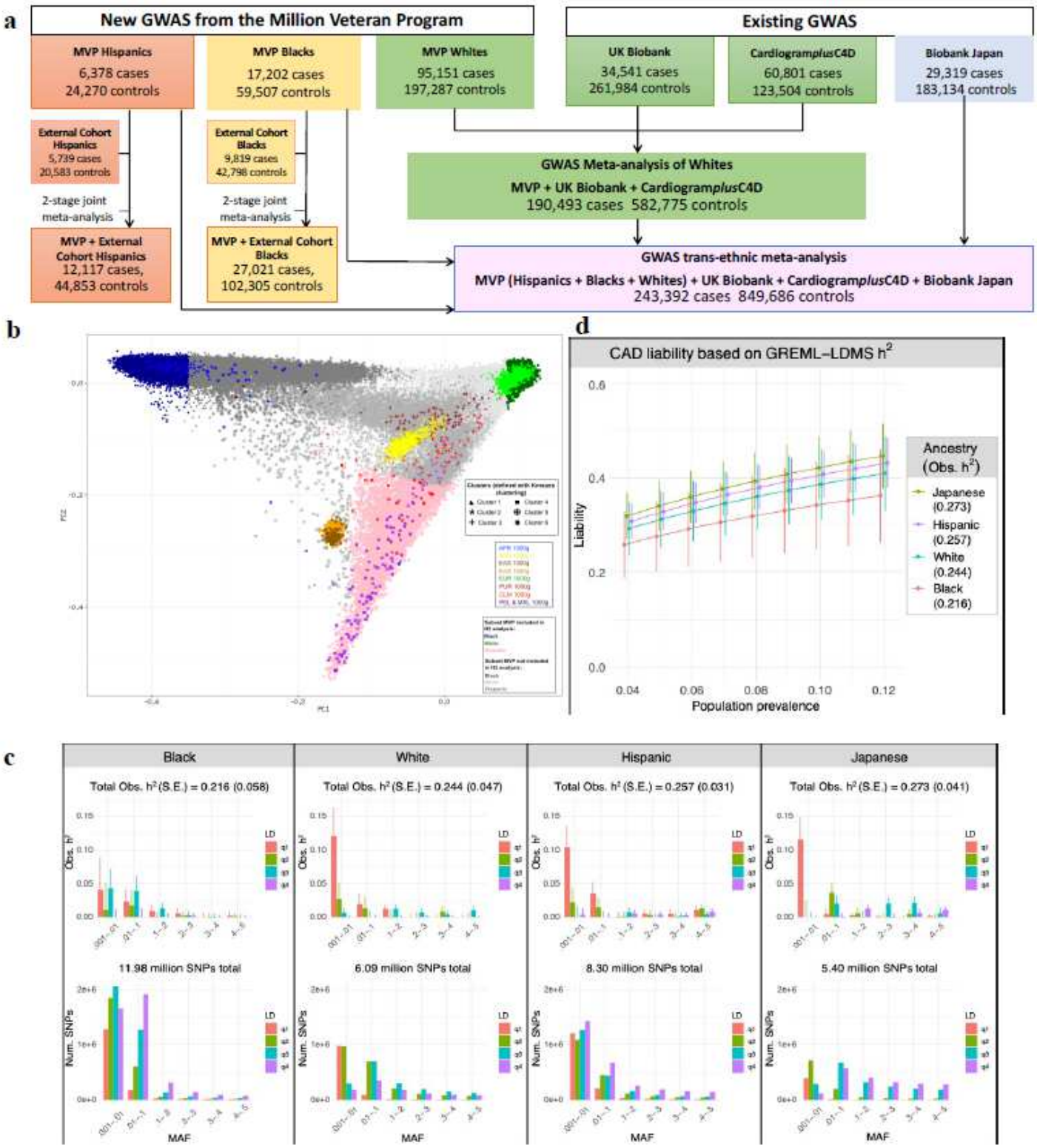


Figure 1

Design of multi-ethnic genome wide association study (GWAS) of coronary artery disease (CAD) and estimates of heritability of CAD using GREML-LDMS-I for 4 racial/ethnic groups. a, Phenotyping using the EHR in the MVP identified ~120,000 genotyped cases with CAD and ~285,000 genotyped controls. GWAS

was first performed stratified by racial/ethnic group. GWAS for Whites was then meta-analyzed with 2 existing GWAS for initial discovery among Whites. The GWAS for MVP Hispanics and MVP Blacks as well as the Biobank Japan GWAS of CAD was further incorporated into a single trans-ethnic meta-analysis. Two-stage joint meta-analysis of the most promising SNPs was performed for the Hispanics and Blacks with multiple external cohorts for racial/ethnic specific discovery. b-d, Heritability (h2) analyses for CAD in four major racial groups using GREML-LDMS-I. b. Principal component analysis of MVP participants combined with 1000 genomes was first performed to identify a random subset of 19,400 Hispanics with the highest proportion of Native American ancestry (pink). A random subset of the least admixed Whites (dark green) and the least admixed Blacks (dark blue), respectively, were then matched 1:1 on case-control status, age of first EHR evidence of CAD, type of CAD presentation, and age of controls to the Hispanics. Similar matching was performed for participants from Biobank Japan study. c, h2 on the observed scale stratified by linkage disequilibrium (LD) score quartile blocks minor allele frequency bins (top panel) with corresponding number of SNPs in the same block passing stringent quality control for binary trait GREML h2 (bottom panel). d, h2 on the liability scale for each racial/ethnic group as a function of the presumed population prevalence of CAD.

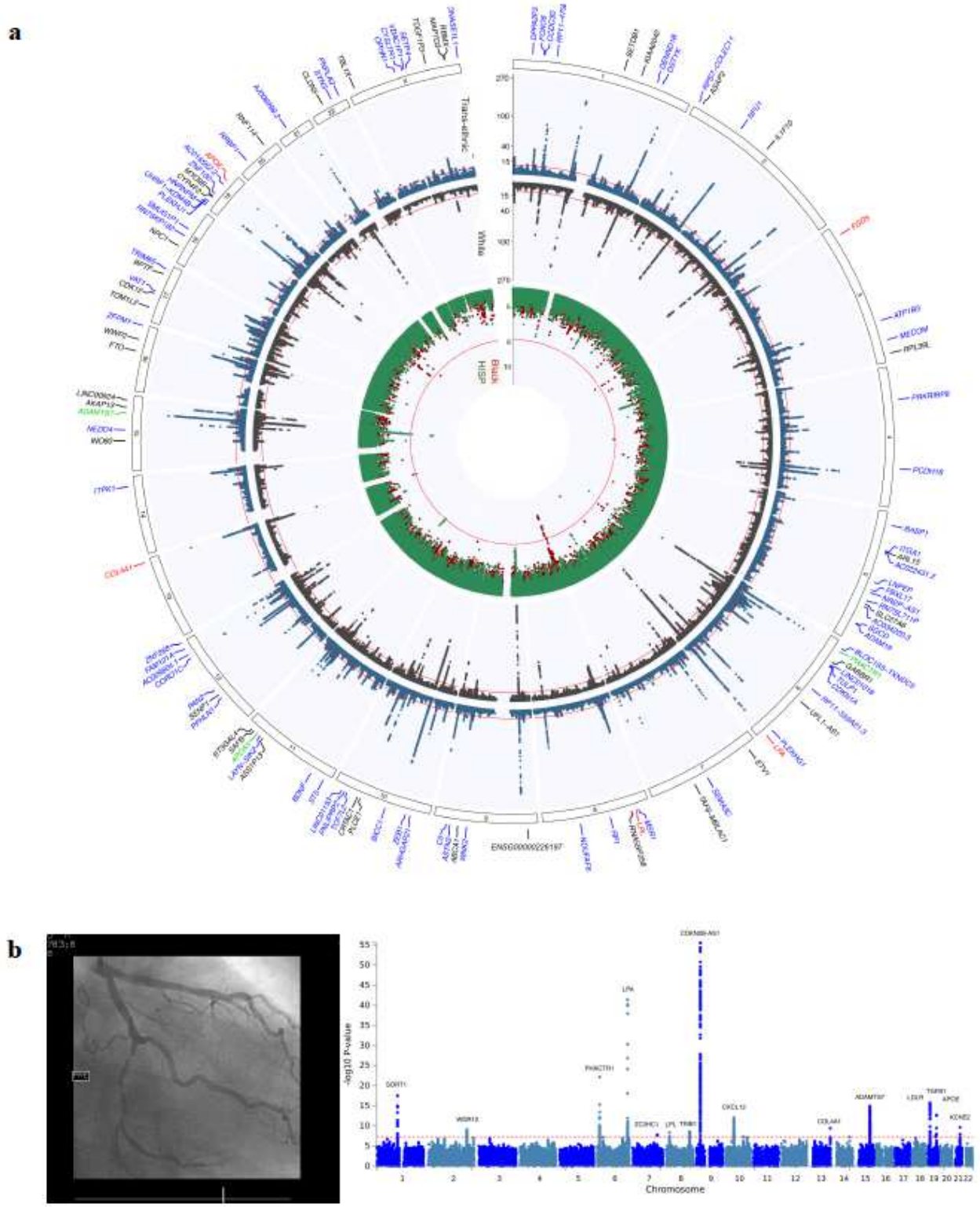


Figure 2

Racial/ethnic specific GWAS and trans-ethnic meta-analysis identifies 107 novel loci for clinical CAD including nine on the X-chromosome, eight previously known loci among Blacks and Hispanics, and 15 previously known loci for angiographically derived burden of coronary atherosclerosis. a, Circos plot indicating the –log10(P) for association with CAD for racial/ethnic specific and trans-ethnic GWAS meta-analyses. The inner track plots the 2-stage meta-analysis association results for MVP Black/African

Americans (AFR) in red and the MVP Hispanic Americans (HISP) in green while the middle track plots the results for the grand meta-analysis of White/European in Black and the trans-ethnic metanalysis further incorporating the MVP AFR, MVP HISP, and Biobank Japan in blue. The red line indicates genome-wide significance (GWS) (P = 5.0 × 10−8). The outer track lists the nearest mapped gene to the lead SNPs reaching GWS in each of these four meta-analyses including five loci in Blacks (red font), three loci in Hispanics (green font), 31 novel loci among Whites (black font), and 71 additional novel loci after the trans-ethnic meta-analysis (blue font). b, Manhattan plot (right) of trans-ethnic meta-analysis of GWAS for burden of coronary atherosclerosis as estimated by the number of coronary obstructions >50% on coronary angiogram (example left, “angiogram one” by j l t is licensed under CC BY-NC-SA 2.0) where 15 loci reach GWS. All but LPL, COL4A1, and TGFB1 reach GWS in Whites and LPL was the only locus to reach GWS in AFR (details not shown). No locus reached GWS in Hispanics alone.

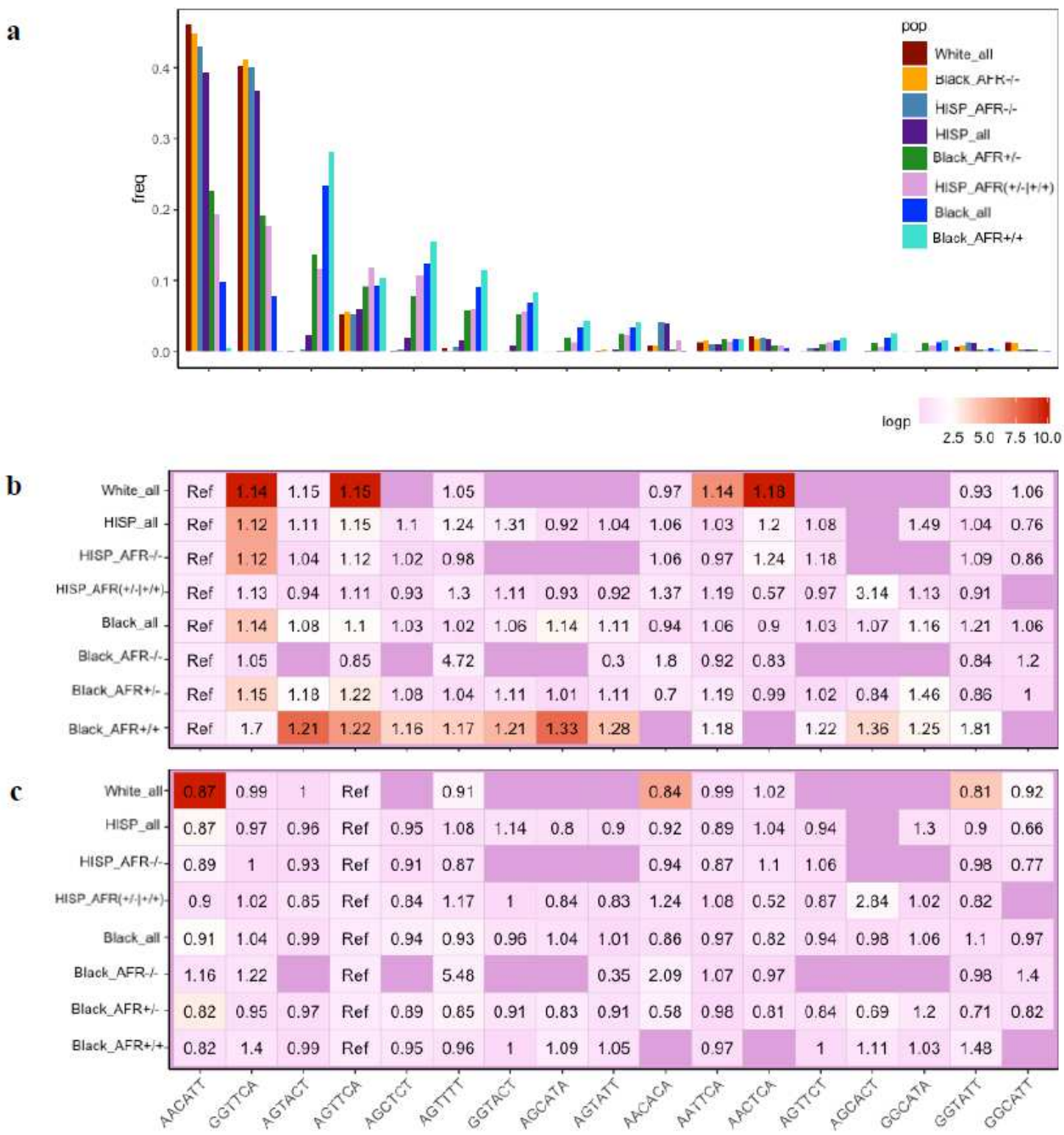


Figure 3

Local ancestry and haplotype analyses reveals a protective haplotype at the 9p21 susceptibility locus for CAD that is virtually absent among African chromosomes. a-c, Black and Hispanic MVP participants were stratified into groups based on the degree of African ancestry at the 9p21 locus for CAD as determined by RFMix. Whites were analyzed as a single non-admixed group. The three subgroups among Blacks formed includes subjects with a high probability of having inherited two African (Black_AFR+/+) derived

chromosomes in the 9p21 region, one African and one European (Black_AFR+/-), or two European chromosomes (Black_AFR-/-). The two subgroups among HA generated included those with high probability of having either 1 or 2 African chromosomes (Hisp_AFR+/-|+/+) vs. those without any African ancestry in this region but rather only Native and/or European American ancestry (Hisp_AFR-/-). Among SNPs in the high-risk region of 9p21 that reached genome wide significance among Whites, six with a minor allele frequency >10% in Black_AFR+/+ were used to infer haplotypes in the region. Each column along the x-axis represents a haplotype, named by the alleles of the six defining SNPs. a, only 17 out of a possible 32 haplotypes were observed to any appreciable frequency (y-axis). The first two haplotypes (AACATT and GGTTCA) dominate in Whites as well as subgroups of Blacks and Hispanics with high proportion of European ancestry at 9p21. However, these same 2 haplotypes are virtually absent among chromosomes of African descent. Most of the remaining haplotypes are present to an appreciable frequency in Black_AFR+/+ (teal) but are virtually absent in Whites. Only one haplotype (AGTTCA) has appreciable frequency in both Whites (~5%) and Black_AFR+/+ (~10%). b-c, summarizes the odds ratio (OR) of CAD and -log10(p) value obtained through a haplotype trend regression analyses where AACATT is the reference haplotype in b and AGTTCA is the reference haplotype in c. The most common haplotype in Whites (AACATT, 47%) is associated with a lower risk of CAD in relation to several other haplotypes but this same haplotype is unable to risk stratify among Blacks given it is virtually absent among Black_AFR+/+. Any signal among Blacks is dependent on the presence of this haplotype through local admixture with Whites, although analyses among the small subgroup Black_AFR-/- do not generate a reliable signal likely because of inadequate power.

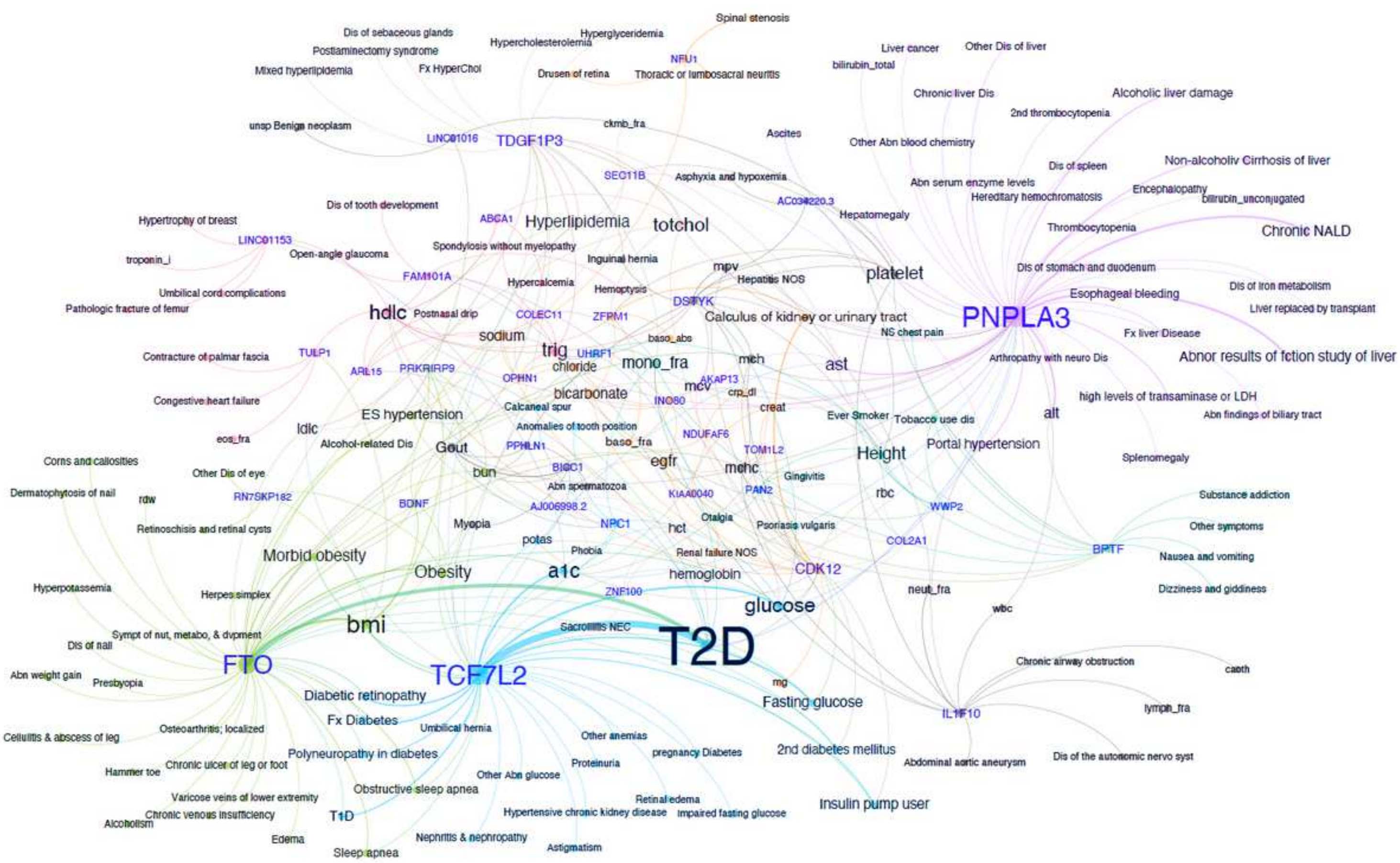


Figure 4

Pleiotropic assessment of 107 novel loci through extended phenome wide association of lead SNPs among White/European controls in the Million Veteran Program highlights and strengthens links between CAD and obesity-insulin resistance-diabetes axis of risk. Network plot of genotype-phenotype associations reaching significance at FDR<0.05 among White/Europeans MVP control participants for the lead SNPs in the 107 novel loci. Nodes are labelled either with the mapped gene for a lead SNP (purple font) or a phenotype tested in the PheWAS (black font). To highlight most pleiotropic SNPs and facilitate interpretation, the plot is restricted to lead SNPs associated with at least three distinct phenotypes. Distinct colors of nodes and edges represent a group of genotypes and phenotypes in the same dominant network. The thickness of the edges is correlated with the strength of the SNP-phenotype association (z-score). The size of the labels is dictated by the number of connections to phenotypes or genes and the strength of association. Network plot was created using Yifan Yu proportional and Atlas 2 layout algorithms as implemented in Gephi software.

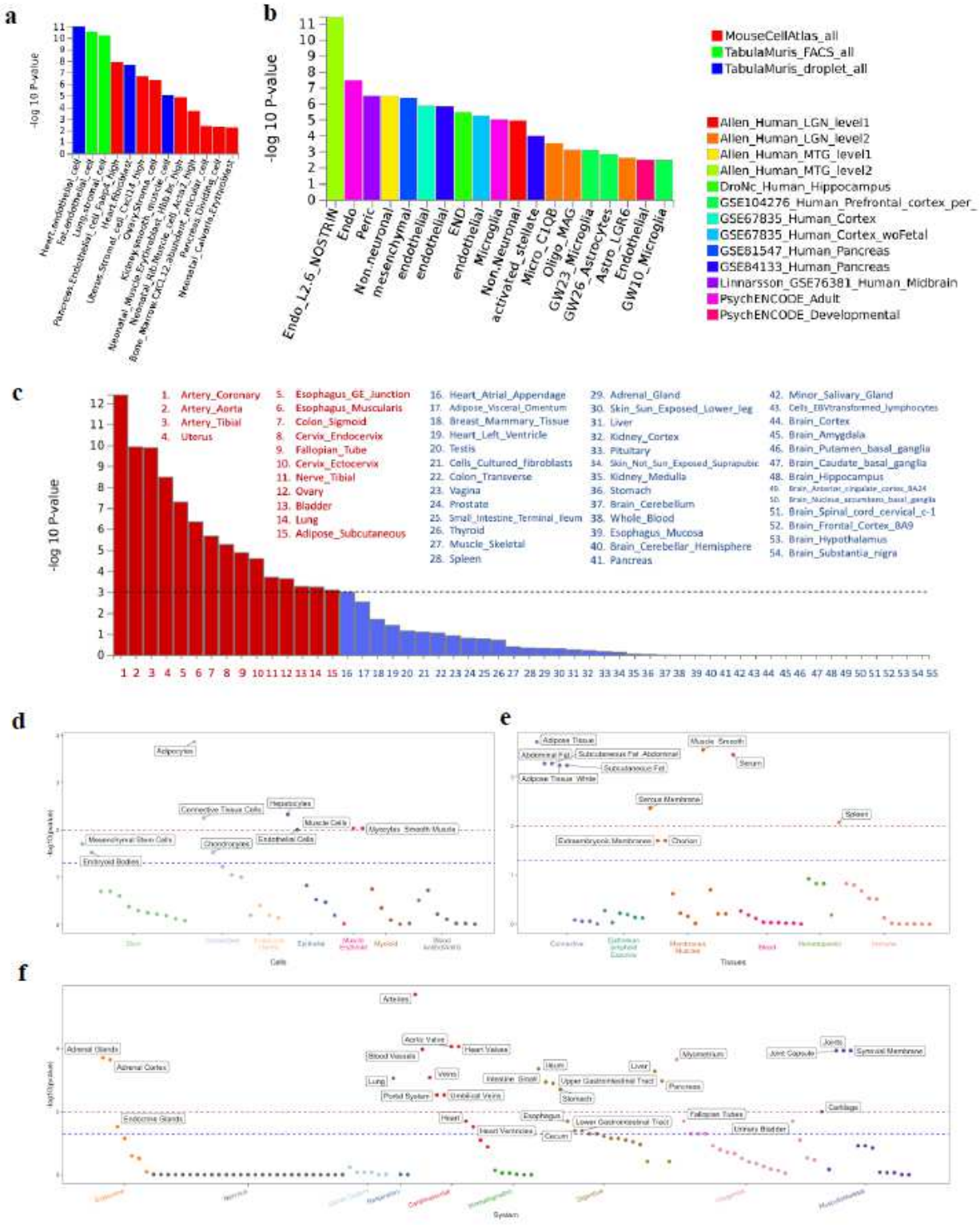


Figure 5

Downstream analyses to prioritize systems, pathways, tissues, and cells relevant to CAD. a-c, MAGMA gene-property analyses to test relationship between expressed genes in specific cells or tissues and genetic associations (meta-analysis of Whites) as implemented in FUMA. Data in a is restricted to three mouse single-cell RNA-seq (sc- RNA) datasets involving a broad range of cell types/organs while data in b is restricted to human datasets mostly involving the brain but also the pancreas and blood. Results

show only independent cell-type associations based on within-dataset conditional analyses ordered by p value across datasets. Data in c shows results for 54 specific tissue from the GTEx RNA-seq dataset v8 in order of p-value significance with red bars and font highlighting statistically significant tissues after adjusting for multiple testing (horizontal black dashed line) while remaining tissues are in blue. d-f, DEPICT following standard algorithm on the same GWAS used for MAGMA analyses in a-c. DEPICT results are separated into d, cells e, tissues, and f, systems. -log10(pvalue) for a false discovery rate (FDR) of <0.05 is demarcated by red dashed line while the FDR <0.2 threshold is shown in blue. Only cells/tissues reaching an FDR<0.2 are labelled. Endothelial, stromal/fibroblast, smooth muscle cells as well as adipocytes and hepatocytes are prioritized as well as multiple tissues rich in these cell types or their derivatives. Please see text for more details on methods, summary and interpretation of results.

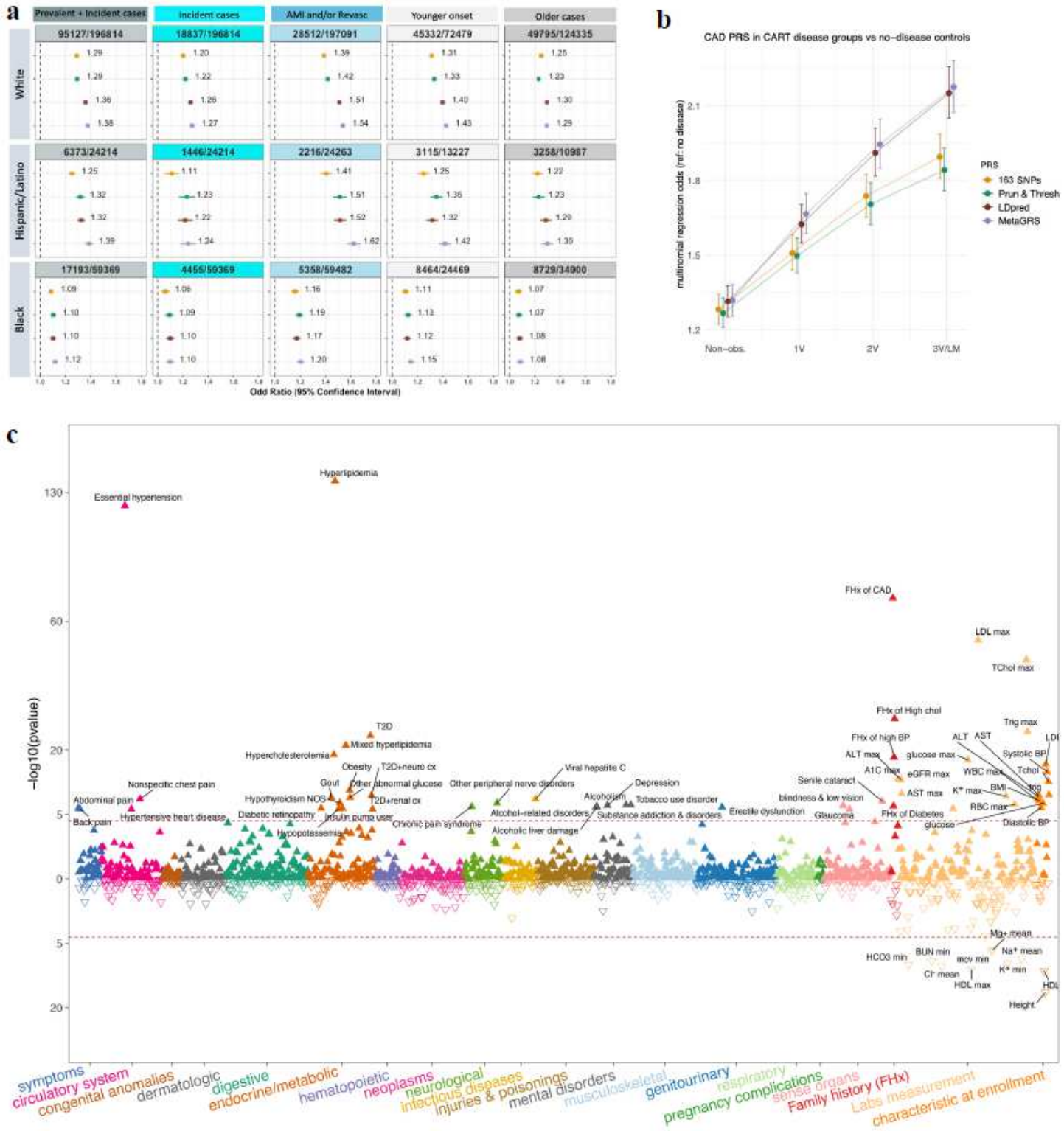


Figure 6

Testing externally validated polygenic risk scores (PRS) for association with clinical CAD, burden of coronary atherosclerosis, and other phenotypes in the Million Veteran Program. a, b Four progressively complex weighted polygenic risk scores (PRS) of CAD are constructed, standardized to mean 0 and unitvariance, and tested for association with clinical CAD and burden of atherosclerosis in MVP using logistic and multinomial regression, respectively and reporting the odds ratio of risk associated with 1

standard deviation increase in PRS. The simplest score, ‘163 SNPs’, is restricted to lead SNPs of genome wide significant as of 2019 from CARDIoGRAMplusC4D and the UK Biobank. The remaining genome-wide PRSs were derived in external datasets using either a standard pruning and thresholding strategy, ‘Prun & Thresh’, modeling linkage disequilibrium, ‘LDPred’, or through the meta-analysis of the weights of 3 separate scores, ‘metaGRS’. a, PRS were tested in MVP Whites, Blacks, and Hispanics, separately. In addition to all cases combined, subgroups of incident only cases (after enrollment), severe cases with evidence of either a myocardial infarction and/or a revascularization procedure, and early onset vs older onset cases (divided by median age of onset) were tested. b, PRS are tested for burden of coronary atherosclerosis only among Whites. The reference group is subjects with normal coronaries on angiography. For progressively higher burdens of disease are tested against the reference group including non-obstructive disease (‘Non-obs.’), 1-vessel disease (1V), 2-vessel disease (2V), and 3-vessel or left main disease (3V/LM). c, The best performing score in a and b, the meta-GRS, is tested for association with Phecodes, clinical labs and anthropomorphic measures, as well as selected components of the baseline questionnaires among whites with no EHR evidence of atherosclerosis related complications at the end of EHR follow up.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MVPGWASCADSupplementaryTextv4.pdf