# Genetics of Blood Lipids Among ~300,000 Multi-Ethnic Participants of the Million Veteran Program

Derek Klarin\*, M.D.1,2,3, Scott M. Damrauer\*, M.D. 4,[[1]](#footnote-1), Kelly Cho, Ph.D., M.P.H. [[2]](#footnote-2), Yan

V. Sun, Ph.D.[[3]](#footnote-3), Tanya M. Teslovich, Ph.D.[[4]](#footnote-4), Jacqueline Honerlaw, R.N., B.S.N.,

M.P.H.6, David R Gagnon, M.D., Ph.D., M.P.H. 6,[[5]](#footnote-5), Scott L. DuVall, Ph.D.[[6]](#footnote-6),[[7]](#footnote-7), Jin Li,

Ph.D.[[8]](#footnote-8),[[9]](#footnote-9), Gina Peloso Ph.D.9, Mark Chaffin, M.Sc., B.S.1,2, Jie Huang, M.D., Ph.D.6,

Hua Tang, Ph.D.[[10]](#footnote-10), Julie Lynch, Ph.D., R.N.,10,[[11]](#footnote-11), Yuk-Lam Ho, M.P.H.6, Dajiang Liu

Ph.D.16, Connor A. Emdin, D.Phil.1,2, Alex H. Li, Ph.D.8, Jennifer S. Lee, M.D.,

Ph.D.12,13, Pradeep Natarajan, M.D., M.M.Sc.1,2,17,18, Rajiv Chowdhury, Ph.D.19, Danish

Saleheen, M.D., Ph.D.5,20, Marijana Vujkovic, Ph.D. 5,20, Aris Baras, M.D.8, Saiju

Pyarajan, Ph.D.6,21, Emanuele Di Angelantonio, Ph.D.19, Benjamin M. Neale, Ph.D. 2,22,23,

Aliya Naheed, Ph.D.24, Amit V. Khera, M.D.1,2, John Danesh, FMedSci19, Kyong-Mi Chang, M.D.5,25, Gonçalo Abecasis, D.Phil.26, Cristen Willer, Ph.D.27-29, Frederick E.

Dewey, M.D.8, David J. Carey, Ph.D. 30, Global Lipids Genetics Consortium, Myocardial

Infarction Genetics (MIGen) Consortium, The Geisinger-Regeneron DiscovEHR Collaboration, The VA Million Veteran Program, John Concato, M.D. 31,32, J. Michael

Gaziano, M.D.6,19, M.P.H., Christopher J. O’Donnell, M.D. 6, Philip S. Tsao, Ph.D.12,13,

Sekar Kathiresan\*, M.D.1,2, Daniel J. Rader\*, M.D.33, Peter W.F. Wilson\*, M.D. 34,35,

Themistocles L. Assimes\*, M.D., Ph.D.12,13

# Affiliations

1. Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston MA, USA
2. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard,

Cambridge MA, USA

1. Department of Surgery, Massachusetts General Hospital, Boston MA, USA
2. Department of Surgery, Perlman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
3. Department of Public Health Sciences, Institute of Personalized Medicine, Penn State

College of Medicine, Hershey PA, USA

1. Cardiovascular Research Center, Massachusetts General Hospital, Harvard Medical

School, Boston MA, USA

1. Department of Medicine, Harvard Medical School, Boston MA, USA
2. Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care,

University of Cambridge, Cambridge UK

1. Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA, USA
2. Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston MA, USA

22Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston MA, USA

1. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard,

Cambridge MA, USA

1. Initiative for Noncommunicable Diseases, Health Systems and Population Studies Division, International Centre for Diarrhoeal Disease Research, Bangladesh.
2. Department of Medicine, Perlman School of Medicine, University of Pennsylvania, Philadelphia PA, USA
3. Center for Statistical Genetics, Department of Biostatistics, University of Michigan

School of Public Health, Ann Arbor MI, USA

1. Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor MI, USA
2. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor MI, USA
3. Department of Human Genetics, University of Michigan, Ann Arbor MI, USA
4. Geisinger Health System. Danville PA 17821
5. Clinical Epidemiology Research Center, VA Connecticut Healthcare System, West Haven CT, USA
6. Department of Medicine, Yale School of Medicine, New Haven CT, USA
7. Institute for Translational Medicine and Therapeutics, Perlman School of Medicine,

University of Pennsylvania, Philadelphia PA, USA

1. Atlanta VA Medical Center, Decatur GA, USA
2. Emory Clinical Cardiovascular Research Institute, Atlanta GA, USA \* These authors contributed equally

Abstract word count: 150, total Main Body Word Count (Excluding Abstract): 3320 **Corresponding Author:**

Themistocles L. Assimes, M.D., Ph.D.,

Stanford University School of Medicine & VA Palo Alto Health Care System,

Suite 300, 1070 Arastradero Road, Palo Alto, CA 94304-1334 Tel: (650) 498-4154

tassimes@stanford.edu

**Subject Codes**: Lipids, population genetics, genome-wide association studies, coronary artery disease

# Abstract

The Million Veteran Program was established in 2011 as a national research initiative to determine how genetic variation influences the health of U.S. veterans. We genotyped 312,571 participants of this program using a custom biobank array and linked the genetic data to laboratory and clinical phenotypes extracted from electronic health records covering a median of 10.0 years of follow up. Among 297,626 veterans with at least 1 blood lipid measure including 57,332 blacks and 24,743 Hispanics, we tested up to ~32 million genetic variants for association with blood lipid levels and identified 118 novel genome-wide significant loci after meta-analysis with data from the Global Lipids Genetics Consortium (total N > 600,000). Through a focus on mutations leading to a loss of gene function and a phenome-wide association study, we propose novel indications for pharmaceutical inhibitors targeting *PCSK9 (*abdominal aortic aneurysm), *ANGPTL4*

(type 2 diabetes), and *PDE3B* (triglycerides and coronary disease).

# Main Text

Large-scale biobanks offer the potential to link genes to health traits documented in electronic health records (EHR) with unprecedented power1. In turn, these discoveries are expected to improve our understanding of the etiology of common and complex diseases as well as our ability to treat and prevent these conditions. To this end, the Million

Veteran Program (MVP) was established in 2011 by the Veteran Affairs (VA) Office of Research and Development as a nationwide research program within the VA healthcare system2. The overarching goal of MVP is to reveal new biologic insights and clinical associations broadly relevant to human health and to enhance the care of veterans through precision medicine.

Blood concentrations of total cholesterol (TC), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL), and triglycerides (TG) are heritable risk factors for atherosclerotic cardiovascular disease3, a highly prevalent condition among U.S. veterans. Genome-wide association studies (GWAS) to date have identified at least 268 loci that influence these levels4-12, many of which are under investigation as

potential therapeutic targets13,14. However, off-target effects have dampened enthusiasm

for some of these molecules15,16. Understanding the full spectrum of clinical consequences of a genetic variant through phenome-wide association scanning

(“PheWAS”17) may shed light on potential unintended effects as well as novel therapeutic indications for some of these molecules.

We first performed a GWAS, including a discovery phase in MVP and a replication phase in the Global Lipids Genetics Consortium (GLGC) (**Fig. 1)**. In the discovery phase, we performed association testing among 297,626 white, black, and Hispanic MVP participants with blood lipids stratified by ethnicity followed by a metaanalysis of results across all three groups. Replication of MVP findings was conducted in one of two independent studies from the GLGC followed by association testing of novel loci with CAD in the Cardiogram+C4D or the Cardiogram+C4D Exome consortia. We then examined novel, genome-wide lipid-associated, low-frequency missense variants unique to black and Hispanic individuals. We focused on results for predicted loss of gene function (pLoF) mutations, as these associations have revealed target pathways for

pharmacologic inactivation and modulation of cardiovascular risk14,18,19. Finally, we performed a PheWAS for a set of DNA sequence variants within genes that have already emerged as therapeutic targets for lipid modulation, leveraging the full catalog of ICD-9 diagnosis codes in the VA EHR to better understand the potential consequences of pharmacologic modulation of these genes or their products.

# Results

*Demographic and Clinical Characteristics of Genotyped Participants in the Million*

## Veteran Program

A total of 353,323 veterans had genetic data available in MVP, with clinical phenotypes recorded in the VA EHR over 3,088,030 patient-years prior to enrollment (median of 10.0 years per participant) and 61,747,974 distinct clinical encounters (median of 99 per participant). We categorized veterans into three mutually exclusive ancestral groups for association analysis: 1) non-Hispanic whites, 2) non-Hispanic blacks, and 3) Hispanics. Admixture plots depicting the genetic background of the black and Hispanic groups are shown in **Supplementary Figures 1 and 2**. Demographics and participant counts for a number of cardiometabolic traits for the 312,571 white, black, and Hispanic MVP participants that passed our quality control are depicted in **Table 1**.

A subset of 297,626 participants passing quality control had at least 1 laboratory measurement of blood lipids in their EHR. These individuals collectively had a total of 15,456,328 lab entries for blood lipids, or a median of 12 measures per lipid fraction per participant. To minimize potential confounding from the use of lipid-altering agents with variable adherence, we selected a participant’s maximum LDL cholesterol, TG, and TC as well as his or her minimum HDL cholesterol for genetic association analysis20. **Table 2** summarizes characteristics at enrollment and the distribution lipid levels for MVP participants included in our analysis**.** As expected, participants were largely male but 28% were of non-European ancestry. While approximately 45% had evidence of a statin prescription at the time of enrollment, only 8 to 9% participants had such evidence at the time of their maximum LDL or TC measurement used for our GWAS analysis.

## Genetic Association Analysis of Lipids

We successfully imputed 19.3, 31.4, and 30.4 million variants in white, black, and

Hispanic veterans, respectively, using the 1000 Genomes Project21 reference panel (**Table 2**). Black and Hispanic participants had substantially more variants available for analysis, reflecting the known greater genetic diversity within these populations21,22 (**Fig 2a,b)**. We also identified 6,657 pLoF variants in 4,294 genes across the three ethnicities (**Fig. 2c**).

We compared the Z scores and effect estimates from the published literature with those observed in MVP for >2,000 previously reported5 genome-wide significant variants for lipids that were imputed using the HapMap reference panel23. We found a strong correlation of genetic associations across all four traits, validating the lipid data secured through the EHR (**Supplementary Fig. 3,4**).

We performed association testing separately among individuals of each of three ancestries (whites, blacks, and Hispanics) in our initial discovery analysis and then metaanalyzed results across ancestry groups using an inverse variance-weighted fixed effects method (**Fig. 1a, Supplementary Fig. 5**). Following trans-ethnic meta-analysis in the discovery phase of our study, a total of 46,526 variants at 188 of the 268 known loci for lipids met the genome-wide significance threshold (P < 5x10-8) (**Supplementary Tables 1-4**). We performed pairwise comparisons of the allele frequencies and effect estimates between whites and blacks as well as between whites and Hispanics for 354 of the 444 previously established independent genome-wide significant variants for lipids which were well imputed in all three ancestral groups in MVP (**Fig. 3**)11. We observed a much stronger correlation between white and Hispanic effect allele frequencies (Pearson correlation coefficient R = 0.96) than between whites and blacks (R= 0.72), likely reflecting the greater European admixture in the MVP Hispanic participants. The correlation in effects estimates among the three ethnicities varied by lipid trait (**Fig. 3, Supplementary Fig. 6**).

We sought replication for variants within MVP with suggestive associations (P < 1 x 10-4) in one of two independent studies (**Figure 1b**): the 2017 GLGC exome array meta-analysis11 or the 2013 GLGC “joint meta-analysis5.” A total of 170,925 variants demonstrated suggestive association (P<10-4) in the MVP discovery analysis. Among these variants, 39,663 were also available for *in silico* replication in at least one of the two GLGC studies involving up to 319,677 additional individuals. We defined significant novel associations as those that were at least nominally significant in replication (P<0.05) with consistent direction of effect and had an overall P < 5 x10-8 (genome-wide significance) in the discovery and replication cohorts combined. Following replication,

118 novel loci (from 142 lead variants) exceeded genome-wide significance (P < 5x10-8, **Supplementary Tables 5-8)**.Minor allele frequencies (MAF) of lead variants ranged from 0.08% to 49.9%, with effect sizes ranging from 0.01 to 0.243 standard deviations.

For example, carriers of a rare missense mutation in the gene encoding Sorting Nexin-8 [*SNX8* p.Ile414Thr, (rs144787122) MAF = 0.35% in MVP] demonstrated a 0.10 standard deviation (3.8 mg/dL) higher plasma LDL cholesterol after testing in 587,481 individuals.

*Low-Frequency Missense Variant Lipid Associations Specific to Blacks and Hispanics* We next focused on ancestry specific low-frequency (MAF < 5%) missense variants, as

these variants have been suggested to have a higher likelihood of causality24,25. We identified several novel low-frequency missense variants associated with one or more lipid levels at genome wide significance that were specific to blacks or Hispanics. We found a total of 5 variants associated with LDL cholesterol and/or TC among blacks

(**Supplementary Table 9)** and 2 associated with HDL cholesterol and/or TC among Hispanics (**Supplementary Table 10**) in *PCSK9*, *LDLR*, *APOB*, and *ABCA1*. All 10 associations were directionally consistent in the 2017 GLGC exome chip meta-analysis with 9 reaching nominal significance (p < 0.05) among 17,009 blacks and 5,084 Hispanics in GLGC. In addition, the 7 variants we identified were either monomorphic or had a MAF of < 0.0005 in the ~215,000 white veterans in MVP. Of note, we observed the low-frequency 443Thr allele in *PCSK9* within Hispanics to be 8 fold more common in blacks (MAF = 0.011 in Hispanics versus 0.092 in blacks). We also found this variant to be associated with TC in blacks at genome-wide significance.

## Predicted Loss of Gene Function Lipid Associations

We focused next on the subset of genotyped or imputed pLoF variants. A total of 15 unique pLoF variants demonstrated genome-wide significant lipid associations across individuals of all three ethnic groups (**Supplementary Table 11**). We replicated known pLoF associations at *PCSK919*, *APOC318*, *ANGPTL88*, *LPL26*, *CD3627*, and *HBB28,* and we observed genome-wide significant associations of comparable magnitude of effect in each of the three ethnic groups for 2 pLoF variants: *APOC3* c.55+1G>A and *LPL* p.Ser747Ter.

We identified one novel pLoF association. Among white MVP participants, carriers of a rare stop-gain mutation in *PDE3B* (p.Arg783Ter; carrier frequency of 1 in 625), exhibited a 4.72 mg/dL (0.41 standard deviations) higher blood HDL cholesterol (P

< 2.8 x 10-16) and 43.3 mg/dL (-0.27 standard deviations) lower blood TG (P = 7.5x10-8). We found this signal to be independent of a previously reported genome-wide significant association in the region involving a common polymorphism, rs103737811 (p.Arg783Ter conditional analysis P = 6.3 x 10-16 for HDL cholesterol, and P = 8.91 x 10-8 for TG). We also identified one individual who was homozygous for p.Arg783Ter. This *PDE3B* “human knockout” was in his sixth decade of life HDL cholesterol and TG levels of 73 and 56 mg/dL, respectively. He was not on lipid-lowering medication and was free of coronary artery disease (CAD). We replicated the TG and HDL associations for this

pLoF variant in an independent sample of ~45,000 participants of the DiscovEHR study

(**Fig. 4a,b**).

## Loss of PDE3B function and risk of Coronary Artery Disease

Hypothesizing that mutations damaging or causing a loss of function in *PDE3B* could protect against the development of CAD based on their association with lifelong lower levels of TG in blood, we conducted a case-control study of CAD involving 5 cohorts: MVP, UK Biobank, Myocardial Infarction Genetics Consortium (MIGen), Penn Medicine Biobank (PMBB), and DiscovEHR. For 3 studies that underwent exome sequencing (MIGen, PMBB, DiscovEHR), we combined pLoF variants with missense variants predicted to be damaging or possibly damaging by *each* of 5 computer prediction algorithms (LRT score, MutationTaster, PolyPhen-2, HumDiv, PolyPhen-2 HumVar, and SIFT) as performed previously26,29. Because damaging mutations are individually rare, we aggregated them in subsequent association analysis with CAD (**Supplementary Table 12**). Among 103,580 individuals with CAD and 566,813 controls available for meta-analysis in these 5 cohorts, carriers of damaging *PDE3B* mutations were found to have a 24% decreased risk of CAD (OR = 0.76, 95% CI = 0.65-0.90, P = 0.0015, **Fig.**

**4c**).

## Novel Lipid Loci and Association with Coronary Disease

To further evaluate whether novel lipid variants identified in our analysis also influence the risk of CAD, we examined the association of lead variants within the 118 novel lipid loci identified in our study with CAD in either the CARDIoGRAMplusC4D 1000 genome GWAS30 or the MIGen and CARDIoGRAM Exome Chip GWAS analysis31

when a variant was not available in the former dataset. In total, 25 of the 118 loci showed at least nominal (P < 0.05) association with CAD in the CARDIoGRAM studies

(**Supplementary Table 13**). Notably, the previously identified lead CAD 9p21 locus (rs1333048, CAD P = 5.7 x 10-94) is also associated with LDL cholesterol and TC at genome-wide significance in our study. However, the LDL raising allele is in the opposite direction of the CAD effect estimate, suggesting that the causal variant(s) at 9p21 may confer CAD risk outside of a lipid pathway as implicated by preliminary functional work at the locus32. We then examined the direction of effect for LDL cholesterol, TG, and HDL cholesterol raising alleles on CAD for the 118 novel loci in our analysis. Consistent with prior observations, the 32 LDL and 63 TG raising alleles (lipid P < 10-4) were more likely to be associated with an increase risk of CAD (two-tailed binomial P = 0.05 and 3.8 x 10-5 for LDL and TG, respectively) but the same was not true for 9 alleles associated with a higher HDL (P < 10-4) but not with LDL or TG (P >

0.05)(two-tailed binomial P = 0.25).

## PheWAS of Low-Frequency variants in genes targeted by lipid therapies

We leveraged a median of 65 unique ICD-9 diagnosis codes per participant prior to enrollment in MVP to explore the spectrum of phenotypic consequences of genetic variation within genes targeted by lipid-lowering medicines. We selected five lipid genes currently being targeted by pharmaceutical agents and identified functional variants in these genes: two nonsense variants (*LPL* p.Ser474Ter, *ANGPTL8* p.Gln121Ter) and three missense variants (*ANGPTL4* p.Glu40Lys, *APOA5* p.Ser19Trp, *PCSK9* p.Arg46Leu). We considered phenotypes to be significantly associated with a variant if they met a Bonferroni corrected P < 4.98 x 10-5 [0.05/1004 traits], a conservative threshold given the correlation structure present among PheWAS phenotypes33.

A total of 176,913 white veterans were available for analysis after quality control. Among these individuals, we identified 33 statistically significant phenotypic associations across the 5 variants (**Supplementary Table 14)**. We replicated known associations with CAD for *LPL26*, *ANGPTL414*, and *PCSK919*. Of note, carriers of TGlowering/HDL cholesterol-raising mutations in *ANGPTL4* (p.Glu40Lys, 7,013 carriers) were also found to have a reduced risk of type 2 diabetes (**Fig. 5)**and carriers of LDL cholesterol-lowering mutations in *PCSK9* (p.Arg46Leu, 5,537 carriers) also demonstrated a reduced risk of abdominal aortic aneurysm (**Fig. 6**). Lastly, we found no association between *ANGPTL4* p.Glu40Lys and abdominal lymphatic disorders (**Supplementary** **Table** **15**) to support recent reports of mesenteric lymphadenopathy secondary to granulomatous lipid accumulations in mice and primates injected with a monoclonal antibody against Angptl415 .

# Discussion

We leveraged clinical and genetic data from the Million Veteran Program to investigate the inherited basis of blood lipids in nearly 300,000 U.S. veterans. Our investigation resulted in four key findings. First, we robustly confirmed 188 previously identified loci while concurrently uncovering an additional 118 novel genome-wide significant loci whose collective relationship to the risk of CAD was found to be consistent with prior studies. Second, we identified ancestry-specific effects of rare coding variation on lipids among white, black, and Hispanic participants. Third, we identified 15 pLoF mutations associated with lipids at a genome-wide level of significance, including a proteintruncating variant in *PDE3B* that lowers TG, raises HDL cholesterol, and protects against CAD. Finally, we examined the full spectrum of phenotypic consequences for mutations in lipid genes emerging as therapeutic targets, identifying protective effects of functional mutations in *PCSK9* for abdominal aortic aneurysm and in *ANGPTL4* for type 2 diabetes.

We glean four main insights through our findings. First, we confirm the enormous potential of a large-scale multi-ethnic biobank built within an integrated health care system in the discovery of the genetic basis of human traits. Specifically, we leveraged the VA’s mature nationwide EHR to efficiently extract existing repeated laboratory measures of lipids collected during the course of clinical care in nearly 300,000 veterans over a median of 10 years for GWAS analysis. Subsequent meta-analysis (combined N>600,000) with existing datasets increased the number of known independent genetic lipid associations to nearly 400. Our results highlight multiple lipid pathways with links to human disease. For example, common variants near genes such as *COL4A2* and *ITGA1* identified for LDL cholesterol/TC suggest links to extracellular matrix and cell

adhesion biology, two pathways recently implicated by GWAS of CAD34,35. We also demonstrated that carriers of a rare missense mutation in the gene encoding Perilipin-1 (*PLIN1* p.Leu90Pro) possess a markedly higher plasma HDL cholesterol (0.243 standard deviations). In humans, Perilipin-1 is required for lipid droplet formation, triglyceride storage, as well as free fatty acid metabolism, and frameshift pLoF mutations Perilipin-1 have been reported to result in severe lipodystrophy36. A variant downstream of *BDNF* (encoding Brain-Derived Neurotrophic Factor) was found to be associated with HDL cholesterol and TG levels, supporting recent evidence linking this gene with metabolic syndrome and diabetes37. These findings not only improve our understanding of the genetic basis of dyslipidemia, but also provide insights into targets for the development of novel therapeutic agents.

Our second insight embraces the benefit of studying individuals with a diverse ethnic background. Such a design can provide valuable incremental information on the nature of previously identified human genetic associations. In MVP, we examined nearly 60,000 black and 25,000 Hispanic veterans for analysis, representing one of the largest - if not the largest - single-cohort GWAS to date for these ethnic groups for any trait. Among these individuals, we compared the effect estimates and allele frequencies of lipid-associated variants across ancestral group and identified 7 novel low-frequency coding variants associated with lipids only in non-European populations. Conversely, we also confirmed a shared genetic architecture across all three racial groups for pLoF variation at the *LPL* and *APOC3* loci. Previous work identifying low-frequency missense and pLoF variation in lipid genes have led to the development of the next generation of

pharmaceutical agents for cardiovascular disease14,15,38,39. Expansion of these efforts to larger sample sizes and additional ancestries may help explain differences in blood lipid levels and risk of atherosclerosis among select populations.

Our third insight centers around our findings for the deleterious exonic variants within *PDE3B*. These findings lend human genetic support to PDE3B inhibition as a therapeutic strategy for atherosclerosis. Cilostazol, an inhibitor of both the 3A and 3B isoforms of the phosphodiesterase enzyme, is known to have anti-platelet40, vasodilatory41, and inotropic42 effects via inhibition of PDE3A, and also has well documented substantial effects on TG and HDL cholesterol levels43 — likely through antagonism of PDE3B. We demonstrate that a *PDE3B* pLoF variant recapitulates the known lipid effects of cilostazol, and extend these findings to show that damaging *PDE3B* mutations are also associated with reduced risk of CAD. Randomized control trials to date have demonstrated cilostazol’s efficacy in intermittent claudication43 and prevention of restenosis following percutaneous coronary intervention44. The drug is also currently used off-label for the prevention of stroke recurrence through a presumed antiplatelet effect45. We note that mice genetically deficient in *Pde3b* display reduced atherosclerosis46 as well as decreased infarct size and improved cardiac function following experimental coronary artery ligation47. In light of our findings, use of cilostazol, or one of its derivatives, for the primary or secondary prevention of CAD deserves further consideration.

Our final insight highlights the potential benefit of phenome-wide association scanning across a large-scale EHR-based biobank to predict both potentially adverse as well as beneficial consequences of artificially inhibiting gene function. Here, we provide evidence that pharmacologic *PCSK9* inhibition may reduce abdominal aortic aneurysm risk in addition to its known effects on atherosclerotic cardiovascular disease13.

Similarly, we expand on the potential indications for *ANGPTL4* inhibition to include type 2 diabetes. Future PheWAS efforts may reveal associations that facilitate prioritization of drugs currently in development, repurposing of therapies already in clinical use, or prediction of adverse or off-target effects prior to investigation through expensive and time-consuming clinical trials.

Several limitations deserve mention. First, our MVP lipid phenotype definitions

are based entirely on EHR data with a high prevalence of use of lipid-lowering therapy at enrollment. We used maximum or minimum values to capture untreated lipid levels, but the possibility of misclassification of lipid levels remains for participants entering the VA healthcare system on therapy. Such misclassification, however, would be expected to generally reduce our power to detect genetic associations. Second, participants in MVP are overwhelmingly male. Although almost 25,000 women were included in our analysis, we did not attempt to detect genetic associations specific to females or heterogeneity of effects between sexes due to suspected limited power. Lastly, power to detect associations with less common diseases in our PheWAS may also be limited despite the overall number of participants included in the analysis.

In conclusion, we identified >100 new genetic signals for blood lipid levels utilizing a biobank that exploits existing EHRs of U.S. veterans. We demonstrate the potential of this approach in the discovery of novel genetic associations and the development of novel therapeutic agents.

# Online Methods

The design of the Million Veteran Program (MVP) has been previously described2.

Briefly, individuals aged 19 to 104 years have been recruited from more than 50 VA Medical Centers nationwide since 2011. Each veteran’s EHR data are being integrated into the MVP biorepository, including inpatient International Classification of Diseases (ICD-9) diagnosis codes, Current Procedural Terminology (CPT) procedure codes, clinical laboratory measurements, and reports of diagnostic imaging modalities. The MVP received ethical and study protocol approval from the VA Central Institutional Review Board (IRB) in accordance with the principles outlined in the Declaration of

Helsinki.

## Genetic Data

DNA extracted from whole blood was genotyped using a customized Affymetrix Axiom biobank array, the MVP 1.0 Genotyping Array. With 723,305 total DNA sequence variants, the array is enriched for both common and rare variants of clinical significance in different ethnic backgrounds. Veterans of three mutually exclusive ethnic groups were identified for analysis: 1) non-Hispanic whites, 2) non-Hispanic blacks, and 3) Hispanics. Quality-control procedures used to assign ancestry, remove low-quality samples and variants, and perform genotype imputation to the 1000 Genomes reference panel21 are described in the supplementary note.

## EHR-Based Lipid Phenotypes

EHR clinical laboratory data were available for MVP participants from as early as 2003. We extracted the maximum LDL cholesterol/TG/TC, and minimum HDL cholesterol for each participant for analysis. These extreme values were selected to approximate plasma lipid concentrations in the absence of lipid lowering therapy as described previously20.

For each phenotype (LDL cholesterol, natural log transformed TG, HDL cholesterol, and TC), residuals were obtained after regressing on age, age2, sex, and 10 principal components of ancestry. Residuals were subsequently inverse normal transformed for association analysis. Statin therapy prescription at enrollment was defined as the presence of a statin prescription in the EHR within 90 days before or after enrollment in MVP. Statin therapy prescription at the maximum lipid measurement was defined as the presence of a statin prescription in the EHR within 90 days prior to the maximum lipid laboratory measurement used in our GWAS analysis. Further details of lipid phenotype quality control are described in the supplementary note.

## MVP Association Analysis

Genotyped and imputed DNA sequence variants with a MAF > 0.0003 were tested for association with the inverse normal transformed residuals of lipid values through linear regression assuming an additive genetic model. In our initial discovery analysis, we performed association testing separately among individuals of each of three genetic ancestries (whites, blacks, and Hispanics) and then meta-analyzed results across ethnic groups using an inverse variance-weighted fixed effects method. For variants with suggestive associations (association P < 10-4), we sought replication of our findings in one of two independent studies: the 2017 GLGC exome array meta-analysis11 or the 2013 GLGC “joint meta-analysis5.” Replication was first performed using summary statistics from the 2017 GLGC exome array study. A total of 242,289 variants in up to 319,677 individuals were analyzed after quality control and were available for replication. If a DNA sequence variant was not available for replication in the above exome array-focused study, we sought replication from publicly available summary statistics from the 2013 GLGC “joint meta-analysis.” An additional 2,044,165 variants in up to 188,587 individuals were available for replication in this study. In total, 2,286,454 DNA sequence variants in up to 319,677 individuals were available for independent replication. If a variant was available for replication in both studies, we prioritized replication using summary statistics from the 2017 GLGC exome array study given its larger sample size. We defined significant novel associations as those that were at least nominally significant in replication (P<0.05) and had an overall P < 5 x10-8 (genome-wide significance) in the discovery and replication cohorts combined. Further details of the association analysis are described in the supplementary note.

## Identification of Independent Low-Frequency Coding Variant Lipid Associations Specific to Blacks and Hispanics

We used the P value and linkage disequilibrium-driven clumping procedure in PLINK version 1.90b (--clump) to identify associations between low-frequency coding variants and lipids specific to blacks and Hispanics. Input included summary lipid association statistics from our MVP 1000 Genomes imputed genome-wide association study of black and Hispanic individuals, and reference linkage disequilibrium panels of 661 African (AFR) and 347 Ad Mixed American (AMR) samples from 1000 Genomes phase 3 whole genome sequencing data. Variants were clumped with stringent r2 (<0.01) and P (< 5 x 10-8) thresholds in a 1 mega-base region surrounding the lead variant at each locus to reveal independent index variants at genome-wide significance. From this list of independent variants, we report novel protein-altering variants specific to blacks and

Hispanics at a MAF < 0.05.

## Loss of Gene Function Analysis

We used the Variant Effect Predictor48 software to identify pLoF DNA sequence variants defined as: premature stop (nonsense), canonical splice-sites (splice-donor or spliceacceptor) or insertion/deletion variants that shifted frame (frameshift). For the pLoF lipids analysis, we then merged these variants with data from the Exome Aggregation Consortium24 (Version 0.3.1), a publicly available catalogue of exome sequence data to confirm consistency in variant annotation. We required that pLoF DNA sequence variants be observed in at least 50 individuals, and set a statistical significance threshold of P < 5 x 10-8 (genome-wide significance).

## Loss of PDE3B Gene Function and Coronary Artery Disease

We identified a novel lipid association for a pLoF mutation in the *PDE3B* gene

(rs150090666, p.Arg783Ter). For carriers of damaging mutations in Phosphodiesterase 3B, we examined the mutation’s effects on risk for CAD using logistic regression in five separate cohorts: MVP, UK Biobank, and 3 cohorts with exome sequencing: the

Myocardial Infarction Genetics Consortium (MIGen), the Penn Medicine Biobank (PMBB), and DiscovEHR. In studies with exome sequencing, we combined pLoF variants with missense variants predicted to be damaging or possibly damaging by each of 5 computer prediction algorithms (LRT score, MutationTaster, PolyPhen-2, HumDiv, PolyPhen-2 HumVar, and SIFT) as performed previously26,29. Because any individual damaging mutation was rare, variants were aggregated together for subsequent phenotypic analysis. We performed logistic regression on disease status, adjusting for age, sex, and principal components of ancestry as appropriate. Effects of *PDE3B* damaging mutations were pooled across studies using an inverse-variance weighted fixed effects meta-analysis. Further details of participating cohorts and CAD case definitions are described in the supplementary note.We set a P < 0.05 threshold for statistical significance.

## Novel Lipid Loci and Association with Coronary Disease

To assess whether novel lipid loci in our study modulate the risk of CAD, we extracted association results for the lead variant at each locus from either the

CARDIoGRAMplusC4D CAD GWAS30 or from the MIGen and CARDIoGRAM Exome

Chip GWAS analysis31 for variants not available in the former. A two-tailed exact binomial test for goodness of fit was performed examining the expected and observed distributions of 1) LDL and 2) TG raising alleles (P < 10-4), and 3) HDL raising alleles (P < 10-4) not associated with LDL or TG (P > 0.05) and their effect on CAD risk. We tested the null hypothesis that the lipid associated variants were equally likely to increase or decrease CAD risk and set a P < 0.05 threshold for statistical significance.

## PheWAS of Variation in Genes Targeted by Lipid Lowering Therapies

For a set of DNA sequence variants within genes targeted by lipid-lowering medicines, we performed a PheWAS leveraging the full catalog of EHR ICD-9 diagnosis codes. We selected five lipid genes currently being targeted by pharmaceutical agents and identified functional variants in these genes: two nonsense variants (*LPL* p.Ser474Ter, *ANGPTL8*

p.Gln121Ter) and three missense variants (*ANGPTL4* p.Glu40Lys, *APOA5* p.Ser19Trp, *PCSK9* p.Arg46Leu). Details of PheWAS quality control, case definitions, and association analysis are described in the supplementary note. We considered phenotypes to be significantly associated with a variant if they met a Bonferroni corrected P < 4.98 x 10-5 [0.05/1004 traits].

**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 the Veterans Administration (VA) Cooperative Studies Program (CSP) award #G002. This research was also supported by three additional Veterans

Administration awards (I01-01BX03340, I01-BX002641, and I01-CX001025) and the NIH (T32 HL007734 and R01HL127564). The content of this manuscript does not represent the views of the Department of Veterans Affairs or the United States

Government.

**Acknowledgements:** Data on coronary artery disease / myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and data on coronary artery disease

/ myocardial infarction have been contributed by the Myocardial Infarction Genetics and CARDIoGRAM Exome investigators. Both datasets were downloaded from

www.CARDIOGRAMPLUSC4D.ORG.

**Author Contributions:**

*Concept and design:* D.K., T.L.A., S.M.D., K.C., K-M. C, P.S.T, S.K., D.J.R., P.W.W., J.C., J.M.G.

## Acquisition, analysis, or interpretation of data: D.K., S.M.D, Y.V.S, K.C., Y.V.S,

T.M.T., J.H., D.R.G, S.L.D., Jin L., G.P., M.C., Jie H., H.T., J.L., Y.H., D.L., C.A.E., A.H.L., J.S.L., R.C., P.N., D.S., M.V., A.B., S.P., E.D., B.M.N., A.N., A.V.K., J.D., K-

M.C., G.A., C.W., F.E.D., D.J.C.

*Drafting of the manuscript:* D.K., T.L.A.

## Critical revision of the manuscript for important intellectual content: S.M.D., K.C., P.N,

C.W., J.L., F.E.D., S.L.D., K-M. C, C.J.O., P.S.T, S.K., D.J.R, P.W.W

*Administrative, technical, or material support:* D.K., K.C., J.H., D.R.G, S.L.D, J.L., Y.H., J.C., J.M.G, C.J.O, P.S.T, P.W.W.

**Disclosures:** Dr. Kathiresan reports grant support from Regeneron and Bayer, grant support and personal fees from Aegerion, personal fees from Regeneron Genetics Center, Merck, Celera, Novartis, Bristol-Myers Squibb, Sanofi, AstraZeneca, Alnylam, Eli Lilly, and Leerink Partners, personal fees and other support from Catabasis, and other support from San Therapeutics outside the submitted work. He is also the chair of the scientific advisory board at Genomics plc. Drs. Teslovich, Alex Li, Baras, Dewey, and Carey are employees of Regeneron Pharmaceuticals.

References:

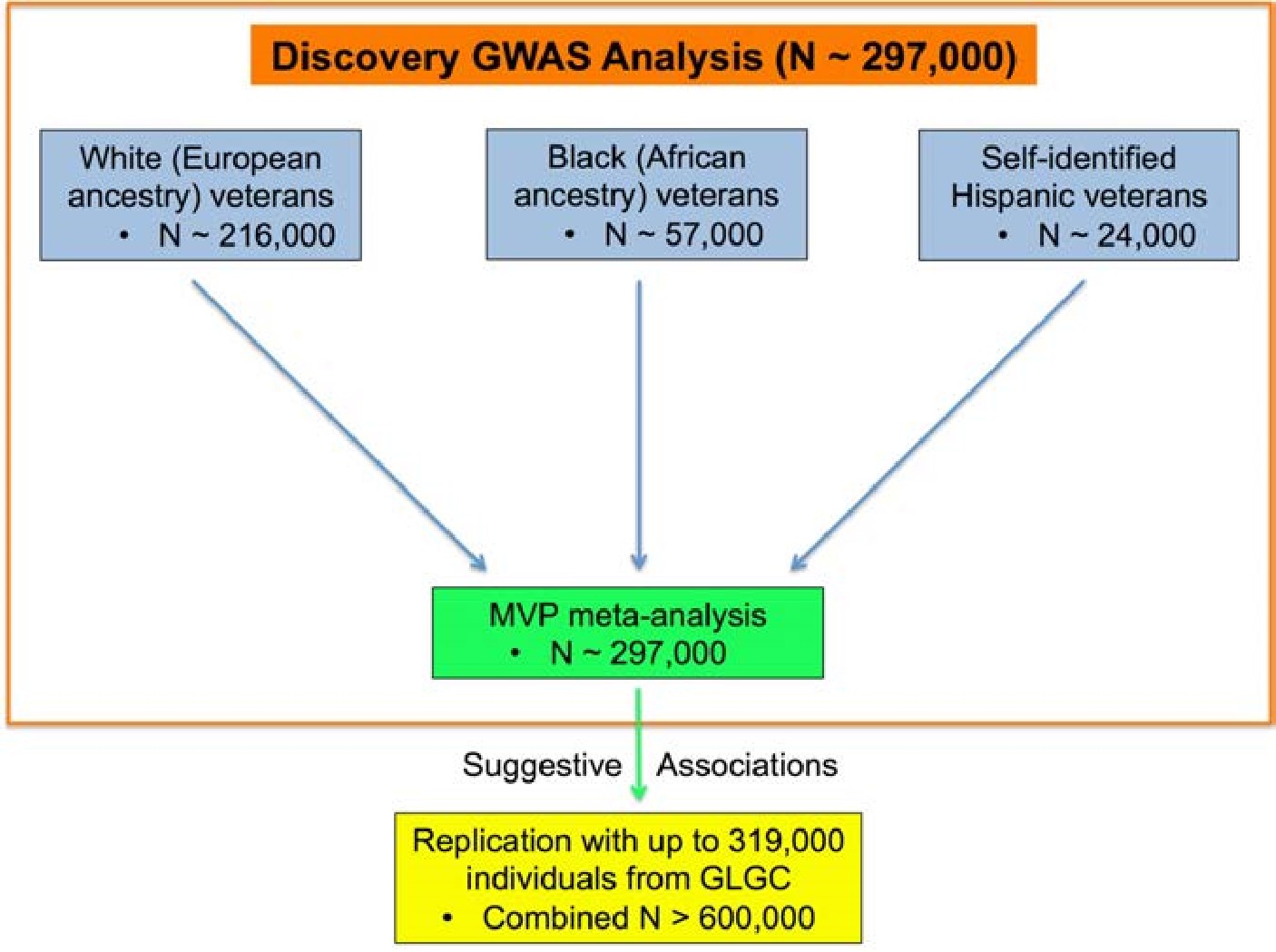
1. Collins, R. What makes UK Biobank special? *The Lancet* **379**, 1173-1174 (2012).
2. Gaziano, J.M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214-23 (2016).
3. Di Angelantonio, E. *et al.* Major lipids, apolipoproteins, and risk of vascular disease. *Jama* **302**, 1993-2000 (2009).
4. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-13 (2010).
5. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-83 (2013).
6. Chasman, D.I. *et al.* Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS Genet* **5**, e1000730 (2009).
7. Albrechtsen, A. *et al.* Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* **56**, 298-310 (2013).
8. Peloso, G.M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet* **94**, 223-32 (2014).
9. Asselbergs, F.W. *et al.* Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am J Hum Genet* **91**, 823-38 (2012).
10. Below, J.E. *et al.* Meta-analysis of lipid-traits in Hispanics identifies novel loci, population-specific effects, and tissue-specific enrichment of eQTLs. *Sci Rep* **6**, 19429 (2016).
11. Liu, D.J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nat Genet* (2017).
12. Lu, X. *et al.* Exome chip meta-analysis identifies novel loci and East Asian-specific coding variants that contribute to lipid levels and coronary artery disease. *Nat Genet* (2017).
13. Sabatine, M.S. *et al.* Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *N Engl J Med* (2017).
14. Myocardial Infarction Genetics and CARDIoGRAM Exome Consortia Investigators. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. *N Engl J Med* **374**, 1134-44 (2016).
15. Dewey, F.E. *et al.* Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease. *N Engl J Med* **374**, 1123-33 (2016).
16. Barter, P.J. *et al.* Effects of torcetrapib in patients at high risk for coronary events. *N Engl J Med* **357**, 2109-22 (2007).
17. Denny, J.C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102-10 (2013).
18. Crosby, J. *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med* **371**, 22-31 (2014).
19. Cohen, J.C., Boerwinkle, E., Mosley, T.H., Jr. & Hobbs, H.H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264-72 (2006).
20. Abul-Husn, N.S. *et al.* Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* **354**(2016).
21. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
22. Tishkoff, S.A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-44 (2009).
23. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).
24. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).
25. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* (2017).
26. Khera, A.V. *et al.* Association of Rare and Common Variation in the Lipoprotein Lipase Gene With Coronary Artery Disease. *JAMA* **317**, 937-946 (2017).
27. Dewey, F.E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**(2016).
28. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet* **47**, 1272-81 (2015).
29. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-90 (2014).
30. CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-

wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121-30 (2015).

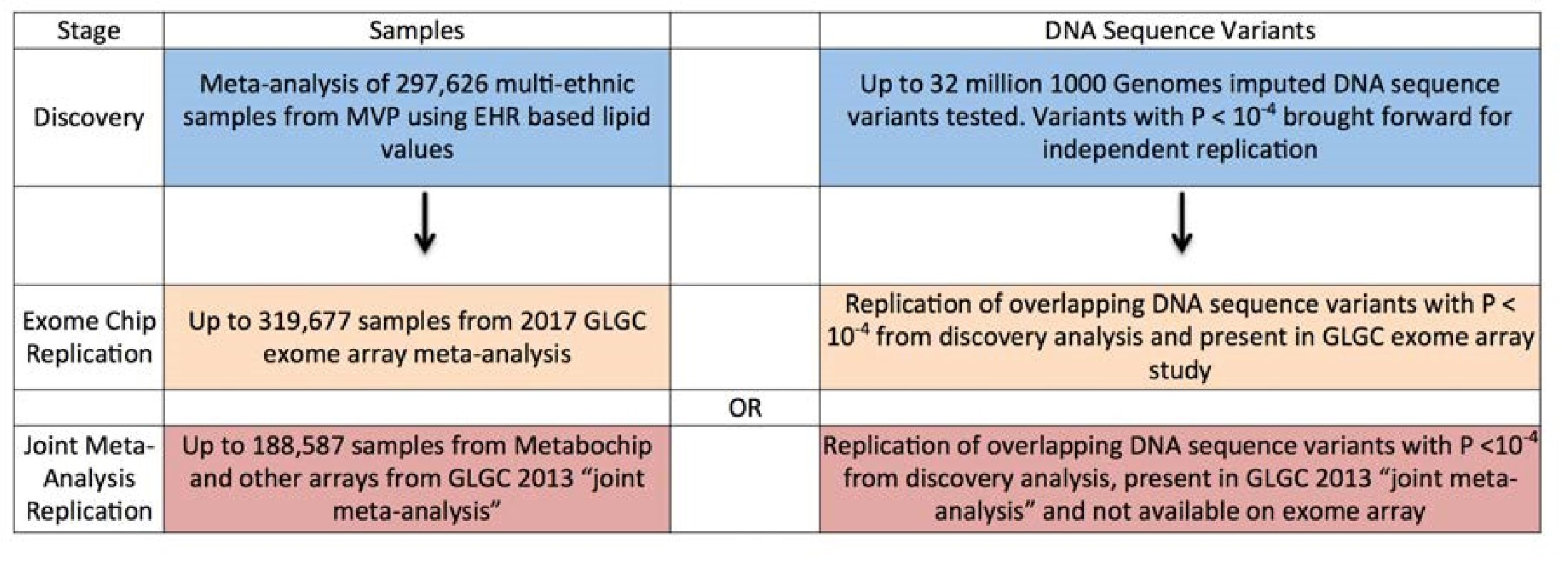
1. Webb, T.R. *et al.* Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease. *J Am Coll Cardiol* **69**, 823-836 (2017).
2. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* **470**, 264-8 (2011).
3. Diogo, D. *et al.* Phenome-wide association studies (PheWAS) across large "real-world data" population cohorts support drug target validation. *bioRxiv* (2017).
4. Klarin, D. *et al.* Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat Genet* (2017).
5. Nelson, C.P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat Genet* **49**, 1385-1391 (2017).
6. Gandotra, S. *et al.* Perilipin deficiency and autosomal dominant partial lipodystrophy. *N Engl J Med* **364**, 740-8 (2011).
7. Rani, J. *et al.* T2DiACoD: A Gene Atlas of Type 2 Diabetes Mellitus Associated Complex Disorders. *Sci Rep* **7**, 6892 (2017).
8. Musunuru, K. *et al.* Exome Sequencing, ANGPTL3 Mutations, and Familial Combined Hypolipidemia *New England Journal of Medicine* (2010).
9. Graham, M.J. *et al.* Cardiovascular and Metabolic Effects of ANGPTL3 Antisense Oligonucleotides. *N Engl J Med* (2017).
10. Zhang, W. & Colman, R.W. Thrombin regulates intracellular cyclic AMP concentration in human platelets through phosphorylation/activation of phosphodiesterase 3A. *Blood* **110**, 1475-82 (2007).
11. Maass, P.G. *et al.* PDE3A mutations cause autosomal dominant hypertension with brachydactyly. *Nat Genet* **47**, 647-53 (2015).
12. Vandeput, F. *et al.* Selective regulation of cyclic nucleotide phosphodiesterase PDE3A isoforms. *Proc Natl Acad Sci U S A* **110**, 19778-83 (2013).
13. Bedenis, R. *et al.* Cilostazol for intermittent claudication. *Cochrane Database Syst Rev*, Cd003748 (2014).
14. Tsuchikane, E. *et al.* Impact of cilostazol on restenosis after percutaneous coronary balloon angioplasty. *Circulation* **100**, 21-6 (1999).
15. Shinohara, Y. *et al.* Cilostazol for prevention of secondary stroke (CSPS 2): an aspirincontrolled, double-blind, randomised non-inferiority trial. *Lancet Neurol* **9**, 959-68 (2010).
16. Ahmad, F. *et al.* Phosphodiesterase 3B (PDE3B) regulates NLRP3 inflammasome in adipose tissue. *Sci Rep* **6**, 28056 (2016).
17. Chung, Y.W. *et al.* Targeted disruption of PDE3B, but not PDE3A, protects murine heart from ischemia/reperfusion injury. *Proc Natl Acad Sci U S A* **112**, E2253-62 (2015).
18. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).

**Figure 1**

# a)



# b)



# Figure 1. GWAS Study Design

1. DNA sequence variants across 3 separate ancestry groups in the Million Veteran Program were meta-analyzed using an inverse-variance weighted fixed effects metaanalysis in the discovery phase. Variants with suggestive association were then brought forward for independent replication.
2. DNA sequence variants with suggestive association (P < 10-4) in discovery were brought forward for independent replication and tested using summary statistics from either 1) the 2017 exome-array focused GLGC meta-analysis (exome chip replication) *or* 2) the 2013 “joint meta-analysis” (joint meta-analysis replication) from the GLGC.

Abbreviations: MVP, Million Veteran Program; GWAS, genome-wide association study; EHR, electronic health record; GLGC, Global Lipids Genetics Consortium

**Figure 2 a)**

# b) c)

Annotation

Ethnicity

Frameshift Variant

Splice Acceptor Variant

Splice Donor Variant

Stop Gained

Number of Variants

Black

Hispanic

White

0

2000

4000

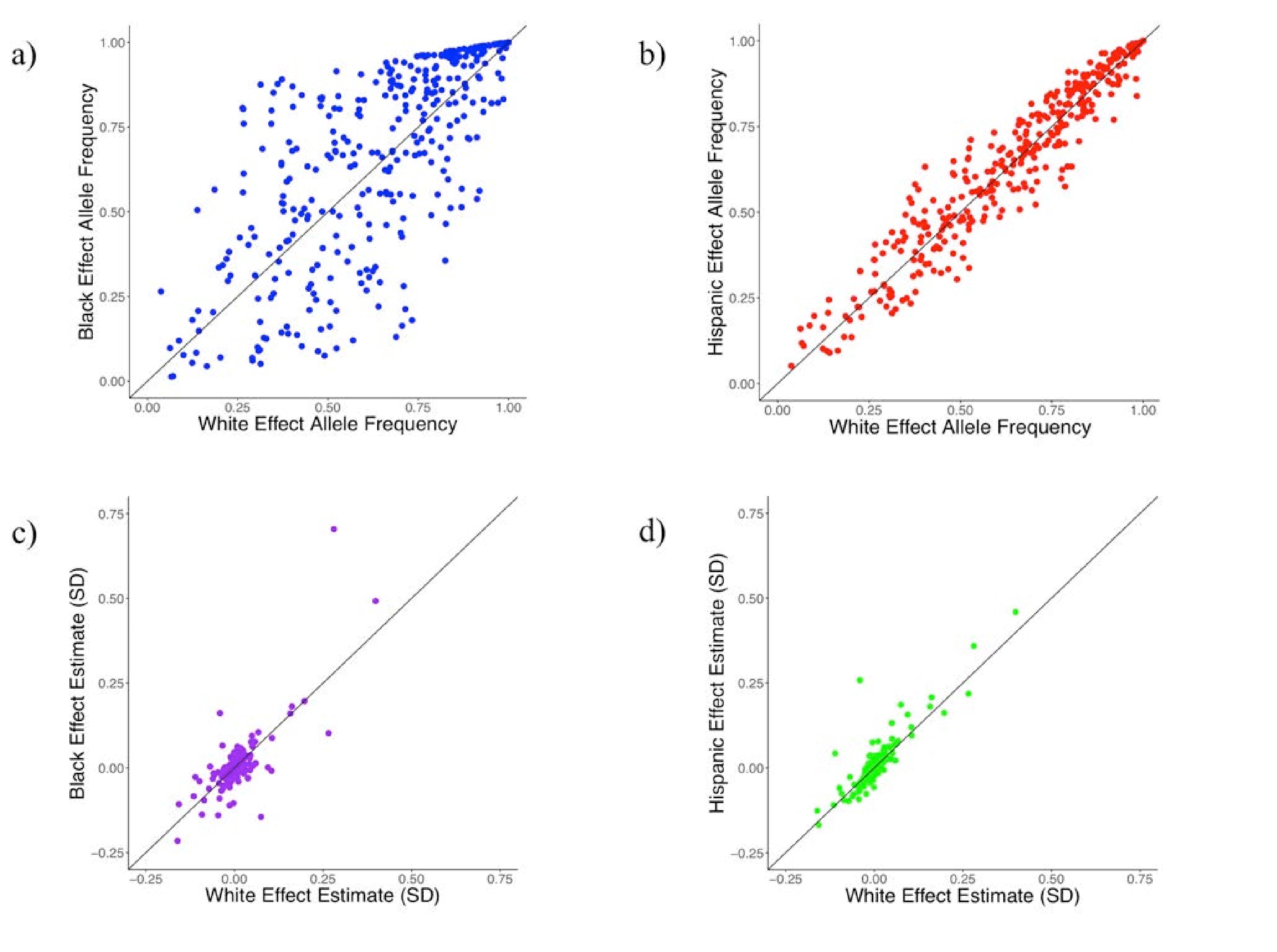
6000

**Figure 2. Genetic Variation in Million Veteran Program Participants** Histogram of rare (MAF 0.0003-0.05, **a**) and common (MAF 0.05-0.5, **b**) variants passing quality control stratified by ethnicity in the MVP.

**c)** The number of pLoF variants passing quality control for white, black, and Hispanic participants in MVP. Each pLoF annotation (frameshift, splice donor/acceptor, stop gain) is depicted by a separate color.

Abbreviations: MVP, Million Veteran Program; pLoF, predicted Loss of Function

# Figure 3



# Figure 3. Comparison of 354 Independent Lipid Associated Variants Across Ethnicities

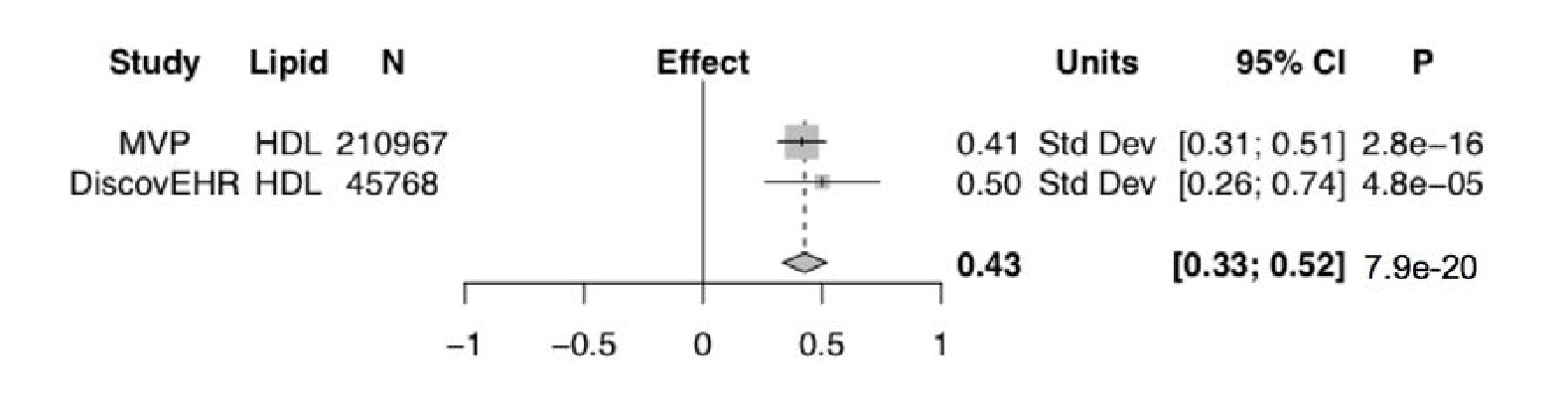
Allele frequencies observed in white individuals (x-axes) compared to black **(a**, R = 0.72,**)** or Hispanic **(b,** R = 0.96**)** individuals for lipid-associated variants.

Effect estimates for LDL cholesterol association in white individuals (x-axes) compared to black **(c,** R = 0.9**)** or Hispanic **(d,** R = 0.83**)** individuals.

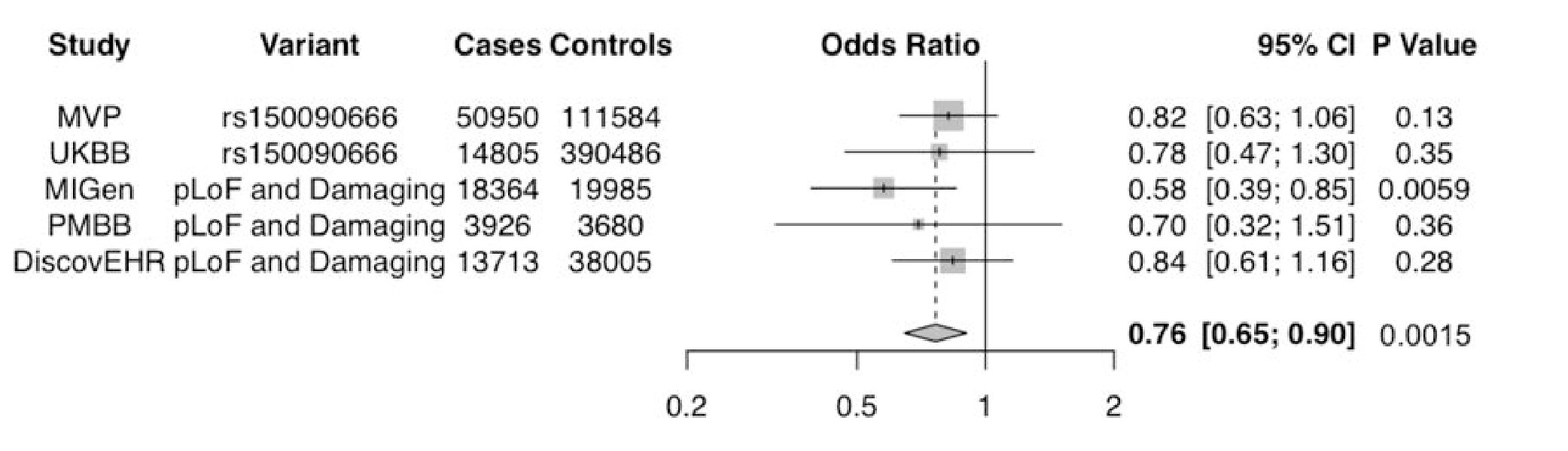
Abbreviations: SD, Standard Deviations; LDL, Low-Density Lipoprotein; R = Pearson correlation coefficient

**Figure 4**

# a)



# b) c)



# Figure 4. *PDE3B* Loss of Gene Function, Lipids, and Coronary Disease

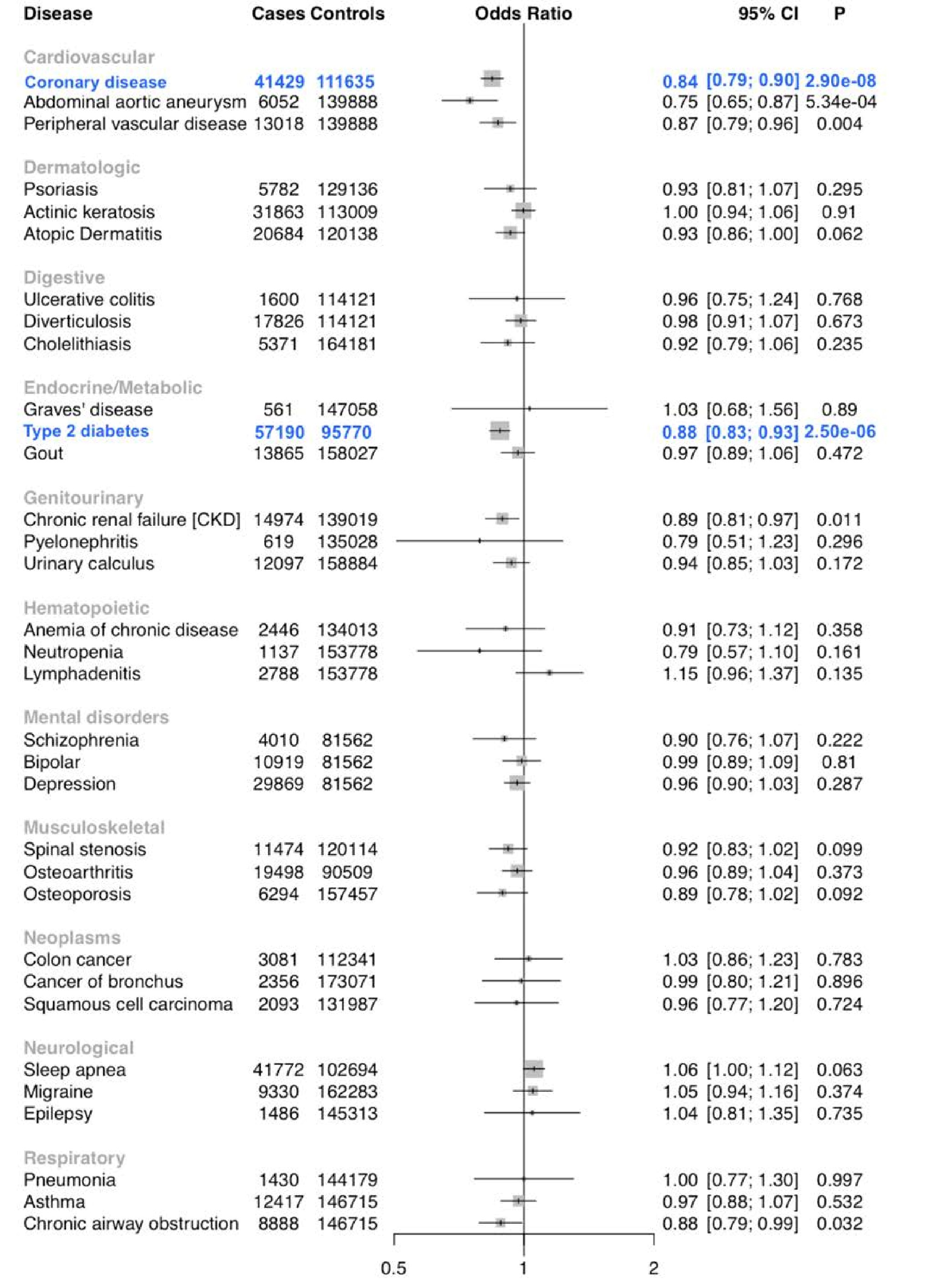
Results for the association of the predicted loss of function mutation p.Arg783Ter in *PDE3B* with HDL cholesterol **(a)** and TG **(b)** for white veterans in MVP with independent replication in the DiscovEHR study.

**c)** Meta-analysis of the association of damaging *PDE3B* mutations and coronary disease across five studies, including three (MIGen, PMBB, DiscovEHR) with exome sequencing. Results were pooled in an inverse-variance weighted fixed effects metaanalysis.

Abbreviations: MVP, Million Veteran Program; HDL, High-Density Lipoprotein; TG,

Triglycerides; UKBB, UK Biobank; MIGen, Myocardial Infarction Genetics Consortium; PMBB, Penn Medicine Biobank

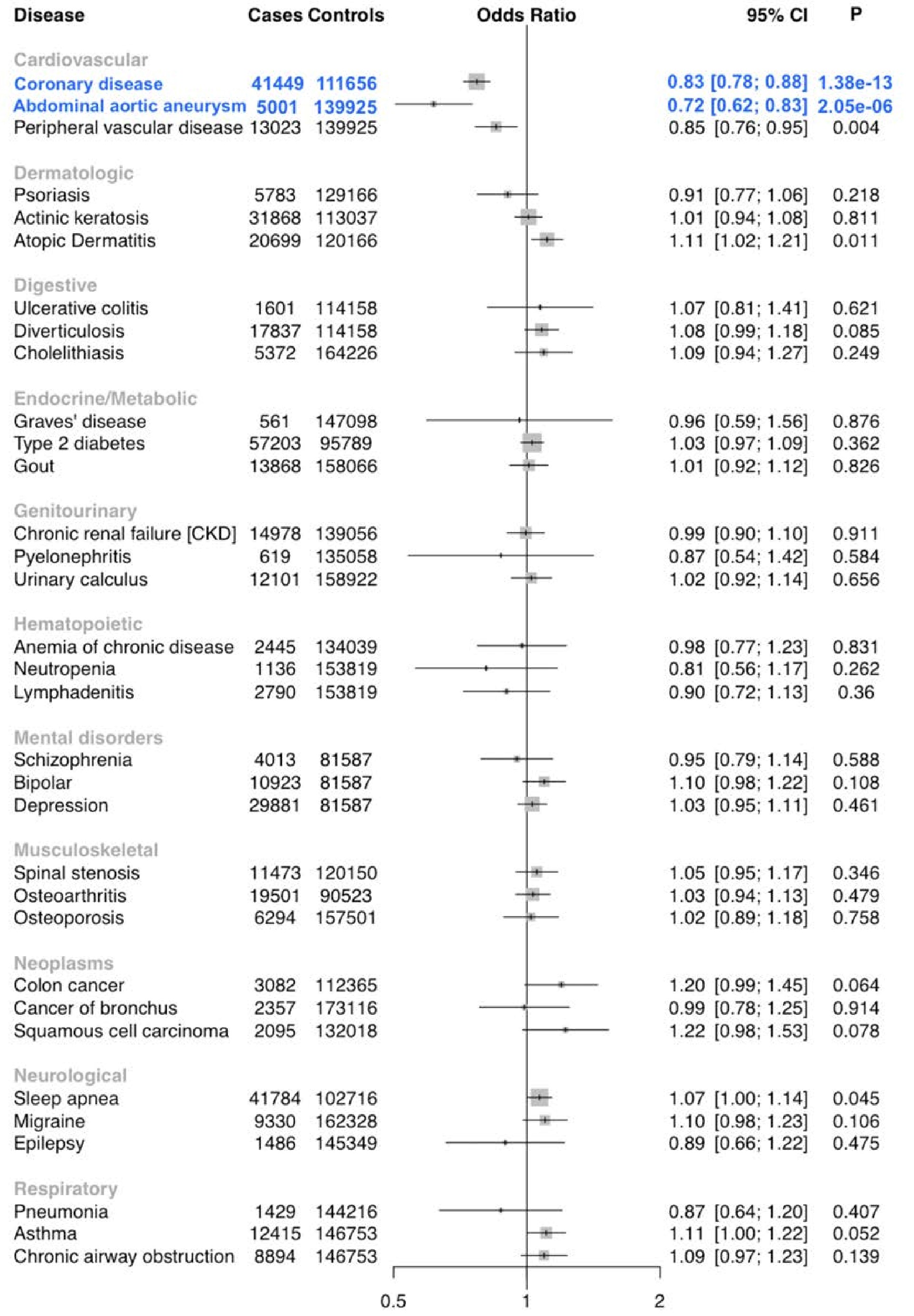
# Figure 5



# Figure 5. PheWAS associations for *ANGPTL4* 40Lys Carriers

Forest plot for a representative 33 of the 1004 disorders tested in the *ANGPTL4* p.Glu40Lys PheWAS. Statistically significant associations are shown in blue.

# Figure 6



# Figure 6. PheWAS associations for *PCSK9* 46Leu Carriers

Forest plot for a representative 33 of the 1004 disorders tested in the *PCSK9* p.Arg46Leu PheWAS. Statistically significant associations are shown in blue.

# Table 1 - Demographic and clinical characteristics of black, white, and Hispanic individuals passing quality control in the Million Veteran Program

**Basic Demographics Genotyped Veterans**

N 312,571

Age at Enrollment ± SD, years 62.4 ± 13.5

Male, n (%) 287,441 (92.0%)

Body Mass Index ± SD, kg/m2 30.3 ± 6.0

Current Smoker, n (%) 59,385 (19.0%)

Former Smoker, n (%)

N with ≥ 1 Measurement of Plasma Lipids, (%)

Number of Lipid Measurements, (Median Per

15,456,328 (12)

Lipid Fraction)

**Race/Ethnicity**

Black, n (%) 59,007 (18.9%)

White, n (%) 227,817 (72.8%)

Hispanic, n (%) 25,747 (8.1%)

# Cardiometabolic Disease at Enrollment\*

Coronary Artery Disease, n (%) 67,912 (21.7%)

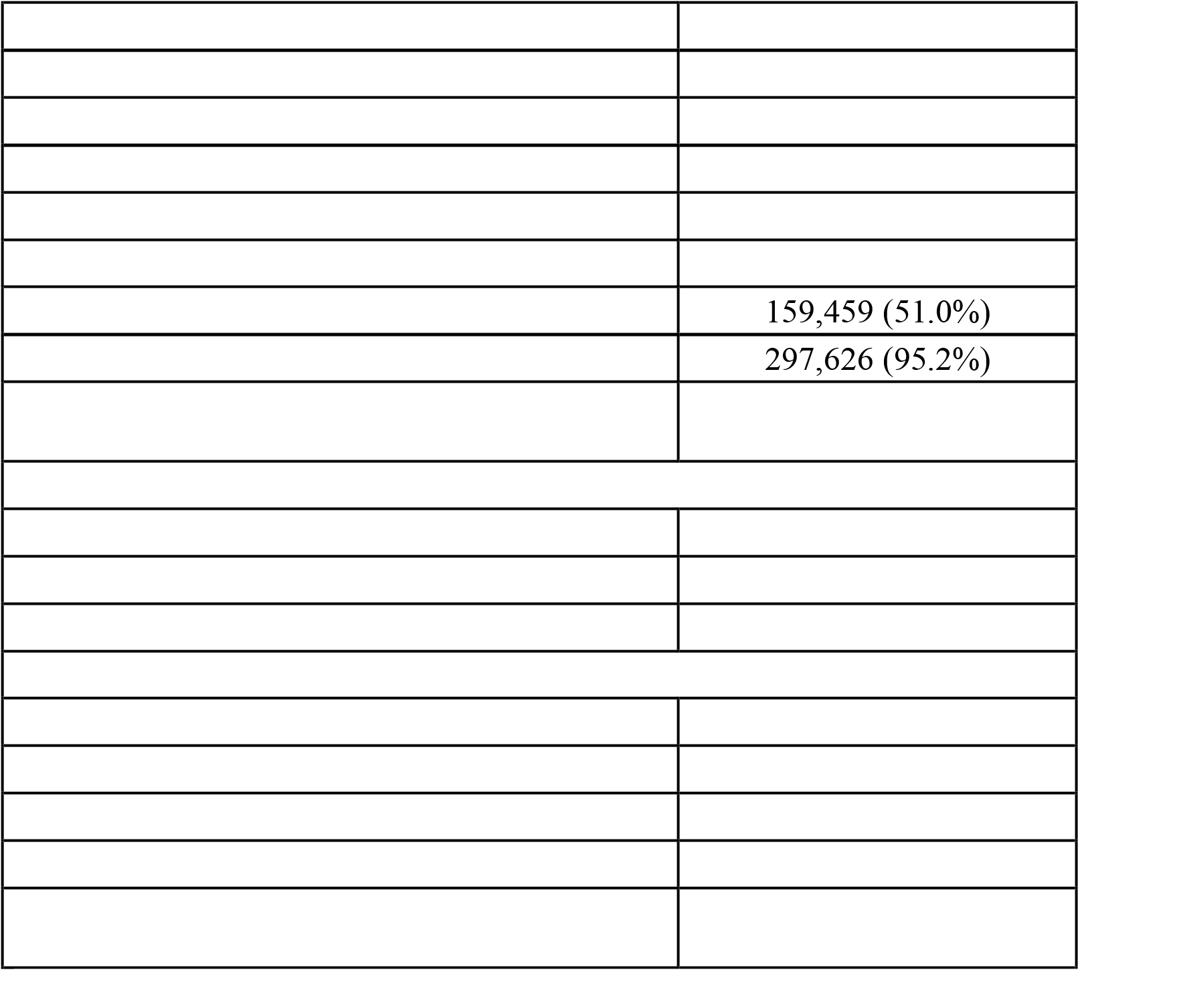
Type 2 Diabetes, n (%) 92,079 (29.5%)

Peripheral Artery Disease, n (%) 21,418 (6.9%)

Abdominal Aortic Aneurysm, n (%) 5,618 (1.8%)

Deep Venous Thrombosis or Pulmonary

7,009 (2.2%) Embolism, n (%)

\* Diseases are defined by *International Classification of Disease*, *Ninth Edition* (ICD-9) diagnosis codes.

Abbreviations: SD, Standard Deviation

# Table 2 - Demographic and clinical characteristics for 297,626 veterans in the Million Veteran Program lipids analysis

|  |  |  |  |
| --- | --- | --- | --- |
|  | **White** | **Black** | **Hispanic** |
| Veterans, N (%) | 215,551 (72.4%) | 57,332 (19.3%) | 24,743 (8.3%) |
| Age at Enrollment ± SD, years | 64.2 ± 13 | 57.7 ± 11.8 | 56.3 ± 15.0 |
| Male, n (%) | 200,900 (93.2%) | 50,059 (87.3%) | 22,601 (91.3%) |
| Body Mass Index ± SD, kg/m2 | 30.1 ± 5.9 | 30.4 ± 6.3 | 30.7 ± 5.8 |
| Statin Therapy Prescription at Enrollment, n (%) | 100,024 (46.4%) | 23,302 (40.6%) | 9,646 (39.0%) |
| Statin Therapy Prescription at time of Max LDL Blood Draw, n (%) | 18,818 (8.7%) | 5,024 (8.8%) | 2,262 (9.1%) |
| Statin Therapy Prescription at time of Max TC Blood Draw, n (%) | 18,433 (8.6%) | 5,027 (8.8%) | 2,162 (8.7%) |
| Mean Min HDL-C ± SD, mg/dL | 36.2 ± 11.4 | 38.9 ± 12.8 | 36.4 ± 11.0 |
| Mean Max LDL-C ± SD, mg/dL | 139 ± 38.4 | 142.2 ± 40.7 | 141.3 ± 38.1 |
| Median Max TG ± IQR, mg/dL | 211 ± 174 | 179 ± 149 | 221 ± 184 |
| Mean Max TC ± SD, mg/dL | 218.6 ± 46.7 | 220.8 ± 47.2 | 221.9 ± 48.0 |
| Variants Included in Analysis | 19,342,852 | 31,448,849 | 30,455,745 |

Abbreviations: Min, Minimum; Max, Maximum; SD, Standard Deviation; HDL-C, HighDensity Lipoprotein Cholesterol; LDL-C, Low-Density Lipoprotein Cholesterol; TG, Triglycerides; TC, Total Cholesterol

1. Corporal Michael Crescenz VA Medical Center, Philadelphia, PA [↑](#footnote-ref-1)
2. Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC),

   VA Boston Healthcare System, Boston MA, USA [↑](#footnote-ref-2)
3. Department of Epidemiology, Rollins School of Public Health, Department of

   Biomedical Informatics, School of Medicine, Emory University, Atlanta GA, USA [↑](#footnote-ref-3)
4. Regeneron Genetics Center. Tarrytown, NY 10591 [↑](#footnote-ref-4)
5. Department of Biostatistics, Boston University School of Public Health, Boston MA,

   USA [↑](#footnote-ref-5)
6. VA Salt Lake City Health Care System, Salt Lake City, UT, USA [↑](#footnote-ref-6)
7. Department of Medicine, University of Utah School of Medicine, Salt Lake City UT, USA [↑](#footnote-ref-7)
8. Department of Medicine, Stanford University School of Medicine, Stanford CA, USA [↑](#footnote-ref-8)
9. VA Palo Alto Health Care System, Palo Alto, CA, USA [↑](#footnote-ref-9)
10. Department of Genetics, Stanford University School of Medicine, Stanford CA, USA [↑](#footnote-ref-10)
11. University of Massachusetts College of Nursing & Health Sciences, Boston MA, USA [↑](#footnote-ref-11)