nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0222-9

# Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program

Derek Klarin[1,2,3,38], Scott M. Damrauer[4,5,38], Kelly Cho[6], Yan V. Sun[7], Tanya M. Teslovich[8], Jacqueline Honerlaw[6], David R. Gagnon[6,9], Scott L. DuVall[10,11], Jin Li[12,13], Gina M. Peloso[9], Mark Chaffin[2], Aeron M. Small[4,14], Jie Huang[6], Hua Tang[15], Julie A. Lynch[10,16], Yuk-Lam Ho[6], Dajiang J. Liu[17], Connor A. Emdin[1,2], Alexander H. Li[8], Jennifer E. Huffman[6], Jennifer S. Lee[12,13], Pradeep Natarajan[1,2,18], Rajiv Chowdhury[19], Danish Saleheen[4,20], Marijana Vujkovic[4,20], Aris Baras[8], Saiju Pyarajan[6,21], Emanuele Di Angelantonio[19], Benjamin M. Neale[2,22,23], Aliya Naheed[24], Amit V. Khera[1,2], John Danesh[19], Kyong-Mi Chang[4,25], Gonçalo Abecasis[26], Cristen Willer[27,28,29], Frederick E. Dewey[8], David J. Carey[30], Global Lipids Genetics Consortium[31], Myocardial Infarction Genetics (MIGen) Consortium[31], The Geisinger-Regeneron DiscovEHR Collaboration[31], The VA Million Veteran Program[31], John Concato[14,32], J. Michael Gaziano[6,21,33], Christopher J. O'Donnell[6,33,40], Philip S. Tsao[12,13,40], Sekar Kathiresan[1,2,40], Daniel J. Rader[25,33,34,35,36,40], Peter W. F. Wilson[37,38,40] and Themistocles L. Assimes[12,13,40]*

**The Million Veteran Program (MVP) was established in 2011 as a national research initiative to determine how genetic variation influences the health of US military veterans. Here we genotyped 312,571 MVP participants 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 one blood lipid measurement, including 57,332 black and 24,743 Hispanic participants, we tested up to around 32 million variants for association with 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 predicted to result in 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).**

Large-scale biobanks offer the potential to link genes to health traits documented in electronic health records (EHR) with unprecedented power[1]. 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 MVP was established in 2011 by the Veteran Affairs Office of Research and Development as a nationwide research program within the Veteran Affairs healthcare system[2]. The overarching goal of MVP is to find new biologic insights and clinical associations broadly relevant to human health and to enhance the care of veterans (former US military personnel) through precision medicine.

Blood concentrations of low-density lipoprotein cholesterol (LDL-C), triglycerides, total cholesterol and high-density lipoprotein cholesterol (HDL-C) are heritable risk factors for atherosclerotic cardiovascular disease[3], a highly prevalent condition among US veterans. Genome-wide association studies (GWAS) to date have identified at least 268 loci that influence these levels[4–12], many of which are under investigation as potential therapeutic targets[13,14]. However, off-target effects have dampened enthusiasm for some of these molecules[15,16]. Understanding the full spectrum of clinical consequences of a genetic variant through phenome-wide association studies (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 (stage 1), we performed association testing among 297,626 white (European ancestry), black (African ancestry) and Hispanic MVP participants with blood lipids stratified by ethnicity followed by a meta-analysis of results across all three groups. Replication of MVP findings was conducted in stages 2a or 2b with data from either one of two independent studies from the GLGC. Next, we leveraged the results of our discovery and meta-analysis to (1) estimate the variance explained by known and newly discovered lipid loci; (2) assess the potential of the use of multiple lipid measurements for discovery within MVP; (3) perform a transcriptome-wide association study (TWAS), a competitive gene-set pathway analysis and a tissue-expression analysis. We then focused on novel, genome-wide lipid-associated, low-frequency missense variants unique to our non-European populations as well as predicted loss of gene function (pLOF) mutations across all ethnic groups, as these associations have identified target pathways for pharmacological inactivation and modulation of

A full list of affiliations appears at the end of the paper.

cardiovascular risk[14,18,19]. Lastly, 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 International Classification of Disease, ninth edition (ICD-9) diagnosis codes in the Veteran Affairs EHR to better understand the potential consequences of pharmacological modulation of these genes or their products. We followed up significant findings from our PheWAS with multivariate Mendelian randomization analyses.

## Results

**Demographics of genotyped MVP participants.** A total of 353,323 veterans had genetic data available in MVP, with clinical phenotypes recorded in the Veteran Affairs EHR for 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 white, (2) non-Hispanic black and (3) Hispanic participants. Admixture plots depicting the genetic background of the black and Hispanic groups are shown in Supplementary Figs. 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 one laboratory measurement of blood lipids in their EHR. These individuals collectively had a total of 15,456,328 laboratory entries for blood lipids, or a median of 12 measurements 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-C, triglycerides and total cholesterol as well as his or her minimum HDL-C for genetic association analysis[20]. Table 2 summarizes characteristics at enrollment and the distribution lipid levels for MVP participants included in our analysis. As expected, most of the participants were male and 28% were of non-European ancestry. While approximately 45% of participants had evidence of a statin prescription at the time of enrollment, only 8–9% participants had such evidence at the time of their maximum LDL-C or total cholesterol measurement used for our GWAS analysis.

**Lipid genetic association and conditional analyses.** We successfully imputed (INFO > 0.3, minor allele frequency (MAF) > 0.0003) 19.3, 31.4, and 30.4 million variants for white, black and Hispanic veterans, respectively, using the 1000 Genomes Project[21] reference panel (Table 2). Black and Hispanic participants had substantially more variants available for analysis, reflecting the known greater genetic diversity within these populations[21,22]. We also identified 6,657 pLOF variants in 4,294 genes across the three ethnicities (Supplementary Fig. 3).

We compared the $Z$ scores and effect estimates from the published literature with those observed in MVP for 444 previously reported[11] exome-wide significant variants for lipids. We found a strong correlation of genetic associations across all four traits, validating the lipid data obtained from the EHR (Supplementary Figs. 4,5).

We performed association testing separately among individuals of each of three ancestries (white, black, and Hispanic) in our initial discovery analysis and then meta-analyzed results across ancestry groups using an inverse-variance-weighted fixed-effects method (Fig. 1a, Supplementary Fig. 6). Following trans-ethnic meta-analysis in the discovery phase of our study (stage 1), a total of 46,526 variants at 188 of the 268 known loci for lipids met the genome-wide significance threshold ($P < 5 \times 10^{-8}$) (Supplementary Tables 1–4). We performed pairwise comparisons of the allele frequencies and effect estimates between white and black participants as well as between white and Hispanic participants for 354 of the 444 previously established independent variants for lipids that were well-imputed in all three ancestral groups in MVP[11] (Fig. 2). We observed a much stronger correlation for effect allele frequencies between white and Hispanic participants (Pearson's correlation coefficient $R = 0.96$) than between white and black participants ($R = 0.72$), likely reflecting the greater European admixture in the MVP Hispanic participants. The effect estimates among the three ethnicities varied by lipid trait (Fig. 2, Supplementary Fig. 7).

We sought replication for variants within MVP with suggestive associations ($P < 1 \times 10^{-4}$) in either stages 2a or 2b (Fig. 1b). We first attempted replication of these variants using summary statistics from the 2017 GLGC exome array meta-analysis (stage 2a)[11]. If association statistics for promising DNA sequence variants from stage 1 were not available for replication in the 2017 exome array-focused study, we sought replication of these variants in publicly

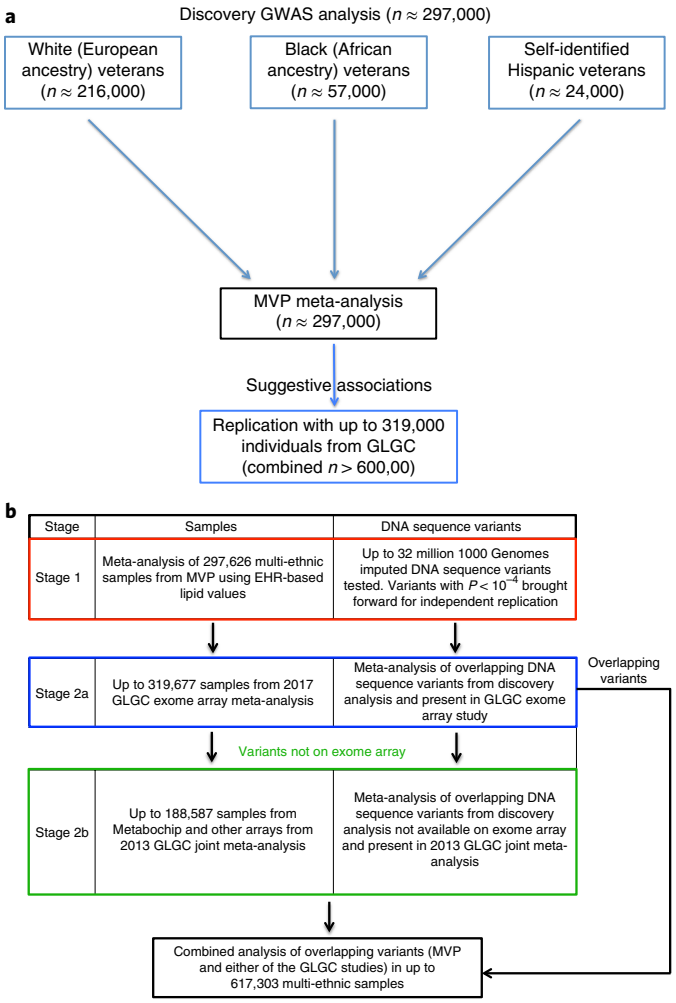


**Fig. 1 | GWAS study design. a**, DNA sequence variants across three separate ancestry groups in the MVP were meta-analyzed using an inverse-variance-weighted fixed-effects method in the discovery phase (stage 1). Variants with suggestive association were then brought forward for independent replication. **b**, DNA sequence variants with suggestive associations (two-sided linear regression $P < 1 \times 10^{-4}$) in the discovery analysis (stage 1) were brought forward for independent replication and tested using summary statistics from the 2017 exome-array-focused GLGC meta-analysis (stage 2a). Only variants with suggestive associations in stage 1 that were not present in the GLGC 2017 exome-array study (stage 2a) were alternatively replicated in the 2013 GLGC joint meta-analysis (stage 2b).

**Table 1 | Demographic and clinical characteristics of black, white and Hispanic individuals passing quality control in the MVP**

| Basic demographics | Genotyped veterans |
|---|---|
| *n* | 312,571 |
| Age at enrollment in years (mean ± σ) | 62.4 ± 13.5 |
| Male, *n* (%) | 287,441 (92.0%) |
| Body mass index in kg m$^{-2}$ (mean ± σ) | 30.3 ± 6.0 |
| Current smoker, *n* (%) | 59,385 (19.0%) |
| Former smoker, *n* (%) | 159,459 (51.0%) |
| *n* with ≥1 measurement of plasma lipids (%) | 297,626 (95.2%) |
| Number of lipid measurements (median per lipid fraction) | 15,456,328 (12) |
| **Race/ethnicity** | |
| Black, *n* (%) | 59,007 (18.9%) |
| White, *n* (%) | 227,817 (72.8%) |
| Hispanic, *n* (%) | 25,747 (8.1%) |
| **Cardiometabolic disease at enrollment**[a] | |
| 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 embolism, *n* (%) | 7,009 (2.2%) |

[a] Diseases are defined by ICD-9 diagnosis codes.

available summary statistics from the 2013 GLGC 'joint meta-analysis' (stage 2b). We did not attempt replication of any variant in both studies given the substantial overlap of participants in these two studies. 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 either stage 2a (GLGC 2017) or stage 2b (GLGC 2013). 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 $P < 5 \times 10^{-8}$ (genome-wide significance) in the discovery and replication cohorts combined. Following replication, 118 novel loci (from 141 lead variants) exceeded genome-wide significance ($P < 5 \times 10^{-8}$, Supplementary Tables 5–8). MAFs of lead variants ranged from 0.08% to 49.9%, with effect sizes ranging from 0.01 to 0.243 standard deviations (σ). For example, carriers of a rare missense mutation in the gene encoding sorting nexin 8 (*SNX8* Ile414Thr, (rs144787122, NC_000007.13: 2296552A>G) MAF = 0.35% in MVP) demonstrated a 0.10σ (3.8 mg dl$^{-1}$) higher plasma LDL-C after testing in 587,481 individuals.

More than one variant may independently affect plasma lipid levels at any given genetic locus. We performed a conditional analysis using combined summary statistics from MVP and publicly available data from GLGC for each lipid trait (Supplementary Fig. 8) and identified a total of 826 independently associated lipid variants across 118 novel and 268 previously identified loci (Supplementary Table 9).

**Variance explained obtained from multiple lipid measurements.** The previously mapped 444 lipid variants explain about 7.5–10.5% of the phenotypic variance in lipid levels in the MVP population. The 118 novel loci in our study explain an additional 0.38–0.74% in phenotypic variance, and the 826 independent variants identified in our conditional analysis increase the overall explained phenotypic variance to 8.8–12.3% (Supplementary Table 10).

We subsequently explored the impact of multiple lipid measurements in an analysis restricted to 171,314 European MVP participants with ≥5 lipid measurements in their EHR. We constructed a weighted genetic risk score (GRS) of 223 variants across 268 of the previously mapped loci with effect estimates available in the 2017 GLGC exome array analysis summary statistics[11] (Supplementary Table 11). Generally across the four lipid traits, the GRS explained a larger proportion of the phenotypic variance with an increasing number of lipid measurements included in the analysis (Supplementary Table 12). In addition, when the maximum/minimum lipid values were used as in our discovery GWAS, the GRS explained more total variance than when using up to five lipid measurements for the LDL-C, triglycerides and total cholesterol phenotypes.

**Transcriptome-wide association study.** We next performed a TWAS[23] using: (1) pre-computed weights from expression array data measured in peripheral blood from 1,245 unrelated control

**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 in years (mean ± σ) | 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 in kg m$^{-2}$ (mean ± σ) | 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 maximum LDL-C blood draw, *n* (%) | 18,818 (8.7%) | 5,024 (8.8%) | 2,262 (9.1%) |
| Statin therapy prescription at time of maximum total cholesterol blood draw, *n* (%) | 18,433 (8.6%) | 5,027 (8.8%) | 2,162 (8.7%) |
| Minimum HDL-C in mg dl$^{-1}$ (mean ± σ) | 36.2 ± 11.4 | 38.9 ± 12.8 | 36.4 ± 11.0 |
| Maximum LDL-C in mg dl$^{-1}$ (mean ± σ) | 139 ± 38.4 | 142.2 ± 40.7 | 141.3 ± 38.1 |
| Median maximum triglycerides ± IQR in mg dl$^{-1}$ | 211 ± 174 | 179 ± 149 | 221 ± 184 |
| Maximum total cholesterol in mg dl$^{-1}$ (mean ± σ) | 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 |

IQR, interquartile range.

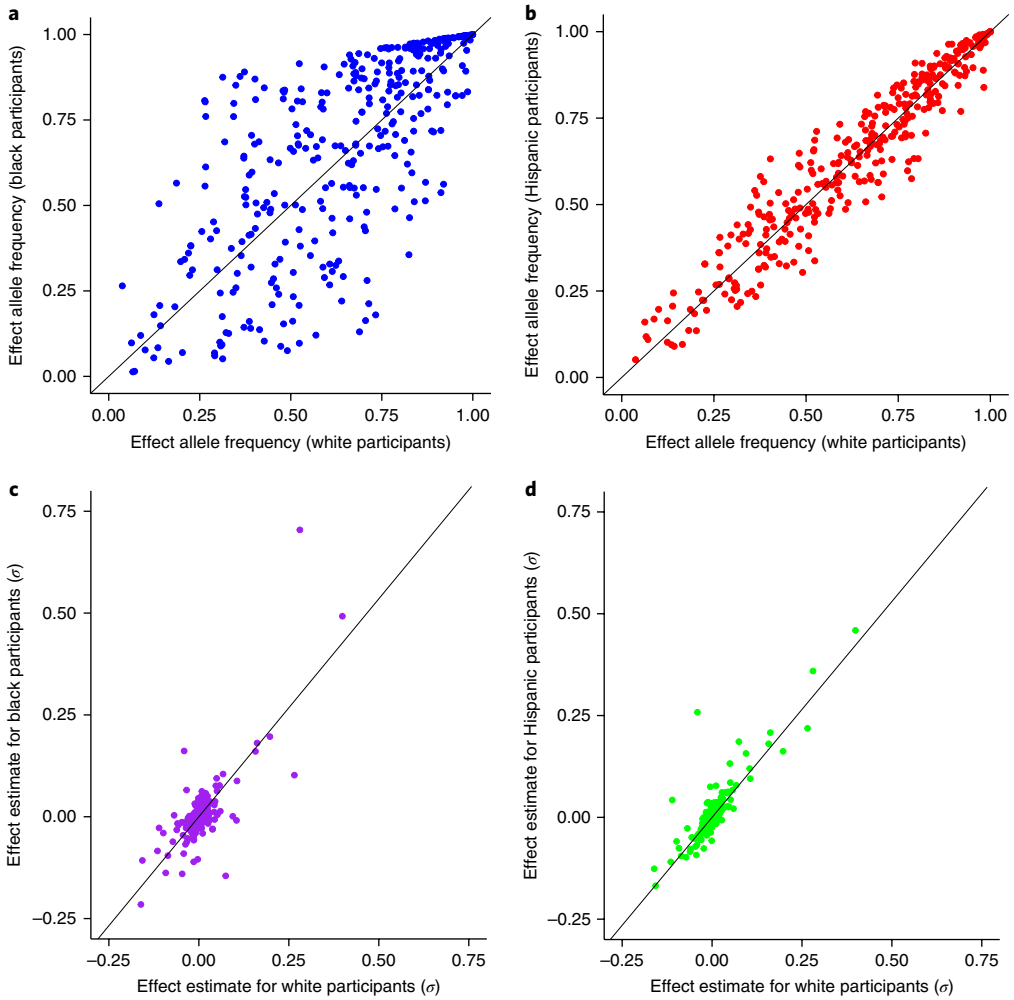


**Fig. 2 | Comparison of 354 independent lipid-associated variants across ethnicities. a**,**b**, Allele frequencies of lipid-associated variants observed in white individuals ($n = 215{,}196$; *x* axes) compared to black (**a**; $n = 57{,}280$; Pearson's $R = 0.72$), or Hispanic (**b**; $n = 24{,}742$; Pearson's $R = 0.96$) individuals. **c**,**d**, Linear regression effect estimates for LDL-C associations in white individuals ($n = 215{,}196$; *x* axes) compared to black (**c**; $n = 57{,}280$; $\beta = 1.07$) or Hispanic (**d**; $n = 24{,}742$; $\beta = 1.06$) individuals.

individuals from The Netherlands Twin Registry[24], RNA-sequencing data measured in adipose tissue from 563 control individuals from the Metabolic Syndrome in Men study[23] and RNA-sequencing data from post-mortem liver (97 individuals) and tibial artery (285 individuals) tissue from the Genotype-Tissue Expression project[25] (GTEx V6); and (2) combined MVP and GLGC summary statistics for each of the four lipid traits. In brief, this approach integrates information from expression reference panels (variant–expression correlation), GWAS summary statistics (variant–trait correlation), and linkage disequilibrium (LD) reference panels (variant–variant correlation) to assess the association between the *cis*-genetic component of expression and phenotype[23]. The results yield candidate causal genes from the GWAS results under the assumption that the causal mechanism of the tested genes involves changes in *cis*-expression.

Our TWAS identified a total of 655 genome-wide significant ($P < 5 \times 10^{-8}$) gene–lipid associations (summed across expression reference panels) in 333 distinct genes, including 194 that were significant in more than one tissue or lipid trait (Supplementary Tables 13–16, Supplementary Figs. 9–10). The 333 distinct genes fell within 122 genomic loci, 117 of which were within a lipid GWAS region (±1 Mb around a mapped sentinel GWAS variant) identified in either a prior analysis or in the current study. However, five genes identified by TWAS fell outside of previously mapped GWAS regions, representing potentially novel genomic loci for lipids (Supplementary Table 17). Previous work has suggested that future lipid GWAS with larger sample sizes will likely confirm the novel lipid loci identified by our TWAS[26]. Results from additional competitive gene-set pathway and tissue expression analyses are available in the Supplementary Note.

**Non-European low-frequency missense variant associations.** We next focused on ancestry specific low-frequency (MAF < 5%) missense variants, as these variants have been suggested to have a higher likelihood of causality[27,28]. We identified several novel low-frequency missense variants associated with one or more lipid levels at genome-wide significance that were specific to black or Hispanic participants. We found a total of five variants associated with LDL-C and/or total cholesterol among black individuals (Supplementary Table 18) and two associated with HDL-C and/or total cholesterol among Hispanic individuals (Supplementary Table 19) in *PCSK9*, *LDLR*, *APOB* and *ABCA1*. All ten associations were directionally consistent with the 2017 GLGC exome chip meta-analysis with nine reaching nominal significance ($P < 0.05$) among 17,009 black and 5,084 Hispanic individuals included in the GLGC study. In addition, the seven variants that we identified were either monomorphic or had a MAF < 0.0005 in the approximately 215,000 white veterans in MVP. Of note, we observed the low-frequency 443Thr allele in

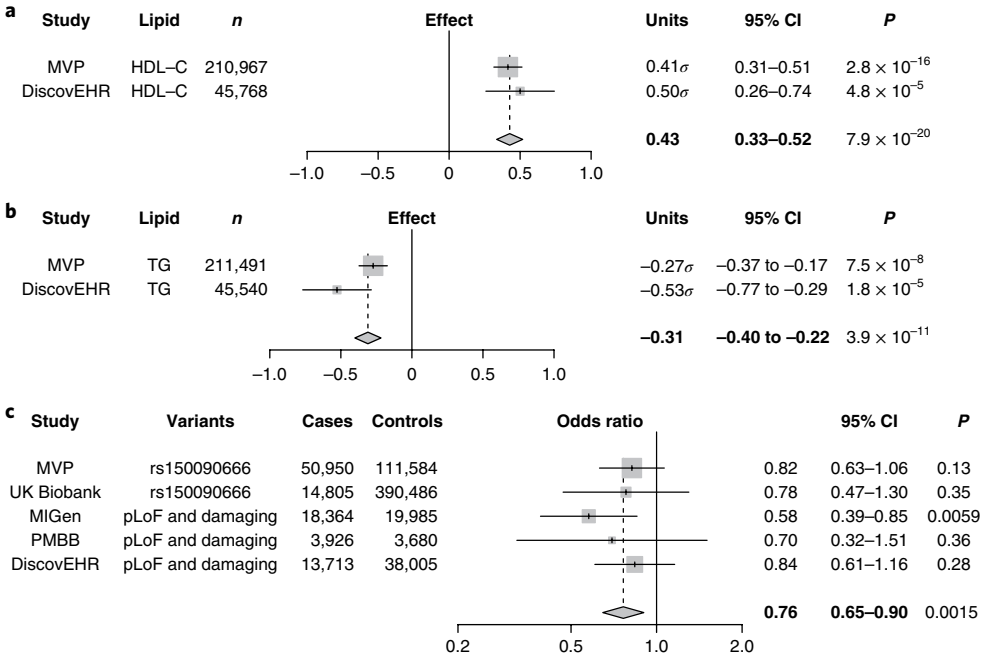


**Fig. 3 | *PDE3B* loss of gene function, lipids and coronary disease. a**,**b**, Linear regression results for the association of the pLOF mutation Arg783Ter in *PDE3B* with HDL-C (**a**) and triglycerides (TG) (**b**) for white veterans in MVP with independent replication in the DiscovEHR study. Two-sided *P* values are displayed. CI, confidence interval. **c**, Meta-analysis of the association of damaging *PDE3B* mutations and coronary artery disease across five studies, including three (MIGen, PMBB and DiscovEHR) with exome sequencing. Logistic regression results were pooled in an inverse-variance-weighted fixed-effects meta-analysis. Minimal evidence of heterogeneity across cohorts was observed ($I^2 = 0\%$). Two-sided *P* values are displayed.

*PCSK9* within Hispanic individuals to be eightfold more common in black individuals (MAF = 0.011 in Hispanic versus 0.092 in black individuals). We also found that this variant was associated with total cholesterol in black individuals at genome-wide significance.

**Predicted loss of gene function lipid associations.** We focused next on the subset of genotyped or imputed pLOF variants (variants that were annotated as premature stop (nonsense), canonical splice sites (splice-donor or splice-acceptor) or insertion/deletion variants that shifted frame (frameshift) by the Variant Effect Predictor software[29]). A total of 15 distinct pLOF variants demonstrated genome-wide significant lipid associations across individuals of all three ethnic groups (Supplementary Table 20). We replicated known pLOF associations at *PCSK9*[19], *APOC3*[18], *ANGPTL8*[8], *LPL*[30], *CD36*[31] and *HBB*[32], and we observed genome-wide significant associations of comparable magnitude of effect in each of the three ethnic groups for two pLOF variants: a base substitution in *APOC3* 55+1G>A and a mutation in *LPL* encoding Ser747Ter.

We identified one novel pLOF association. Among white MVP participants, carriers of a rare stop-gain mutation in *PDE3B* (encoding Arg783Ter; carrier frequency of 1 in 625), exhibited 4.72 mg dl$^{-1}$ ($0.41\sigma$) higher blood HDL-C levels ($P < 2.8 \times 10^{-16}$) and 43.3 mg dl$^{-1}$ ($-0.27\sigma$) lower blood triglyceride levels ($P = 7.5 \times 10^{-8}$). We found this signal to be independent of a previously reported genome-wide significant association in the region involving a common polymorphism, rs1037378[11] (Arg783Ter; conditional analysis $P = 6.3 \times 10^{-16}$ for HDL-C and $P = 8.91 \times 10^{-8}$ for triglycerides). We also identified one individual who was homozygous for Arg783Ter. This *PDE3B* 'human knockout' was in his sixth decade of life and had HDL-C and triglycerides levels of 73 and 56 mg dl$^{-1}$, respectively. He was not on lipid-lowering medication and was free of coronary artery disease (CAD). We replicated the triglyceride and HDL-C associations for this pLOF variant in an independent sample of approximately 45,000 participants of the DiscovEHR study (Fig. 3a,b).

**Loss of PDE3B function and risk of coronary artery disease.** Hypothesizing that mutations that were damaging or causing loss of function in *PDE3B* could protect against the development of CAD based on their association with lifelong lower levels of triglycerides in blood, we conducted a case-control study of CAD involving five cohorts: MVP, UK Biobank, Myocardial Infarction Genetics Consortium (MIGen), Penn Medicine Biobank (PMBB) and DiscovEHR. For three studies that underwent exome sequencing (MIGen, PMBB and DiscovEHR), we combined pLOF variants with missense variants that were predicted to be damaging or possibly damaging by each of five computer prediction algorithms (LRT score, MutationTaster, PolyPhen-2, HumDiv, PolyPhen-2 HumVar, and SIFT) as performed previously[30,33]. Because damaging mutations are individually rare, we aggregated them in subsequent association analysis with CAD (Supplementary Table 21). Among 103,580 individuals with CAD and 566,813 controls available for meta-analysis in these five cohorts, carriers of damaging *PDE3B* mutations were found to have a 24% decreased risk of CAD (odds ratio, 0.76; 95% confidence interval, 0.65–0.90; $P = 0.0015$; Fig. 3c). Data from an additional analysis examining the association of all novel lipid loci identified in our study with CAD are available in the Supplementary Note.

**PheWAS of 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-associated genes currently being targeted by pharmaceutical agents and identified functional variants in these genes: two nonsense variants (*LPL* Ser474Ter and *ANGPTL8* Gln121Ter) and three missense variants (*ANGPTL4* Glu40Lys, *APOA5* Ser19Trp, *PCSK9* Arg46Leu). We considered phenotypes to be significantly associated with a variant if they met a Bonferroni corrected $P < 4.98 \times 10^{-5}$ (0.05/1,004

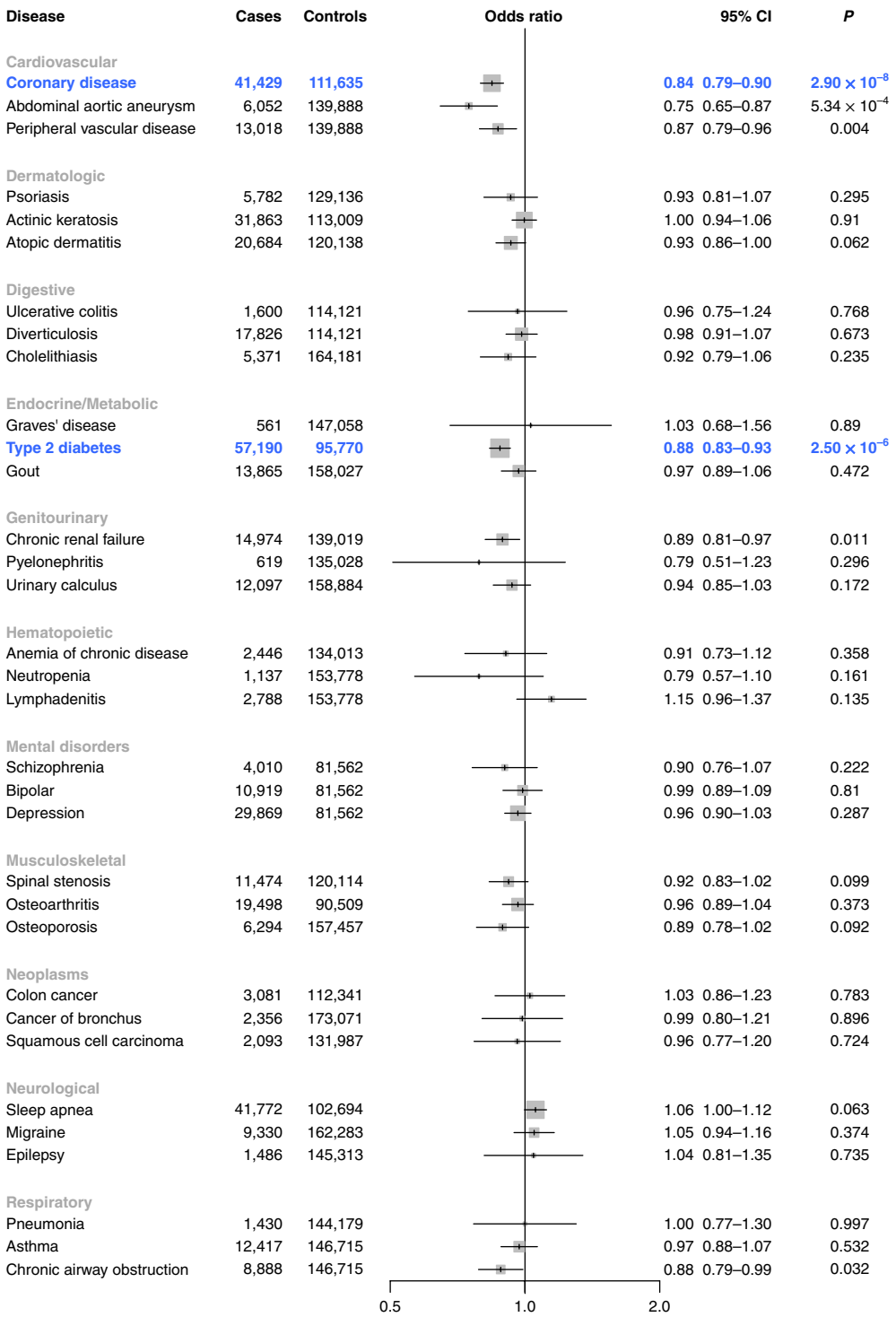


**Fig. 4 | *ANGPTL4* 40Lys carrier disease associations.** Forest plot for a representative 33 of the 1,004 disorders tested in the *ANGPTL4* Glu40Lys PheWAS. Statistically significant logistic regression associations are shown in blue. Two-sided *P* values are displayed.

traits), a conservative threshold given the correlation structure present among PheWAS phenotypes[34].

Data from 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 five variants, all of which are correlated with lipids (Supplementary Table 22). We replicated known associations with CAD for *LPL*[30], *ANGPTL4*[14] and *PCSK9*[19]. Notably, carriers of triglyceride-lowering and/or HDL-C-raising mutations in *ANGPTL4* (Glu40Lys; 7,013 carriers) were also found to have a reduced risk of type 2 diabetes (Fig. 4). We replicated the type 2 diabetes association for the *ANGPTL4* E40K variant in an independent sample of approximately 452,000 participants in the recently published trans-ethnic diabetes GWAS[35](odds ratio, 0.89; 95% confidence interval, 0.86–0.93; $P = 9.24 \times 10^{-10}$; Supplementary Fig. 11). In addition, carriers of LDL-C-lowering mutations in *PCSK9* (Arg46Leu; 5,537 carriers) also demonstrated a reduced risk of abdominal aortic aneurysm (AAA, Fig. 5).

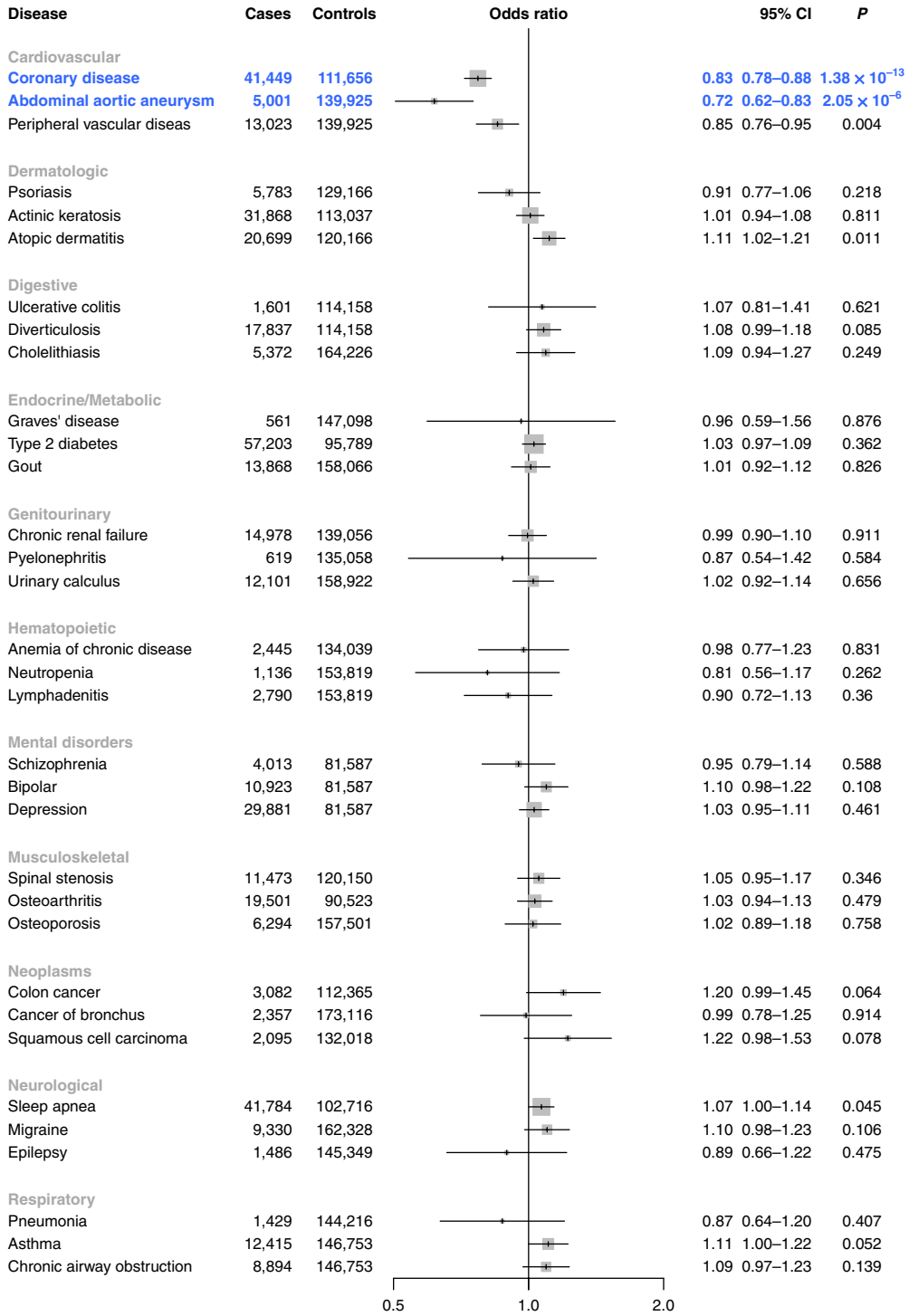


**Fig. 5 | *PCSK9* 46Leu carrier disease associations.** Forest plot for a representative 33 of the 1,004 disorders tested in the *PCSK9* Arg46Leu PheWAS. Statistically significant logistic regression associations are shown in blue. Two-sided *P* values are displayed.

**Lipids and AAA Mendelian randomization analysis.** To further explore the causal relationship of lipids on AAA development, we performed a multivariate Mendelian randomization analysis using a weighted GRS of 223 lipid-associated variants and summary data from a GWAS of 5,002 AAA cases and 139,968 controls in MVP. Consistent with our PheWAS results, a 1$\sigma$ genetically elevated LDL-C was associated with an increased risk of AAA (odds ratio, 1.47; 95% confidence interval, 1.28–1.68; $P=4.4\times10^{-8}$). Furthermore, a 1$\sigma$ genetically elevated HDL-C level was associated with a decreased risk of AAA (odds ratio, 0.79; 95% confidence interval, 0.68–0.91; $P=0.001$); and a 1$\sigma$ genetically elevated triglyceride level was associated with an increased risk of AAA (odds ratio, 1.40, 95% confidence interval, 1.18–1.66; $P=8.5\times10^{-5}$; Fig. 6). An MR-Egger analysis[36] indicated no pleiotropic bias of our lipid genetic instruments (MR-Egger intercept $P>0.05$ for all three lipid fractions (Supplementary Table 23)).

## Discussion

We leveraged clinical and genetic data from the MVP to investigate the inherited basis of blood lipids in nearly 300,000 US veterans. Our investigation resulted in several key findings. First, we robustly

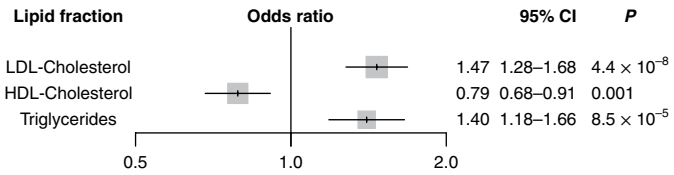


**Fig. 6 | Lipid associations with abdominal aortic aneurysm.** Logistic regression association results of the 223 variant lipid genetic risk score with abdominal aortic aneurysm in a multivariable Mendelian randomization analysis. Odds ratios are displayed per 1$\sigma$ genetically increased lipid fraction. Two-sided *P* values are displayed.

confirmed 188 previously identified loci while concurrently uncovering an additional 118 novel genome-wide significant loci. Next, we identified a total of 826 independent lipid-associated variants increasing the phenotypic variance explained by nearly 2%. We performed a TWAS in four tissues identifying five additional novel lipid loci at a genome-wide level of significance and performed a pathway analysis highlighting lipid transport mechanisms in our GWAS results. We identified ancestry-specific effects of rare coding variation on lipids among white, black and Hispanic participants, and observed 15 pLOF mutations associated with lipids at a genome-wide level of significance, including a protein-truncating variant in *PDE3B* that lowers triglycerides, raises HDL-C 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 obtained 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 Veteran Affairs' 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. Our results highlight the expected increase in variance explained by known loci when repeated lipid measurements are considered but also demonstrate the efficiency of examining the single most extreme lipid value least likely influenced by the use of lipid-altering medications. Subsequent meta-analysis (combined $n > 600{,}000$) with existing datasets increased the number of known independent genetic lipid loci to nearly 400, including several lipid pathways with links to human disease. For example, common variants near genes such as *COL4A2* and *ITGA1* identified for LDL-C and/or total cholesterol suggest links to extracellular matrix and cell adhesion biology, two pathways recently implicated by GWAS of CAD[37,38]. We also demonstrated that carriers of a rare missense mutation in the gene encoding perilipin 1 (*PLIN1* Leu90Pro) possess a markedly higher plasma HDL-C (0.243$\sigma$). In humans, perilipin 1 is required for lipid-droplet formation, triglyceride storage, as well as free fatty-acid metabolism, and frameshift pLOF mutations in the *PLIN1* gene have been reported to result in severe lipodystrophy[39]. A variant downstream of *BDNF* (encoding brain-derived neurotrophic factor) was found to be associated with HDL-C and triglycerides levels, supporting recent evidence linking this gene with metabolic syndrome and diabetes[40]. 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 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 groups and identified seven novel low-frequency coding variants associated with lipids only in non-European populations. Conversely, we also confirmed a shared genetic architecture across all three ethnic 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 disease[14,15,41,42]. Expansion of these efforts to larger sample sizes and additional ancestries may help to 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-platelet[43], vasodilatory[44] and inotropic[45] effects through inhibition of PDE3A, and also has well-documented, substantial effects on triglycerides and HDL-C levels[46]—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 the efficacy of cilostazol in intermittent claudication[46] and prevention of restenosis following percutaneous coronary intervention[47]. The drug is also currently used off-label for the prevention of stroke recurrence through a presumed anti-platelet effect[48]. We note that mice genetically deficient in *Pde3b* display reduced atherosclerosis[49] as well as decreased infarct size and improved cardiac function following experimental coronary artery ligation[50]. 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 PheWAS across a large-scale EHR-based biobank to predict both potentially adverse and 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 disease[13]. This finding is further supported by: our Mendelian randomization results; a recently published analysis using an independent AAA dataset[51]; and a recent report demonstrating that a *Pcsk9* gain-of-function mutation augments AAA development in a mouse model[52]. However, we also recognize the possibility that these results may be a consequence of pleiotropic effects induced by a high phenotypic correlation between AAA and the presence of advanced atherosclerotic disease. Thus, additional studies are necessary before definitive conclusions can be made on causality. Similarly, we expand on the potential indications for ANGPTL4 inhibition to include type 2 diabetes. Future PheWAS efforts may identify 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 to be mentioned. 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 Veteran Affairs 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 discovery analysis, we did not attempt to detect genetic associations specific to females or

heterogeneity of effects between sexes due to suspected limited power. Third, our TWAS identifies candidate causal genes under the assumption that the causal mechanism of the tested genes involves changes in *cis*-expression. However, we are unable to discriminate between instances of pleiotropy (when a given variant may alter gene expression and affect lipid levels independently) with TWAS alone and further functional analysis may be necessary. Fourth, our analysis demonstrating a lack of association between HDL-C raising alleles and CAD risk may be underpowered given the small number of examined alleles, although this finding has been demonstrated consistently in previous studies[53,54]. Lastly, power to detect associations for less common diseases in our PheWAS may also be limited despite the overall number of participants included in the analysis.

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

**URLs.** R statistical software, www.R-project.org; EasyQC, https://www.uni-regensburg.de/medizin/epidemiologie-praeventivmedizin/genetische-epidemiologie/software/; PheWAS, https://github.com/PheWAS/PheWAS; GCTA, http://cnsgenomics.com/software/gcta/#Overview; FUMA, http://fuma.ctglab.nl/; ExAC browser, http://exac.broadinstitute.org/; SNPTEST software program, http://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html; CARDIoGRAMplusC4D and MIGen and CARDIoGRAM Exome investigators datasets, http://www.cardiogramplusc4d.org.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0222-9.

Received: 6 February 2018; Accepted: 3 August 2018;
Published online: 01 October 2018

## References

1. Collins, R. What makes UK Biobank special? *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**, 214–223 (2016).
3. The Emerging Risk Factors Collaboration. Major lipids, apolipoproteins, and risk of vascular disease. *J. Am. Med. Assoc.* **302**, 1993–2000 (2009).
4. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
5. Global Lipids Genetics Consortium.. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (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**, 223–232 (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**, 823–838 (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.* **49**, 1758–1766 (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.* **49**, 1722–1730 (2017).
13. Sabatine, M. S. et al. Evolocumab and clinical outcomes in patients with cardiovascular disease. *N. Engl. J. Med.* **376**, 1713–1722 (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**, 1134–1144 (2016).
15. Dewey, F. E. et al. Inactivating variants in *ANGPTL4* and risk of coronary artery disease. *N. Engl. J. Med.* **374**, 1123–1133 (2016).
16. Barter, P. J. et al. Effects of torcetrapib in patients at high risk for coronary events. *N. Engl. J. Med.* **357**, 2109–2122 (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**, 1102–1111 (2013).
18. The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute.. 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**, 1264–1272 (2006).
20. Abul-Husn, N. S. et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* **354**, aaf7000 (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**, 1035–1044 (2009).
23. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
24. Wright, F. A. et al. Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).
25. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
26. Mancuso, N. et al. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).
27. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
28. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
29. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
30. Khera, A. V. et al. Association of rare and common variation in the lipoprotein lipase gene with coronary artery disease. *J. Am. Med. Assoc.* **317**, 937–946 (2017).
31. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
32. Sidore, C. et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* **47**, 1272–1281 (2015).
33. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
34. Diogo, D. et al. Phenome-wide association studies (PheWAS) across large "real-world data" population cohorts support drug target validation. Preprint at https://www.biorxiv.org/content/early/2017/11/13/218875 (2017).
35. Mahajan, A. et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
36. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
37. Klarin, D. et al. Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat. Genet.* **49**, 1392–1397 (2017).
38. 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).
39. Gandotra, S. et al. Perilipin deficiency and autosomal dominant partial lipodystrophy. *N. Engl. J. Med.* **364**, 740–748 (2011).
40. Rani, J. et al. T2DiACoD: a gene atlas of type 2 diabetes mellitus associated complex disorders. *Sci. Rep.* **7**, 6892 (2017).
41. Musunuru, K. et al. Exome sequencing, *ANGPTL3* mutations, and familial combined hypolipidemia. *N. Engl. J. Med.* **363**, 2220–2227 (2010).
42. Graham, M. J. et al. Cardiovascular and metabolic effects of *ANGPTL3* antisense oligonucleotides. *N. Engl. J. Med.* **377**, 222–232 (2017).
43. Zhang, W. & Colman, R. W. Thrombin regulates intracellular cyclic AMP concentration in human platelets through phosphorylation/activation of phosphodiesterase 3A. *Blood* **110**, 1475–1482 (2007).
44. Maass, P. G. et al. PDE3A mutations cause autosomal dominant hypertension with brachydactyly. *Nat. Genet.* **47**, 647–653 (2015).

45. Vandeput, F. et al. Selective regulation of cyclic nucleotide phosphodiesterase PDE3A isoforms. *Proc. Natl Acad. Sci. USA* **110**, 19778–19783 (2013).
46. Bedenis, R. et al. Cilostazol for intermittent claudication. *Cochrane Database Syst. Rev.* **10**, CD003748 (2014).
47. Tsuchikane, E. et al. Impact of cilostazol on restenosis after percutaneous coronary balloon angioplasty. *Circulation* **100**, 21–26 (1999).
48. Shinohara, Y. et al. Cilostazol for prevention of secondary stroke (CSPS 2): an aspirin-controlled, double-blind, randomised non-inferiority trial. *Lancet Neurol.* **9**, 959–968 (2010).
49. Ahmad, F. et al. Phosphodiesterase 3B (PDE3B) regulates NLRP3 inflammasome in adipose tissue. *Sci. Rep.* **6**, 28056 (2016).
50. Chung, Y. W. et al. Targeted disruption of PDE3B, but not PDE3A, protects murine heart from ischemia/reperfusion injury. *Proc. Natl Acad. Sci. USA* **112**, E2253–E2262 (2015).
51. Harrison, S. C. et al. Genetic association of lipids and lipid drug targets with abdominal aortic aneurysm: a meta-analysis. *JAMA Cardiol.* **3**, 26–33 (2018).
52. Lu, H. et al. Hypercholesterolemia induced by a PCSK9 gain-of-function mutation augments angiotensin II-induced abdominal aortic aneurysms in C57BL/6 mice—brief report. *Arterioscler. Thromb. Vasc. Biol.* **36**, 1753–1757 (2016).
53. Voight, B. F. et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**, 572–580 (2012).
54. Do, R. et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45**, 1345–1352 (2013).

## Acknowledgements

Data on patients with coronary artery disease and myocardial infarctions have been contributed by the CARDIoGRAMplusC4D investigators and the Myocardial Infarction Genetics and CARDIoGRAM Exome investigators. Both datasets were obtained online (see URLs). This research is based on data from the MVP, Office of Research and Development, Veterans Health Administration, and was supported by the Department of Veterans Affairs Cooperative Studies Program award G002. This research was also supported by three additional Department of Veterans Affairs awards (1I0101BX003340, 1I01BX003362, and 1I01CX001025) and the NIH (T32 HL007734, K01HL125751, R01HL127564). The content of this manuscript does not represent the views of the Department of Veterans Affairs or the United States Government.

## 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.F.W., J.C. and J.M.G. Acquisition, analysis or interpretation of data: D.K., S.M.D., Y.V.S., K.C., T.M.T., J.Ho., D.R.G., S.L.D., J.L., G.M.P., M.C., A.M.S., J.Hu., H.T., J.S.L., Y.-L.H., D.J.L., C.A.E., A.H.L., J.A.L., R.C., P.N., D.S., M.V., A.B., S.P., E.D.A., B.M.N., A.N., A.V.K., J.D., K.-M.C., G.A., C.W., F.E.D., J.E.H. and D.J.C. Drafting of the manuscript: D.K. and T.L.A. Critical revision of the manuscript for important intellectual content: S.M.D., Y.V.S., K.C., P.N., C.W., J.A.L., F.E.D., S.L.D., K.-M.C., C.J.O., P.S.T., S.K., D.J.R. and P.W.W. Administrative, technical or material support: D.K., Y.V.S., K.C., J.Ho., D.R.G., S.L.D., J.A.L., Y.H., J.C., J.M.G., C.J.O., P.S.T, J.E.H., and P.W.W.

## Competing interests

S.K. 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. T.M.T., A.H.L., A.B., F.E.D. and D.J.C. are employees of Regeneron Pharmaceuticals. G.A. has received consulting income from Regeneron Genetics Center, 23andMe and Helix. S.L.D. has received research grant support from the following for-profit companies through the University of Utah or the Western Institute for Biomedical Research (VA Salt Lake City's affiliated non-profit): AbbVie Inc., Anolinx LLC, Astellas Pharma Inc., AstraZeneca Pharmaceuticals LP, Boehringer Ingelheim International GmbH, Celgene Corporation, Eli Lilly and Company, Genentech Inc., Genomic Health Inc., Gilead Sciences Inc., GlaxoSmithKline PLC, Innocrin Pharmaceuticals Inc., Janssen Pharmaceuticals Inc., Kantar Health, Myriad Genetic Laboratories Inc., Novartis International AG and PAREXEL International Corporation.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41588-018-0222-9.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Correspondence and requests for materials** should be addressed to T.L.A.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional 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. [3]Boston VA Healthcare System, Boston, MA, USA. [4]Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, USA. [5]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [6]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA, USA. [7]Department of Epidemiology, Rollins School of Public Health, Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, USA. [8]Regeneron Genetics Center, Tarrytown, NY, USA. [9]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [10]VA Salt Lake City Health Care System, Salt Lake City, UT, USA. [11]Department of Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA. [12]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [13]VA Palo Alto Health Care System, Palo Alto, CA, USA. [14]Department of Medicine, Yale School of Medicine, New Haven, CT, USA. [15]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [16]University of Massachusetts College of Nursing and Health Sciences, Boston, MA, USA. [17]Department of Public Health Sciences, Institute of Personalized Medicine, Penn State College of Medicine, Hershey, PA, USA. [18]Cardiovascular Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [19]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [20]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [21]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [22]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [23]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [24]Initiative for Noncommunicable Diseases, Health Systems and Population Studies Division, International Centre for Diarrheal Disease Research, Dhaka, Bangladesh. [25]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [26]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [27]Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI, USA. [28]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [29]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [30]Geisinger Health System, Danville, PA, USA. [31]A list of members and affiliations appears in the Supplementary Note.[32]Clinical Epidemiology Research Center, VA Connecticut Healthcare System, West Haven, CT, USA. [33]Department of Medicine, Harvard Medical School, Boston, MA, USA. [34]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [35]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [36]Cardiovascular Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [37]Atlanta VA Medical Center, Decatur, GA, USA. [38]Emory Clinical Cardiovascular Research Institute, Atlanta, GA, USA. [39]These authors contributed equally: Derek Klarin, Scott M. Damrauer. [40]These authors jointly supervised: Christopher J. O'Donnell, Philip S. Tsao, Sekar Kathiresan, Daniel J. Rader, Peter W. F. Wilson, Themistocles L. Assimes. *e-mail: tassimes@stanford.edu

## Methods

The design of the MVP has been previously described[2]. In brief, individuals aged 19–104 years have been recruited from more than 50 Veterans Affairs Medical Centers nationwide since 2011. Each veteran's EHR data are being integrated into the MVP biorepository, including inpatient ICD-9 diagnosis codes, Current Procedural Terminology procedure codes, clinical laboratory measurements and reports of diagnostic imaging modalities. The MVP received ethical and study protocol approval from the Veteran Affair Central Institutional Review Board in accordance with the principles outlined in the Declaration of Helsinki. Informed consent was obtained from all participants of the MVP study.

**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 importance in different ethnic backgrounds. Veterans of three mutually exclusive ethnic groups were identified for analysis: (1) non-Hispanic white veterans (European ancestry), (2) non-Hispanic black veterans (African ancestry) and (3) Hispanic veterans. Further details on the methods used to assign ancestry and perform sample quality control are described in the Supplementary Note.

**Variant quality control.** Prior to imputation, variants that were poorly called (genotype missingness > 5%) or that deviated from their expected allele frequency based on reference data from the 1000 Genomes Project[21] were excluded. After pre-phasing using EAGLE[55] v.2, genotypes from the 1000 Genomes Project[21] phase 3, v.5 reference panel were imputed into MVP participants via Minimac3 software[56]. Ethnicity-specific principal component analysis was performed using the EIGENSOFT software[57].

Following imputation, variant-level quality control was performed using the EasyQC R package[58] (see URLs), and the following exclusion metrics were used: ancestry-specific Hardy–Weinberg equilibrium[59] $P < 1 \times 10^{-20}$, posterior call probability < 0.9, imputation quality/INFO < 0.3, MAF < 0.0003, call rate < 97.5% for common variants (MAF > 1%) and call rate < 99% for rare variants (MAF < 1%). Variants were also excluded if they deviated by >10% from their expected allele frequency based on reference data from the 1000 Genomes Project[21].

**EHR-based lipid phenotypes.** EHR clinical laboratory data were available for MVP participants from as early as 2003. We extracted the maximum LDL-C, triglycerides and total cholesterol values and minimum HDL-C values for each participant for analysis. These extreme values were selected to approximate plasma lipid concentrations in the absence of lipid-lowering therapy as described previously[20]. For each phenotype (LDL-C, natural log-transformed triglycerides, HDL-C and total cholesterol), residuals were obtained after regressing on age, age$^2$, 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 the 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 on 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 (stage 1), 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-analysis[11] (stage 2a) or the 2013 GLGC joint meta-analysis[5] (stage 2b). Replication was first attempted using summary statistics from the 2017 GLGC exome array study (stage 2a). 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 (stage 2b). 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 in either stage 2a or stage 2b. We emphasize that if a variant was available for replication in both studies, replication was performed only 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 $P < 5 \times 10^{-8}$ (genome-wide significance) in the discovery and replication cohorts combined. Novel loci were defined as being greater than 1 Mb away from a known genome-wide-associated lead variant for lipids. Additionally, LD information from the 1000 Genomes Project[21] was used to determine independent variants for which a locus extended beyond 1 Mb. All association *P* values were two-sided. Further details on the association analysis are described in the Supplementary Note.

**Conditional analysis.** We used the COJO-GCTA software (see URLs) to perform an approximate, stepwise conditional analysis to identify independent variants within lipid-associated loci given that individual level data for the prior GLGC lipid analyses are not publicly available. We used summary statistics of ~1.9 million overlapping variants that we meta-analyzed across either one of the two GLGC datasets (predominantly European) and the European MVP dataset to conduct this analysis (Supplementary Fig. 8) combined with an LD matrix obtained from 10,000 unrelated European individuals randomly sampled from the UK Biobank interim release.

**Variance explained using multiple lipid measurements.** We estimated the proportion of variance explained by the set of 444 previously mapped independent lipid variants, the 118 novel lipid loci identified in our study, and the 826 independent lipid variants identified from conditional analysis using ridge regression with the glmnet R package. The variance explained was determined after tuning the hyperparameter (lambda) to approximate an optimal value, and then calculating the model $R^2$ after performing linear regression with the inverse-normal-transformed lipid outcome and each set (444, 118, 826) of independent genome-wide lipid variants as predictors.

We estimated the variance explained for a GRS of 223 previously described GWAS lipid variants weighted by their previously reported effect sizes[11] (Supplementary Table 11) as a function of the number of lipid measurements in MVP to assess the potential impact of using multiple lipid measurements in discovery. We performed this analysis using the mean of one, two, three, four and five lipid measurements for each individual starting with their measurement closest to enrollment and moving towards the past. To account for the use of statin therapy, individuals with evidence of a statin prescription in their EHR at the time of enrollment had their LDL-C and total cholesterol values adjusted by dividing by 0.7 and 0.8, respectively, as previously described[5]. In addition, we also calculated the variance explained by the single maximum triglycerides, LDL-C and/or total cholesterol levels and minimum HDL-C levels from the EHR without adjustment for lipid-lowering therapy. Our analyses were restricted to a subset of 171,314 European MVP participants with ≥5 lipid measurements.

**Lipid TWAS.** We performed a TWAS using summary statistics after a meta-analysis of ~1.9 million overlapping variants among GLGC (predominantly European) and European MVP datasets (Supplementary Fig. 8) and four gene-expression reference panels (whole blood from The Netherlands Twin Registry, adipose tissue from the Metabolic Syndrome in Men study, and tibial artery and liver from GTEx) in independent samples as previously described[23]. In brief, for a given gene, variant-expression weights in the 1-Mb *cis* locus were first computed with BSLMM[60], which "models effects on expression as a mixture of normal distributions to account for the sparse expression architecture. Given weights *w*, lipid *Z* scores *Z*, and variant-correlation (LD) matrix *D*; the association between predicted expression and lipids (that is, the TWAS statistic) was estimated as $Z_{TWAS} = w'Z/(w'Dw)^{1/2}$" (details have been previously described[23]). We computed TWAS statistics by using either the variants genotyped in each expression reference panel or imputed HapMap3 variants. To account for multiple hypotheses, we applied a genome-wide significant *P* value threshold (two-sided $P < 5 \times 10^{-8}$), considerably more stringent than previously used Bonferroni corrections in prior TWAS[26]. We defined novel TWAS loci as a TWAS gene falling outside of a previously identified lipid GWAS region (±1 Mb around a mapped sentinel GWAS variant).

**Identification of independent low-frequency coding variant lipid associations specific to blacks and hispanics.** We used the *P* value and LD-driven clumping procedure in PLINK version 1.90b (--clump) to identify associations between low-frequency coding variants and lipids specific to black and Hispanic individuals. Input-included summary lipid association statistics from our MVP 1000 Genomes imputed genome-wide association study of black and Hispanic individuals, and reference LD panels of 661 African and 347 admixed American samples from 1000 Genomes phase 3 whole-genome sequencing data. Variants were clumped with stringent $r^2$ (<0.01) and *P* ($<5 \times 10^{-8}$) thresholds in a 1-Mb 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 black and Hispanic individuals at a MAF < 0.05.

**Loss of gene function analysis.** We used the Variant Effect Predictor[29] software to identify pLOF DNA sequence variants defined as: premature stop (nonsense), canonical splice-sites (splice-donor or splice-acceptor) or insertion/deletion variants that shifted frame (frameshift). For the pLOF lipids analysis, we then merged these variants with data from the Exome Aggregation Consortium[27] (v.0.3.1, see URLs), 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 $P < 5 \times 10^{-8}$ (genome-wide significance).

**Loss of *PDE3B* gene function and CAD.** We identified a novel lipid association for a pLOF mutation in the *PDE3B* gene (rs150090666, Arg783Ter). For carriers of damaging mutations in phosphodiesterase 3B, we examined the effects of the mutation on risk for CAD using logistic regression in five separate cohorts: MVP, UK Biobank and three cohorts with exome sequencing: the MIGen, the 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 five computer prediction algorithms (LRT score, MutationTaster, PolyPhen-2, HumDiv, PolyPhen-2 HumVar and SIFT) as performed previously[30,33]. 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 on participating cohorts and CAD case definitions are described in the Supplementary Note. We set a two-sided $P < 0.05$ 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* Ser474Ter and *ANGPTL8* Gln121Ter) and three missense variants (*ANGPTL4* Glu40Lys, *APOA5* Ser19Trp, *PCSK9* Arg46Leu). Details on 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 two-sided $P < 4.98 \times 10^{-5}$ (0.05/1,004 traits). For replication of our *ANGPTL4* Glu40Lys type 2 diabetes finding, we combined the PheWAS results with publicly available data from the recently published trans-ethnic type 2 diabetes GWAS[35] using an inverse-variance-weighted fixed-effects method.

**Lipids and AAA Mendelian randomization analysis.** Summary-level data for 223 genome-wide lipid-associated variants were obtained from publicly available data from the Global Lipids Genetics Consortium[11]. We then utilized results from a GWAS of 5,002 AAA cases and 139,968 controls performed in white MVP participants using the previously proposed definition[17]. The effect alleles were matched with all lipid and AAA summary data and three different Mendelian randomization analyses were performed: (1) inverse-variance-weighted; (2) multivariable; and (3) MR-Egger to account for pleiotropic bias. First, we performed inverse-variance-weighted Mendelian randomization using each set of variants for each lipid trait as instrumental variables. This method, however, does not account for possible pleiotropic bias. Therefore, we next performed inverse-variance-weighted multivariable Mendelian randomization. This method adjusts for possible pleiotropic effects across the included lipid traits in our analyses using effect estimates from the variant–AAA outcome and effect estimates from variant–LDL-C, variant–HDL-C and variant–triglycerides as predictors in one multivariable model. We additionally performed MR-Egger as previously described[36]. This technique can be used to detect bias secondary to unbalanced pleiotropy in Mendelian randomization studies. In contrast to inverse-variance-weighted analysis, the regression line is unconstrained, and the intercept represents the average pleiotropic effects across all variants. Bonferroni-corrected two-sided *P* values ($P = 0.016$ (0.05/3)) for three tests were used to declare statistical significance.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The full summary-level association data from the trans-ancestry meta-analysis for each lipid trait from this report are available through dbGaP, with accession number phs001672.v1.p1.

## References

55. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
56. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
57. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
58. Winkler, T. W. et al. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
59. Hyde, C. L. et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* **48**, 1031–1036 (2016).
60. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).

nature research

Corresponding author(s): Themistocles Assimes

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☐ | ☒ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |
| ☐ | ☒ | Clearly defined error bars<br>*State explicitly what error bars represent (e.g. SD, SE, CI)* |

*Our web collection on statistics for biologists may be useful.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Phenotypic data was collected from the electronic health record and genetic data using the Million Veteran Program (MVP) Axiom array. All data was collated using R-3.2 as documented in the URLs section |
| Data analysis | Data was collected using the EasyQC package (exemplar code link documented in the URLs section), and SNPTEST software program as outlined in the supplementary methods (exemplar code link documented in the URLs section) |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data is to be posted online

# Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | All samples available of three ancestries (European, African, Hispanic)were used for analysis (after quality control). Sample size was determined based on genetic data available from MVP. |
| Data exclusions | Data were excluded if they did not pass our QC metrics, or if they did not fall within the three main ancestries used for analysis |
| Replication | Replication was performed using data from one of 2 sources: 1) GLGC 2017 exome chip GWAS summary statistics, or 2) GLGC 2013 joint-meta-analysis GWAS summary statistics |
| Randomization | N/A |
| Blinding | N/A |

# Reporting for specific materials, systems and methods

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Unique biological materials |
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | 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. |
| Recruitment | Individuals aged 19 to 104 years have been recruited voluntarily from more than 50 VA Medical Centers nationwide for participation in the Million Veteran Program biobank study. |