Assessment of multi-population polygenic risk scores for lipid traits in African Americans

Domenica E. Drouet[1], Shiying Liu[2] and Dana C. Crawford[2,3,4]

[1] Department of Medicine, Case Western Reserve University, Cleveland, OH, United States of America
[2] Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, United States of America
[3] Cleveland Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, United States of America
[4] Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH, United States of America

Corresponding author
Dana C. Crawford,
dana.crawford@case.edu

ABSTRACT

Polygenic risk scores (PRS) based on genome-wide discoveries are promising predictors or classifiers of disease development, severity, and/or progression for common clinical outcomes. A major limitation of most risk scores is the paucity of genome-wide discoveries in diverse populations, prompting an emphasis to generate these needed data for trans-population and population-specific PRS construction. Given diverse genome-wide discoveries are just now being completed, there has been little opportunity for PRS to be evaluated in diverse populations independent from the discovery efforts. To fill this gap, we leverage here summary data from a recent genome-wide discovery study of lipid traits (HDL-C, LDL-C, triglycerides, and total cholesterol) conducted in diverse populations represented by African Americans, Hispanics, Asians, Native Hawaiians, Native Americans, and others by the Population Architecture using Genomics and Epidemiology (PAGE) Study. We constructed lipid trait PRS using PAGE Study published genetic variants and weights in an independent African American adult patient population linked to de-identified electronic health records and genotypes from the Illumina Metabochip ($n = 3{,}254$). Using multi-population lipid trait PRS, we assessed levels of association for their respective lipid traits, clinical outcomes (cardiovascular disease and type 2 diabetes), and common clinical labs. While none of the multi-population PRS were strongly associated with the tested trait or outcome, PRS_{LDL-C} was nominally associated with cardiovascular disease. These data demonstrate the complexity in applying PRS to real-world clinical data even when data from multiple populations are available.

INTRODUCTION

Polygenic risk scores (PRS) are generally defined as the sum of an individual's additive genetic risk for an outcome or disease of interest (*Cooke Bailey & Igo Jr, 2016*; *Igo Jr, Kinzy & Cooke Bailey, 2019*; *Osterman, Kinzy & Cooke Bailey, 2021*). Early genetic risk scores for cardiovascular disease (*Kathiresan et al., 2008*), type 2 diabetes (*Hivert et al., 2011*), and

age-related macular degeneration (*Osterman, Kinzy & Cooke Bailey, 2021*), to name a few, suggest genome-wide association study (GWAS)-identified variants in aggregate have the potential for patient risk reclassification or identification prior to development of disease or progression to severe disease, spurring interest in the development of clinical applications of these basic research findings (*Torkamani, Wineiger & Topol, 2018*; *Christoffersen & Tybjærg-Hansen, 2021*). Despite the promising preliminary data and the now 15 years' worth of genome-wide discoveries, PRS are still largely limited to or based on GWAS data from European-descent populations. Recent initiatives that encourage genomic discovery in diverse populations will enable the assessment of multi-population PRS for outcomes or traits as population-specific and shared genetic variants emerge from the most recent GWAS stemming from these efforts (*Hindorff et al., 2018*).

The Population Architecture using Genomics and Epidemiology (PAGE) Study, supported by the National Human Genome Research Institute (NHGRI), was first established to generalize GWAS findings in diverse populations (*Dumitrescu et al., 2011*; *Matise et al., 2011*; *Carlson et al., 2013*). The PAGE Study has since expanded to include discovery of genetic associations in diverse populations for common outcomes and traits (*Wojcik et al., 2019*). In one such recent PAGE Study, array-based GWAS for high density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TG), and total cholesterol (TC), common clinical biomarkers ordered to assess risk for the development of a variety of outcomes including cardiovascular disease (CVD) and type 2 diabetes (T2D), was conducted in > 45,000 ancestrally diverse participants (*Hu et al., 2020*). A total of nine novel lipid loci were identified. Further, comparisons of effect sizes across populations revealed smaller genetic effects in African Americans and Hispanic compared with the same genetic associations identified in European descent participants (*Hu et al., 2020*).

The genetic associations identified for lipid traits in the PAGE Study provide a still unique opportunity to calculate and apply multi-population lipid trait PRS for diverse populations. To seize this opportunity, we accessed clinical data from electronic health records (EHRs) linked to a biorepository with genetic data for almost 12,000 African American patients (*Crawford et al., 2015*). Here, we calculate multi-population PRS for lipids based on the PAGE Study discovery efforts and apply them to real world clinical data in an independent African American patient sample. We anticipated that the availability of genetic association data from diverse populations that include African Americans would result in robust PRS. The resulting data highlight many of the challenges that need to be addressed before PRS research can be translated into clinical practice (*Li et al., 2020*).

MATERIALS & METHODS

Study population

We accessed an existing and previously described dataset (*Crawford et al., 2015*) of de-identified EHRs linked to biospecimens genotyped on the Illumina Metabochip (*Buyske et al., 2012*; *Voight et al., 2012*). These data are a subset of the larger BioVU, Vanderbilt University Medical Center's (VUMC) biobank established in 2007 (*Roden et al., 2008*).
BioVU links discarded biospecimens collected for clinical purposes during outpatient visits to a de-identified version of the patient's EHR. The de-identification process includes the scrubbing of identifiers, data shifting, and linkage to biospecimens *via* a one-way hash that combined make re-identification difficult and unlikely (*Roden et al., 2008*). BioVU as a VUMC program has received approval from the Vanderbilt Institutional Review Board (IRB) and was reviewed in detail by the federal Office for Human Research Protections (OHRP), who agreed with the non-human subjects regulatory designation for both the resource and subsequent research (*Roden et al., 2008*).

The data accessed for the current study were collected under an opt-out participation model described to the patient annually as part of the consent to treat process and represent DNA samples extracted from blood collected between 2007 and 2011 and clinical data collected through 2011. As a study site of the PAGE Study, a consortium focused on genetic association studies in diverse populations (*Matise et al., 2011*), we selected at the time all DNA samples in BioVU from non-European descent patients for Illumina Metabochip genotyping ($n = 15,863$), the majority of which are from African American or Black patients (73%; *Crawford et al., 2015*). As we have previously described (*Crawford et al., 2015*), BioVU is representative of the Nashville, Tennessee VUMC patient population, a population that is on average older and of European descent compared with the surrounding Davidson County population. The dataset described here is de-identified and is considered to be non-human subjects research (Vanderbilt University IRB #110579) (*Pulley et al., 2010*).

Electronic phenotyping

Analyses were limited to African American adults aged 18 years or older. All demographic and clinical data were extracted using a combination of structured (*e.g.*, International Classification of Diseases 9th edition, Clinical Modification (ICD-9-CM) codes, Current Procedural Terminology (CPT) codes, problems lists, and laboratory values) and unstructured (clinical notes) data available in the EHR. Race/ethnicity in BioVU was administratively assigned, and these assignments are highly correlated with genetic ancestry for some (*e.g.*, European and African Americans) but not all (*e.g.*, Hispanics and East Asians) groups (*Dumitrescu et al., 2010*; *Hall et al., 2014*; *Farber-Eger et al., 2017*). Age was calculated as the age of the patient in 2011 based on the patient's year of birth. First mention (*e.g.*, the first clinic visit associated with a lipid lab result) of each patient's serum HDL-C, LDL-C, TG, and TC was extracted and recorded as either a "pre-medication" or "post-medication" value. Pre-medication values were defined as free of evidence of concurrent lipid lowering medication usage. As previously described (*Crawford et al., 2015*), lipid lowering medications included statins (also known as HMG CoA reductase inhibitors, atorvastatin (Lipitor®), fluvastatin (Lescol®), lovastatin (Mevacor®, Altoprev™), pravastatin (Pravachol®), rosuvastatin calcium (Crestor®), simvastatin (Zocor®), lovastatin + niacin (Advicor®), atorvastatin + amlodipine (Caduet®), and simvastatin + ezetimibe (Vytorin™)), selective cholesterol absorption inhibitors (ezetimibe (Zetia®)), resins (cholestyramine (Questran®, Questran® Light, Prevalite®, Locholest®, Locholest® Light), colestipol (Colestid®), colesevelam Hcl

(WelChol®)), fibrates (gemfibrozil (Lopid®), fenofibrate (Antara®, Lofibra®, Tricor®, and Triglide™), clofibrate (Atromid-S)), and niacin. Fasting status is unknown. Heights and weights closest to the first mention of the lipid trait were extracted to calculate body mass index (BMI; kg/m^2). Prevalent cardiovascular disease was defined as mention of heart attack keywords in the problems list ("MI", "myocardial infarction", or "heart attack"), CPT codes (*Crawford et al., 2018*) for coronary artery bypass graft (33510–33514; 33515; 33516; 33517–33519; 33520; 33521–33523; and 33534–33536) and angioplasty and/or coronary stents (92980–92981; 92982; 92984; 92995; and 92996), or three mentions of International Classification of Diseases (ICD)-9-CM codes (410 or 410.*) (*Dumitrescu et al., 2015a*) for myocardial infarction. Type 2 diabetes cases status was defined using a combination of ICD codes, labs, and medications from the previously described and validated electronic Medical Records & Genomics (eMERGE) Network algorithm (*Kho et al., 2012*).

Genotyping and quality control

The Illumina Metabochip was genotyped by Vanderbilt Technologies for Advanced Genomics (VANTAGE), formerly the Vanderbilt University Center for Human Genetics Research DNA Resources Core, for the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study, a study site of PAGE I (*Matise et al., 2011*). Genotyping was conducted following the manufacturer's protocol (Illumina, Inc.; San Diego, CA.). Per PAGE I protocol, 360 HapMap samples were also genotyped for quality control (*Crawford et al., 2012*). Genotype calling and standard quality control measures (call rates, duplicate checks, Hardy-Weinberg Equilibrium, *etc.*) have been described by Buyske et al. 2011 and more recently by *Kaur et al. (2021)*. Metabochip genotypes linked to the EAGLE BioVU de-identified clinical data are available through the database of Genotypes and Phenotypes (dbGaP) accession number phs002767.v1.p1.

Polygenic risk score calculations

Unweighted and weighted multi-population PRS for HDL-C, LDL-C, TG, and TC were calculated for each patient based on published summary statistics for diverse populations from the PAGE Study (*Hu et al., 2020*). We first extracted Metabochip genotypes available for SNPs previously associated HDL-C (44), LDL-C (36), TG (51), TC (48) reported by the PAGE Study at genome-wide significance ($p \le 5 \times 10^{-8}$) (Table S7 from *Hu et al., 2020*; Supplementary Table). After removing SNPs out of Hardy Weinberg Equilibrium at $p < 10^{-4}$, we calculated unweighted PRS for each patient and each lipid trait by counting the number of risk alleles (N) for 42, 34, 50, and 46 HDL-C, LDL-C, TG, and TC-associated SNPs (k), respectively. PRS were weighted (PRS$_w$) using the absolute value of the betas ($|\beta|$) for the corresponding associations published by the PAGE Study (*Hu et al., 2020*).

$$\text{PRS}_w = \sum_{i=1}^{k} |\beta_i| N_i.$$

Statistical methods

Principal components (PCs) were generated as previously described (*Buyske et al., 2012*; *Kaur et al., 2021*). We tested each SNP for an association with the lipid trait previously associated as described in *Hu et al. (2020)*. Single SNP tests of association were performed in PLINK version 1.90 using linear regression, unadjusted and adjusted for sex, age, BMI, and the first ten PCs. TG levels were transformed (natural log) prior to tests of association. Summary statistics were visualized using Synthesis View (*Pendergrass et al., 2010b*; *Pendergrass et al., 2010a*; Supplementary File). We also tested for associations between the lipid traits and their respective unweighted and weighted PRS using linear regression, unadjusted and adjusted for age, sex, BMI, and the first ten PCs. Tests of association between CVD and T2D and unweighted and weighted PRS were performed using 2x2 tables (unadjusted) and logistic regression (CVD and LDL-C weighted PRS adjusted by LDL-C). A limited phenome-wide association study (PheWAS) was performed for unweighted and weighted PRS and platelet counts, glucose, serum creatinine, serum albumin, serum albumin creatinine ratio, blood urea nitrogen (BUN), uric acid, urine creatinine, urine albumin, and estimated glomerular filtration rate (eGFR). eGFR was calculated using EHR-extracted serum creatinine, age closest to serum creatinine, sex, and race and the CKD-EPI equation (*Levey et al., 2009*). eGFR was then categorized as stage 1 ($\geq$90 mL/min per 1.73 m^2), stage 2 (60–89 mL/min per 1.73 m^2), and stage 3 (<60 mL/min per 1.73 m^2) for the limited PheWAS. When necessary, laboratory values were transformed as the natural log of the quantitative value plus one per the PAGE I study PheWAS protocol (*Pendergrass et al., 2011*). All PheWAS tests of association were performed using linear regression except for eGFR where multinomial logistic regression was used to test for association using stage 1 as the referent. All statistical analyses were performed using R version 4.1.0 (*R Core Team, 2021*) and PLINK 1.90.

RESULTS

Study population

A total of 11,521 African American patients were genotyped on the Illumina Metabochip as part of EAGLE BioVU (*Crawford et al., 2015*). Among these patients, 3,254 were adults with height, weight, and Metabochip genotypes (Table 1). Many (61.16%) were female with an average age of 46.95 years ($\pm$15.32 years standard deviation) and an average BMI (31.68 kg/m^2; $\pm$8.52 kg/m^2 standard deviation) considered obese. Average pre-medication and post-medication lipid lab values were within the range expected for a general US adult African American population (*Cohen et al., 2010*; *Carroll et al., 2012*). Each patient had at least one risk allele for each lipid trait (Fig. S1). On average, African American adult patients had a higher TG PRS (5.95) compared with TC (3.52), HDL-C (3.39) and LDL-C (2.01) (Table 1).

Lipid trait genetic associations

We first performed unadjusted (Fig. S2) and adjusted single SNP tests of association for each lipid trait, stratified by lipid lowering medication exposure as described in **Materials & Methods**. Among African Americans with pre-medication lipid labs, four out of 42

Table 1 Study population characteristics for African American adults with at least one lipid lab extracted from the electronic health record. Study population characteristics are based on African American adult (18 years or older) patients who had a mention HDL-C, LDL-C, TG, or TC in the electronic health record (EHR) and who were also genotyped using the Illumina Metabochip. First mention of each lipid lab was extracted and recorded as either a "pre-medication" (free of evidence of concurrent lipid lowering medication) or "post-medication" value as described in Methods and Materials. Body mass index was calculated using the closest height and weight recorded in the EHR compared to the clinic date of the lipid lab. Weighted polygenic risk scores (PRSw) were calculated as described in Methods and Materials.

Variable	Mean (± standard deviation) or %
Female	61.16%
Age	46.95 (± 15.32) years
Body mass index	31.68 (± 8.52) kg/m^2
HDL-C	Pre-medication ($n = 1{,}464$): 53.90 (± 18.16) mg/dL Post-medication ($n = 428$): 50.34 (± 19.39) mg/dL
LDL-C	Pre-medication ($n = 1{,}464$): 105.23 (± 40.72) mg/dL Post-medication ($n = 433$): 101.82 (± 49.98) mg/dL
Triglycerides	Pre-medication ($n = 1{,}535$): 115.51 (± 81.13) mg/dL Post-medication ($n = 441$): 134.15 (± 79.56) mg/dL
ln(Triglycerides)	Pre-medication ($n = 1{,}535$): 4.56 (± 0.60) mg/dL Post-medication ($n = 441$): 4.76 (± 0.53) mg/dL
Total Cholesterol	Pre-medication ($n = 1{,}640$): 182.20 (± 48.72) mg/dL Post-medication ($n = 464$): 177.88 (± 58.10) mg/dL
HDL-C PRSw	3.39 (± 0.61)
LDL-C PRSw	2.01 (± 0.45)
TG PRSw	5.95 (± 0.51)
TC PRSw	3.52 (± 0.43)

HDL-C, three out of 34 LDL-C, five out of 50 TG, and eight out of 46 TC SNPs were associated with their respective lipid trait at $p < 0.05$ adjusted for sex, age, BMI, and PCs (Figs. 1–4). Compared with the literature and accounting for coded alleles (*Hu et al., 2020*), all associations identified here at $p < 0.05$ for LDL-C were in the expected directions of effect. Most of the associations identified here for HDL-C were in the opposite direction
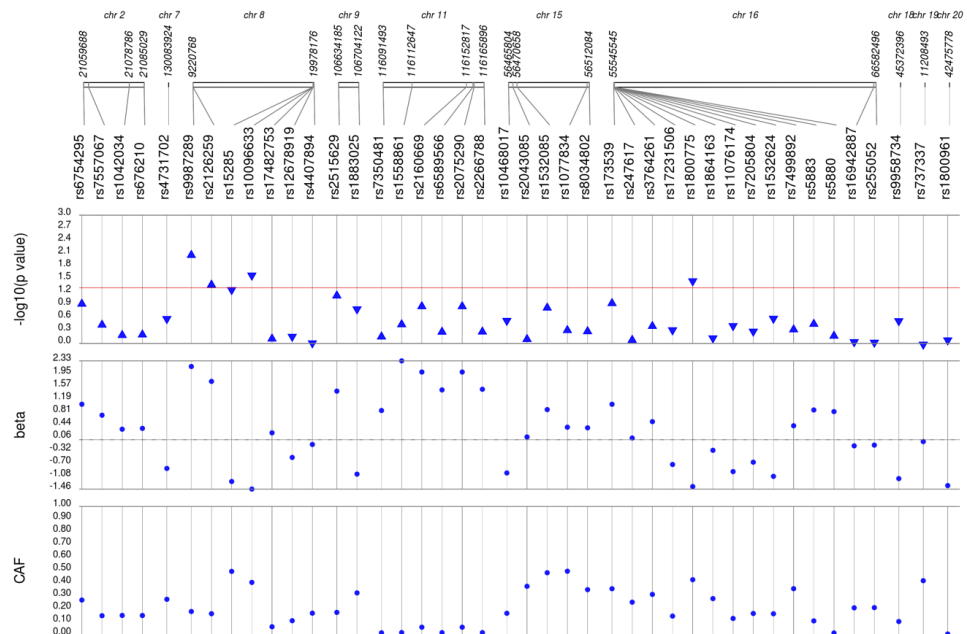
Figure 1 **Results of adjusted single SNP tests of associations among African American adults for HDL-C levels, pre-medication.** Lipid labs were extracted from EHRs and represent the first mention of the laboratory value free of evidence of concurrent lipid lowering medication usage ("pre-medication"). Each SNP was tested for an association with each pre-medication lipid lab using linear regression assuming an additive genetic model adjusted for age, sex, body mass index, and the first 10 principal components. SNP genomic location is given on the x-axis, and p-values ($-\log_{10}$ transformed) are plotted along the y-axis using Synthesis View. The direction of the arrows corresponds to the direction of the beta-coefficient. The significance threshold is indicated by the red line at $p = 0.05$. Also plotted are the betas and the coded allele frequencies (CAF).

compared with *Hu et al. (2020)*, with only rs1800775 associated in the expected direction. For transformed triglycerides (rs1077834) and total cholesterol (rs8106922) each, all but one of the associated SNPs were in the expected direction of effect.

Among African Americans with post-medication lipid labs, two out of 42 HDL-C, four out of 34 LDL-C, and one out of 49 TG SNPs were associated with their respective lipid trait at $p < 0.05$ adjusted for sex, age, BMI, and PCs (Fig. S3). None of the SNPs tested were associated with post-medication TC at $p < 0.05$. Overall, fewer associations at $p < 0.05$ were identified among post-medication lipid labs (eight) compared with pre-medication lipid labs (20), most likely due to differences in sample size (Table 1). Despite the smaller sample sizes, three associations were identified among African Americans with post-medication labs but not among African Americans with pre-medication labs: rs1864163 (HDL-C), rs688 (LDL-C), and rs4407894 (transformed TG) (Figs. 1–4 *versus* Fig. S3). Of these three associations, rs1864163 was associated in the opposite direction compared with *Hu et al. (2020)*. Among the five overlapping pre- and post-medication associations, all were consistent in direction of effects with the exception of rs1800775, which was associated with decreased HDL-C levels among African Americans with pre-medication labs but with increased levels among African Americans with post-medication labs.
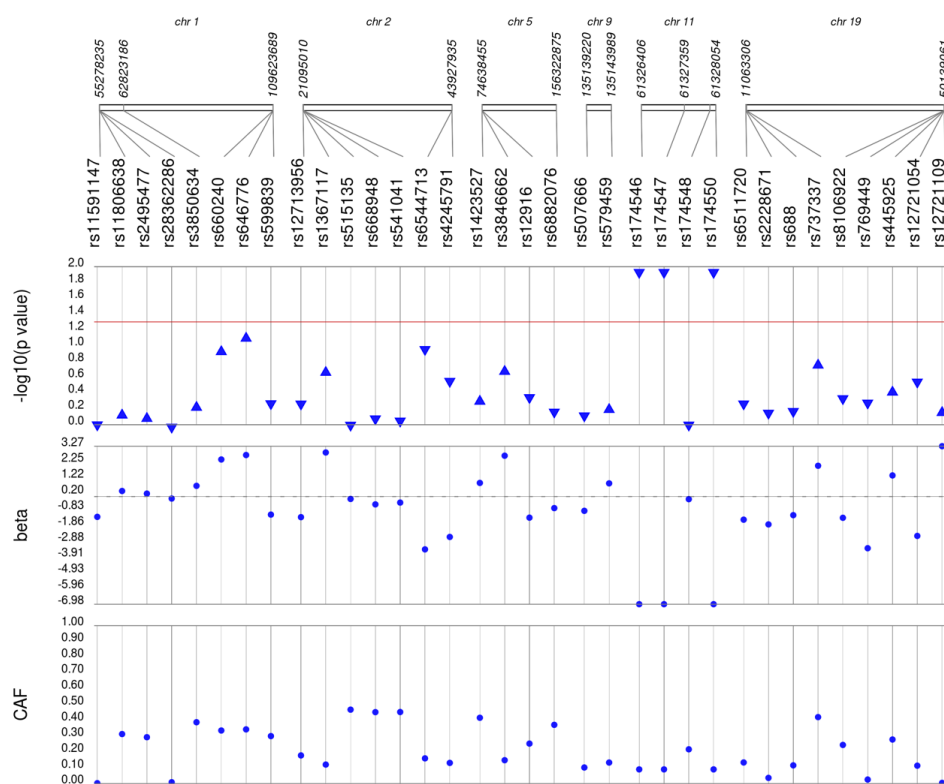
Figure 2 Results of adjusted single SNP tests of associations among African American adults for LDL-C levels, pre-medication. Lipid labs were extracted from EHRs and represent the first mention of the laboratory value free of evidence of concurrent lipid lowering medication usage ("pre-medication"). Each SNP was tested for an association with each pre-medication lipid lab using linear regression assuming an additive genetic model adjusted for age, sex, body mass index, and the first 10 principal components. SNP genomic location is given on the x-axis, and p-values ($-\log_{10}$ transformed) are plotted along the y-axis using Synthesis View. The direction of the arrows corresponds to the direction of the beta-coefficient. The significance threshold is indicated by the red line at $p = 0.05$. Also plotted are the betas and the coded allele frequencies (CAF).

Lipid PRS, CVD, and T2D

We calculated unweighted and weighted PRS for each lipid trait using genomic discovery data specific for diverse populations as described in **Materials & Methods**. Among unadjusted and adjusted tests of association, none were associated with their respective lipid traits at $p < 0.05$ (Table 2). All PRS tested here, regardless of significance, were associated with decreased lipid lab values.

To explore the potential clinical utility of lipid PRS, we tested each lipid weighted PRS for association with CVD and T2DM (Table 3). In this African American patient population, 29.5% had one or more CVD event(s) recorded in their EHR (MI (1.7%), heart attack (29.4%). CABG (1.0%), CSA (1.5%)). Similarly, almost one-third of African American adult patients in the present study had T2D (28.4%). We did not detect significant associations between any of the lipid weighted PRS and T2D at $p < 0.05$ (Table 3). For LDL-C PRS, CVD cases had a higher proportion of extreme LDL-C weighted PRS (10.9%)
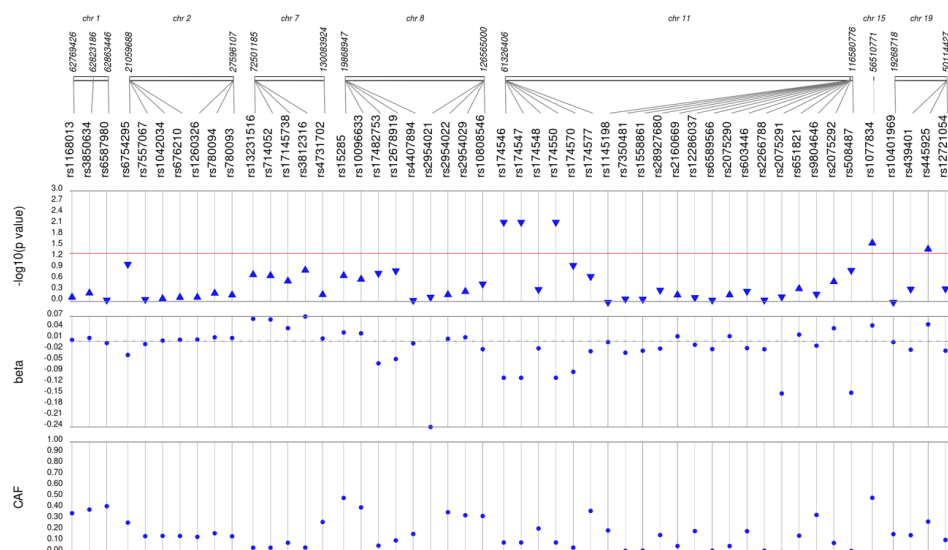
Figure 3 Results of adjusted single SNP tests of associations among African American adults for triglyceride levels, pre-medication. Lipid labs were extracted from EHRs and represent the first mention of the laboratory value free of evidence of concurrent lipid lowering medication usage ("pre-medication"). Each SNP was tested for an association with each pre-medication lipid lab using linear regression assuming an additive genetic model adjusted for age, sex, body mass index, and the first 10 principal components. Triglyceride levels were transformed (natural log) prior to tests of association. SNP genomic location is given on the x-axis, and p-values ($-\log_{10}$ transformed) are plotted along the y-axis using Synthesis View. The direction of the arrows corresponds to the direction of the beta-coefficient. The significance threshold is indicated by the red line at $p = 0.05$. Also plotted are the betas and the coded allele frequencies (CAF).

compared with controls (9.6%) (OR = 1.15, 95% Cl [1.01–1.32]; $p = 0.04$). The association was no longer nominally significant when adjusted for LDL-C levels ($p > 0.05$).

Limited PheWAS

To identify possible pleiotropic relationships between genetic determinants of lipid traits and other commonly ordered clinical labs, we performed a limited PheWAS for each lipid PRS and EHR-extracted labs related to biochemistry, liver, and kidney functions. Nominal associations ($p < 0.05$) were identified for HDL-C PRS and albumin creatinine ratio; serum creatinine and LDL-C and TC PRS; and stage 2 eGFR and TC PRS (Table 4). However, after adjusting significance thresholds for multiple testing (four lipid trait PRS ×10 labs), none of the limited PheWAS associations remained significant.

DISCUSSION

PRS have the potential to identify patients at-risk for developing health outcomes early in the disease process (*Torkamani, Wineiger & Topol, 2018*; *Kullo et al., 2022*). Given genomic discovery for common human diseases remains biased by genetic ancestry (*Sirugo, Williams & Tishkoff, 2019*; *Peterson et al., 2019*), there is much interest in assessing the generalizability of current PRS in diverse populations (*Wang et al., 2022*). In this study
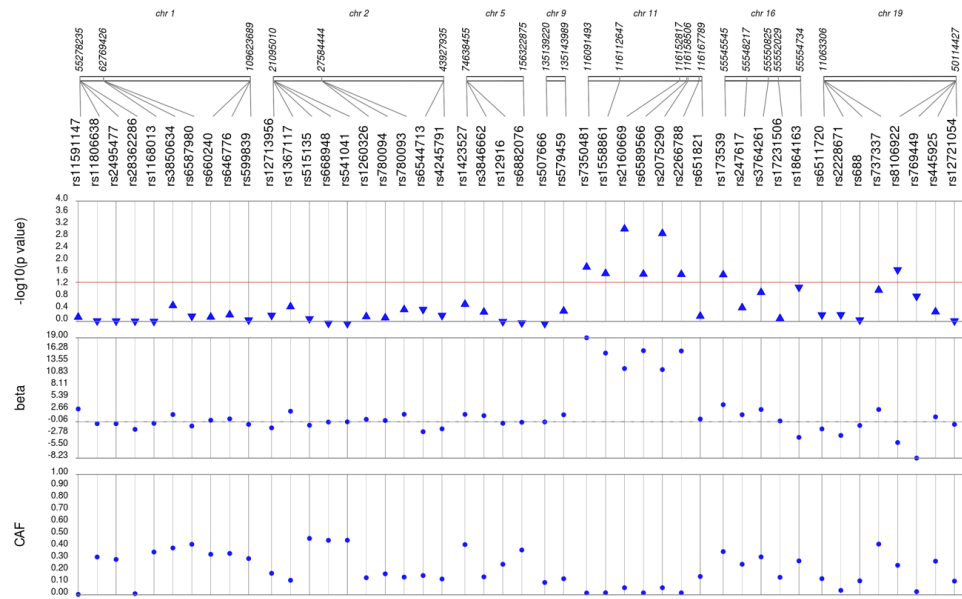
Figure 4 **Results of adjusted single SNP tests of associations among African American adults for total cholesterol levels, pre-medication.** Lipid labs were extracted from EHRs and represent the first mention of the laboratory value free of evidence of concurrent lipid lowering medication usage ("pre-medication"). Each SNP was tested for an association with each pre-medication lipid lab using linear regression assuming an additive genetic model adjusted for age, sex, body mass index, and the first 10 principal components. SNP genomic location is given on the *x*-axis, and *p*-values (−log₁₀ transformed) are plotted along the *y*-axis using Synthesis View. The direction of the arrows corresponds to the direction of the beta-coefficient. The significance threshold is indicated by the red line at $p = 0.05$. Also plotted are the betas and the coded allele frequencies (CAF).

Table 2 **Results of weighted and unweighted polygenic risk score test of associations among African American adults for HDL-C, LDL-C, TG, and TC levels, pre-medication.** Both weighted and unweighted polygenic risk scores (PRS) were calculated as described in Materials & Methods for all patients with a lipid lab value. Lipid labs were extracted from EHRs and represent the first mention of the laboratory value free of evidence of concurrent lipid lowering medication usage ("pre-medication"). Tests of association between PRS and their respective lipid labs were performed using linear regression unadjusted (A), adjusted for sex, age, and body mass index (BMI) (B), and adjusted linear regression for sex, age, BMI, and principal components (C). TG levels were transformed (natural log) prior to tests of association.

PRS	HDL-C		LDL-C		TG		TC	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
A. Unadjusted Linear Regression								
Weighted	−0.0029(0.7626)	0.9970	−3.7776(2.3805)	0.1128	−0.0311 (0.0293)	0.2887	−2.1158 (2.8190)	0.4530
Unweighted	−0.0453 (0.1033)	0.6615	−0.4448(0.2754)	0.1065	−0.0051 (0.0034)	0.1277	−0.1399 ((0.2663))	0.5994
B. Adjusted Linear Regression								
Weighted	−0.3625(0.7247)	0.6169	−3.7223(2.3595)	0.1149	−0.0306 (0.0280)	0.2747	−1.9797(2.7868)	0.4776
Unweighted	−0.0994(0.0982)	0.3118	−0.4472(0.2731)	0.1017	−0.0050 (0.0032)	0.1199	−0.1284 (0.2632)	0.6256
C. Adjusted Linear Regression + Principal Components								
Weighted	−0.3102(0.7293)	0.6707	−3.0489 (2.4835)	0.2198	−0.0346 (0.0289)	0.2314	−1.6617 (2.8394)	0.5585
Unweighted	−0.0928(0.0990)	0.3490	−0.3855 (0.2917)	0.1865	−0.0052 (0.0033)	0.1147	−0.1120 (0.2683)	0.6764
Table 3 Results from tests of association between weighted polygenic risk scores for lipid traits and cardiovascular disease and type 2 diabetes (T2D). Prevalent CVD in African American adults was determined using International Classification of Diseases (ICD) 9th edition, clinical modification (CM) codes and current procedural terminology (CPT) codes as well as search terms using problems list and clinical notes as described in Materials & Methods. Type 2 diabetes (T2D) case status was determined using an algorithm developed and validated by the eMERGE Network. For both CVD and T2D case and control groups, we determined the proportion of African American patients in the upper 90th percentile of each lipid trait weighted polygenic risk score (PRSw) as well as the proportion of patients in the lower 90th percentiles. Tests of association were performed as a 2×2 table, and shown are the resulting odds ratios, 95% confidence intervals (CI), and pvalues.

wPRS	Outcome and sample sizes	% Cases High PRSw/ % Cases Lower PRSw	% Controls High PRSw/ % Controls Lower PRSw	Odds Ratio (95% CI)	*P*-value
HDL-C	CVD $n_{cases} = 3{,}248$ $n_{controls} = 7{,}759$	10.7/89.3	9.7/90.3	1.12 (0.98, 1.29)	0.094
	T2D $n_{cases} = 1{,}330$ $n_{controls} = 4{,}053$	10.5/89.5	9.8/90.2	1.07 (0.87, 1.31)	0.53
LDL-C	CVD $n_{cases} = 3{,}244$ $n_{controls} = 7{,}743$	10.9/89.1	9.6/90.4	1.15 (1.01, 1.32)	0.04
	T2D $n_{cases} = 1{,}330$ $n_{controls} = 4{,}031$	9.4/90.6	10.2/89.8	0.91 (0.73, 1.13)	0.43
TG	CVD $n_{cases} = 3{,}154$ $n_{controls} = 7{,}522$	9.8/90.2	10.1/89.9	0.97 (0.84, 1.12)	0.72
	T2D $n_{cases} = 1{,}297$ $n_{controls} = 3{,}952$	8.6/91.4	10.4/89.6	0.81 (0.65, 1.01)	0.06
TC	CVD $n_{cases} = 3{,}231$ $n_{controls} = 7{,}712$	10.2/89.8	9.9/90.1	1.03 (0.90, 1.19)	0.62
	T2D $n_{cases} = 1{,}323$ $n_{controls} = 4{,}016$	9.1/90.9	10.3/89.7	0.88 (0.71, 1.09)	0.27

we calculated multi-population PRS for HDL-C, LDL-C, TG, and TC using (1) specific genetic variants and (2) their respective weights associated with these traits identified in a diverse population that includes a substantial African American sample (*Hu et al., 2020*) and applied them to an independent African American patient population from a biobank linked to real-world clinical data. Overall, we did not observe strong associations between unweighted or weighted lipid PRS for the lipid traits tested here or for T2D and other common clinical labs. We did, however, observe a nominal association between PRS_{LDL-C} and CVD.

The lack of associations between lipid PRS and their respective lipid traits was somewhat unexpected as previous reports have demonstrated strong associations between lipid traits and genetic or polygenic risk scores. In contrast to the present study, lipid trait PRS in Biobank Japan were strongly associated to their respective lipid labs and in the expected directions (*Tam et al., 2021*). Similarly, lipid PRS developed in the diverse Multi-Ethnic Study of Atherosclerosis (MESA) using variants discovered in European-descent GWAS were strongly associated with their respective lipid traits in the expected direction for

Drouet et al. (2023), *PeerJ*, DOI 10.7717/peerj.14910

Table 4 Limited phenome-wide association study for lipid trait weighted PRS and common clinical labs in African American adult patients. Lipid trait weighted polygenic risk scores (wPRS) were calculated as described in Materials & Methods. Lab values were transformed (*) when necessary, using log10(x+1). All tests of association were performed using linear regression for each lipid trait PRSw and each quantitative trait lab with the exception of estimated glomerular filtration rate (eGFR). We categorized eGFR where patients with eGFR ≥ 90 mL/min/1.73m2 are considered within normal limits and the reference group here (REF). Patients with eGFR 60-90 and < 60 were considered stage 2 and stage 3, respectively. Stage 2 and stage 3 were compared with stage 1, respectively, using logistic regression. Shown are test of association betas, standard errors (SE), and p-values by lipid trait PRSw and organized by clinical lab category.

	Weighted PRS							
	HDL-C		LDL-C		TG		TC	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Biochemistry								
Platelet Count	0.1469 (1.4862)	0.9213	−2.2072 (2.1179)	0.2974	0.4775 (1.8077)	0.7917	−2.1575 (2.1881)	0.3242
Glucose	1.5967 (0.8484)	0.0599	−0.3872 (1.2069)	0.7484	0.2874 (1.0440)	0.7831	−0.0652 (1.2427)	0.9582
Serum Creatinine	0.0067 (0.0053)	0.2057	0.0161 (0.0076)	0.0340	0.0053 (0.0065)	0.4122	0.0203(0.0078)	0.0092
Liver								
Serum Albumin *	−0.0004 (0.0006)	0.5824	0.0001 (0.0009)	0.8930	−0.0014 (0.0008)	0.0865	0.0001 (0.0009)	0.8876
Kidney								
Albumin Creatinine Ratio *	0.0952 (0.0344)	0.0057	−0.0190 (0.0494)	0.7008	0.0229 (0.0430)	0.5945	0.0156 (0.0508)	0.7596
Blood Urea Nitrogen *	0.0070 (0.0041)	0.0852	0.0039 (0.0058)	0.4969	0.0036 (0.0050)	0.4778	0.0088 (0.0059)	0.1390
Uric acid	−0.0780 (0.0664)	0.2403	0.1171 (0.0956)	0.2207	−0.0313 (0.0839)	0.7094	0.1376 (0.0988)	0.1640
Urine Creatinine *	0.0401 (0.0382)	0.2962	−0.0018 (0.0508)	0.9714	0.0600 (0.0498)	0.2307	0.0273 (0.0575)	0.6360
Urine Albumin *	0.0002 (0.0205)	0.9934	−0.0332 (0.0282)	0.2397	0.0348 (0.0250)	0.1653	−0.0152 (0.0293)	0.6038
eGFR								
Stage 2	−0.0136 (0.0437)	0.7552	0.1095 (0.0619)	0.0770	0.0557 (0.0537)	0.2995	0.1365 (0.0638)	0.0324
Stage 3	0.0745 (0.0468)	0.1117	0.04967 (0.0670)	0.4577	0.0568 (0.0580)	0.3275	0.0744 (0.0689)	0.2808

all populations except for TG in African Americans (*Johnson et al., 2015*). In a recent generalization study, European-derived LDL-C, HDL-C, and TG PRS were also strongly associated with their respective lipid traits in populations from Europe, Asia, and Africa (*Kuchenbaecker et al., 2019*). Strong associations between European-derived LDL-C, HDL-C, TG, and TC PRS have also been reported for a study of adolescents from the Netherlands (*Xie et al., 2020*).

Literature specific to lipid trait PRS derived from (all or in part) and applied to African Americans is limited. A recent assessment LDL-C, HDL-C, TG, and TC PRS derived from European, multi-ancestral, and sub-Saharan African discovery meta-analyses all demonstrated strong associations in the expected directions for their respective lipid traits in an independent sample of 7,103 participants from the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen) cohort (*Choudhury et al., 2022*). Of note is a recent study from BioVU where chip-wide LDL-C, HDL-C, and TG PRS were developed and applied to African American patients who largely overlap with the present study (*Dennis et al., 2021*). Strong associations were observed between each PRS and its respected lipid trait (*Dennis et al., 2021*). Consistent with the larger BioVU study, we did not identify strong associations between lipid trait PRS and common clinical labs in the limited PheWAS (*Dennis et al., 2021*). Differences between the larger BioVU study and the present study are numerous, including differences in sample size, genotyping array, computable phenotyping approach, year of clinical data extraction, and methods used to calculate PRS.

As the two studies accessing BioVU demonstrate, direct comparison of PRS associations in the literature is challenging for multiple reasons. First and foremost, there is no consensus on how to calculate or define PRS. As of this writing, the Polygenic Risk Score Catalog (*Lambert et al., 2021*) reports 66 PRS available in the literature for lipid traits, including 27,10, 9, and 7 for LDL-C, HDL-C, TG, and TC, respectively. PRS, even for the same trait (*e.g.*, LDL-C), differ by the number of variants included and how variants are chosen. PRS also differ by the discovery data from which they draw both variants and weights, much of which is still focused on European-descent populations. Finally, there are no standards on the evaluation and reporting of PRS performance. In the present study, we selected index variants associated with lipid traits in a diverse population that includes African Americans and calculated unweighted and weighted PRS using the absolute values of the betas reported in the discovery study. We applied the PRS to an independent sample of African American patients and reported the betas and *p*-values from unadjusted and adjusted tests of association. Collectively, the comparison challenges we experienced here highlight the need to develop approaches to standardize PRS being developed for implementation in a clinical setting to benefit patients (*Wand, Knowles & Clarke, 2021*).

While we did not observe strong associations between the PRS and their lipid traits, we did observe a nominal association between PRS_{LDL-C} and CVD. Our results are consistent with an early PRS_{LDL-C} applied to the Malmö Diet and Cancer Study, a prospective study of European-descent participants where PRS_{LDL-C} constructed from nine SNPs predicted risk of first CVD event independent of known risk factors (*Kathiresan et al., 2008*). Here, we observed a greater proportion of CVD cases among the African American patients in the 90th PRS_{LDL-C} percentile compared with African American patients in the same 90th

percentile without CVD. Despite differences in genetic ancestry, these data are consistent with PRS_{LDL} associations and incident CVD recently observed in the UK Biobank (*Tromp et al., 2022*). The trend observed here is also consistent with data in Biobank Japan, which suggested associations between coronary heart disease among T2D patients and PRS_{LDL}, PRS_{TG}, and PRS_{TC} (*Tam et al., 2021*). To the best of our knowledge, no similar studies yet exist that are specific to African Americans with CVD or lipid trait PRS examined for relationships to T2D risk in any population.

The present study has several limitations. The sample size is small, resulting in low statistical power. With an initial dataset of almost 12,000 African Americans, the dataset was reduced to ~2,000 per lipid trait due to missing data and data cleaning as is characteristic of clinical data. While previous published assessments suggest the computable phenotyping approaches employed here result in high quality, research grade variables (*Crawford et al., 2015*; *Crawford et al., 2018*; *Dumitrescu et al., 2015b*; *Dumitrescu et al., 2015c*; *Dumitrescu et al., 2015a*; *Goodloe et al., 2017*), the CVD variable is a mix of prevalent and incident cases, limiting the assessment of PRS risk. Other limitations are related to PRS calculation. We drew from the genotyped Metabochip and selected index variants associated with lipid traits regardless of linkage disequilibrium. Prevailing PRS methods discourage the inclusion of genetic variants in strong linkage disequilibrium, which can lead to artificial score inflation. The mostly null associations reported here suggest score inflation was not substantial. In addition to the already-noted small sample sizes, the mostly null associations observed here may be a result of an imperfect ancestral match (*Privé et al., 2022*) between the diverse and heterogeneous (*Choudhury et al., 2020*) populations of *Hu et al. (2020)* and the African American patient population studied here. While we based PRS calculations on data that include African Americans, we acknowledge that PRS best practices for diverse populations are actively being developed and evaluated (*Graham et al., 2021*).

CONCLUSIONS

Despite the challenges of sample size, this is one of the few studies of multi-population lipid PRS developed from a genomic discovery effort that includes African Americans and applied to an independent African American patient population. In general, larger, prospective diverse cohorts are needed to assess clinical utility of PRS for a variety of common clinic outcomes.

Funding

This study was funded by the National Human Genome Research Institute (NHGRI) grant U01HG004798 (to DCC), by institutional funds from the CWRU Cleveland Institute for Computational Biology, and by the Clinical and Translational Science Collaborative of Cleveland UL1TR000439 from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health (NIH). Domenica E. Drouet was supported by the Heart, Lung and Blood Summer Research Program 2R25HL103152-11.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

The National Human Genome Research Institute (NHGRI): U01HG004798.

The CWRU Cleveland Institute for Computational Biology.

The National Center for Advancing Translational Sciences (NCATS): UL1TR000439.

The Heart, Lung and Blood Summer Research Program: 2R25HL103152-11.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Domenica E. Drouet performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Shiying Liu performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Dana C. Crawford conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Human Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

The Vanderbilt University Institutional Review Board considers these de-identified data and study protocol (IRB #110579) to be non-human subjects research.

Data Availability

The NHGRI PAGE EAGLE MetaboChip study is available at dbGaP.

- The individual-level data are available for download by authorized investigators: https://view.ncbi.nlm.nih.gov/dbgap-controlled

- Data dictionaries, variable summaries, and documents are available on the dbGaP FTP site: https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs002767/phs002767.v1.p1

- The public summary-level phenotype data may be browsed at the dbGaP study report page: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002767.v1.p1

The NIH Genomic Data Sharing (GDS) Policy sets the expectation that the informed consent under which the data or samples were collected is the basis for determining the appropriateness of sharing data through unrestricted-access databases or NIH designated controlled-access data repositories (e.g.,dbGaP).

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.14910#supplemental-information.

REFERENCES

Buyske S, Wu Y, Carty CL, Cheng I, Assimes TL, Dumitrescu L, Hindorff LA, Mitchell S, Ambite JL, Boerwinkle E, Buzkova P, Carlson CS, Cochran B, Duggan D, Eaton CB, Fesinmeyer MD, Franceschini N, Haessler J, Jenny N, Kang HM, Kooperberg C, Lin Y, Marchand LL, Matise TC, Robinson JG, Rodriguez C, Schumacher FR, Voight BF, Young A, Manolio TA, Mohlke KL, Haiman CA, Peters U, Crawford DC, North KE. 2012. Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study. *PLOS ONE* **7**:e35651 DOI 10.1371/journal.pone.0035651.

Carlson CS, Matise TC, North KE, Haiman CA, Fesinmeyer MD, Buyske S, Schumacher FR, Peters U, Franceschini N, Ritchie MD, Duggan DJ, Spencer KL, Dumitrescu L, Eaton CB, Thomas F, Young A, Carty C, Heiss G, Le Marchand L, Crawford DC, Hindorff LA, Kooperberg CL, PAGE Consortium. 2013. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLOS Biology* **11**:e1001661 DOI 10.1371/journal.pbio.1001661.

Carroll MD, Kit BK, Lacher DA, Shero ST, Mussolino ME. 2012. Trends in lipids and lipoproteins in US adults, 1988–2010. *Journal of American Medical Association* **308**:1545–1554 DOI 10.1001/jama.2012.13260.

Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, Wells G, Kumuthini J, Shriner D, Fakim YJ, Ghoorah AW, Dareng E, Odia T, Falola O, Adebiyi E, Hazelhurst S, Mazandu G, Nyangiri OA, Mbiyavanga M, Benkahla A, Kassim SK, Mulder N, Adebamowo SN, Chimusa ER, Muzny D, Metcalf G, Gibbs RA, Trypano GEN Research Group, Rotimi C, Ramsay M, H3Africa Consortium, Adeyemo AA, Lombard Z, Hanchard NA. 2020. High-depth African genomes inform human migration and health. *Nature* **586(7831)**:741–748 DOI 10.1038/s41586-020-2859-7.

Choudhury A, Brandenburg JT, Chikowore T, Sengupta D, Boua PR, Crowther NJ, Agongo G, Asiki G, Gómez-Olivé FX, Kisiangani I, Maimela E, Masemola-Maphutha M, Micklesfield LK, Nonterah EA, Norris SA, Sorgho H, Tinto H, Tollman S, Graham SE, Willer CJ, AWI-Gen study, H3Africa Consortium, Hazelhurst S, Ramsay M. 2022. Meta-analysis of sub-Saharan African studies provides insights into genetic architecture of lipid traits. *Nature Communications* **13**:2578 DOI 10.1038/s41467-022-30098-w.

Christoffersen M, Tybjærg-Hansen A. 2021. Polygenic risk scores: how much do they add? *Current Opinion in Lipidology* **32**:157–162 DOI 10.1097/MOL.0000000000000759.

Cohen JD, Cziraky MJ, Cai Q, Wallace A, Wasser T, Crouse JR, Jacobson TA. 2010. 30-year trends in serum lipids among United States adults: results from the National Health and Nutrition Examination Surveys II, III, and 1999–2006. *American Journal of Cardiology* **106**:969–975 DOI 10.1016/j.amjcard.2010.05.030.

Cooke Bailey JN, Igo Jr RP. 2016. Genetic risk scores. *Current Protocols in Human Genetics* **91(1)**:29.1–1.29.9 DOI 10.1002/cphg.20.

Crawford DC, Goodloe R, Brown-Gentry K, Wilson S, Roberson J, Gillani NB, Ritchie MD, Dilks HH, Bush WS. 2012. Characterization of the metabochip in diverse populations from the international hapmap project in the epidemiologic architecture for genes linked to environment (eagle) project. *Pacific Symposium on Biocomputing* **2013**:188–199 DOI 10.1142/9789814447973_0019.

Crawford DC, Goodloe R, Farber-Eger E, Boston J, Pendergrass SA, Haines JL, Ritchie MD, Bush WS. 2015. Leveraging epidemiologic and clinical collections for genomic studies of complex traits. *Human Heredity* **79**:137–146 DOI 10.1159/000381805.

Crawford DC, Restrepo NA, Diggins KE, Farber-Eger E, Wells QS. 2018. Frequency and phenotype consequence of APOC3 rare variants in patients with very low triglyceride levels. *BMC Medical Genomics* **11**:66 DOI 10.1186/s12920-018-0387-1.

Dennis JK, Sealock JM, Straub P, Lee YH, Hucks D, Actkins K, Faucon A, Feng Y-CA, Ge T, Goleva SB, Niarchou M, Singh K, Morley T, Smoller JW, Ruderfer DM, Mosley JD, Chen G, Davis LK. 2021. Clinical laboratory test-wide association scan of polygenic scores identifies biomarkers of complex disease. *Genome Medicine* **13**:6 DOI 10.1186/s13073-020-00820-8.

Dumitrescu L, Carty CL, Taylor K, Schumacher FR, Hindorff LA, Ambite JL, Anderson G, Best LG, Brown-Gentry K, Bůžková P, Carlson CS, Cochran B, Cole SA, Devereux RB, Duggan D, Eaton CB, Fornage M, Franceschini N, Haessler J, Howard BV, Johnson KC, Laston S, Kolonel LN, Lee ET, MacCluer JW, Manolio TA, Pendergrass SA, Quibrera M, Shohet RV, Wilkens LR, Haiman CA, Le Marchand L, Buyske S, Kooperberg C, North KE, Crawford DC. 2011. Genetic determinants of lipid traits in diverse population from the Population Architecture using Genomics and Epidemiology (PAGE) study. *PLOS Genetics* **7**:e1002138 DOI 10.1371/journal.pgen.1002138.

Dumitrescu L, Diggins KE, Goodloe R, Crawford DC. 2015a. Testing population-specific quantitative trait associations for clinical outcome relevance in a biorepository linked to electronic health records: LPA and myocardial infarction in African Americans. *Pacific Symposium on Biocomputing* **2016**:96–107 DOI 10.1142/9789814749411_0010.

Dumitrescu L, Goodloe R, Bradford Y, Farber-Eger E, Boston J, Crawford DC. 2015b. The effects of electronic medical record phenotyping details on genetic association studies: HDL-C as a case study. *BioData Mining* **8**:15 DOI 10.1186/s13040-015-0048-2.

Dumitrescu L, Restrepo NA, Goodloe R, Boston J, Farber-Eger E, Pendergrass SA, Bush WS, Crawford DC. 2015c. Towards a phenome-wide catalog of human clinical traits impacted by genetic ancestry. *BioData Mining* **8**:35 DOI 10.1186/s13040-015-0068-y.

Dumitrescu L, Ritchie MD, Brown-Gentry K, Pulley JM, Basford M, Denny JC, Oksenberg JR, Roden DM, Haines JL, Crawford DC. 2010. Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genetics in Medicine* **12**:648–650 DOI 10.1097/GIM.0b013e3181efe2df.

Farber-Eger E, Goodloe R, Boston J, Bush WS, Crawford DC. 2017. Extracting country-of-origin from electronic health records for gene- environment studies as part of the

epidemiologic architecture for genes linked to environment (EAGLE) study. *AMIA Summits on Translational Science Proceedings* **2017**:50–57.

Goodloe R, Farber-Eger E, Boston J, Crawford DC, Bush WS. 2017. Reducing clinical noise for body mass index measures due to unit and transcription errors in the electronic health record. *Annual Medical Informatics Association Joint Summits on Translational Science Proceedings* **2017**:102–111.

Graham SE, Clarke SL, Wu KH, Kanoni S, Zajac GJM, Ramdas S, Surakka I, Ntalla I, Vedantam S, Winkler TW, Locke AE, Marouli E, Hwang MY, Han S, Narita A, Choudhury A, Bentley AR, Ekoru K, Verma A, Trivedi B, Martin HC, Hunt KA, Hui Q, Klarin D, Zhu X, Thorleifsson G, Helgadottir A, Gudbjartsson DF, Holm H, Olafsson I, Akiyama M, Sakaue S, Terao C, Kanai M, Zhou W, Brumpton BM, Rasheed H, Ruotsalainen SE, Havulinna AS, Veturi Y, Feng Q, Rosenthal EA, Lingren T, Pacheco JA, Pendergrass SA, Haessler J, Giulianini F, Bradford Y, Miller JE, Campbell A, Lin K, Millwood IY, Hindy G, Rasheed A, Faul JD, Zhao W, Weir DR, Turman C, Huang H, Graff M, Mahajan A, Brown MR, Zhang W, Yu K, Schmidt EM, Pandit A, Gustafsson S, Yin X, Luan J, Zhao JH, Matsuda F, Jang HM, Yoon K, Medina-Gomez C, Pitsillides A, Hottenga JJ, Willemsen G, Wood AR, Ji Y, Gao Z, Haworth S, Mitchell RE, Chai JF, Aadahl M, Yao J, Manichaikul A, Warren HR, Ramirez J, Bork-Jensen J, Kårhus LL, Goel A, Sabater-Lleal M, Noordam R, Sidore C, Fiorillo E, McDaid AF, Marques-Vidal P, Wielscher M, Trompet S, Sattar N, Møllehave LT, Thuesen BH, Munz M, Zeng L, Huang J, Yang B, Poveda A, Kurbasic A, Lamina C, Forer L, Scholz M, Galesloot TE, Bradfield JP, Daw EW, Zmuda JM, Mitchell JS, Fuchsberger C, Christensen H, Brody JA, Feitosa MF, Wojczynski MK, Preuss M, Mangino M, Christofidou P, Verweij N, Benjamins JW, Engmann J, Kember RL, Slieker RC, Lo KS, Zilhao NR, Le P, Kleber ME, Delgado GE, Huo S, Ikeda DD, Iha H, Yang J, Liu J, Leonard HL, Marten J, Schmidt B, Arendt M, Smyth LJ, Cañadas Garre M, Wang C, Nakatochi M, Wong A, Hutri-Kähönen N, Sim X, Xia R, Huerta-Chagoya A, Fernandez-Lopez JC, Lyssenko V, Ahmed M, Jackson AU, Irvin MR, Oldmeadow C, Kim HN, Ryu S, Timmers PRHJ, Arbeeva L, Dorajoo R, Lange LA, Chai X, Prasad G, Lorés-Motta L, Pauper M, Long J, Li X, Theusch E, Takeuchi F, Spracklen CN, Loukola A, Bollepalli S, Warner SC, Wang YX, Wei WB, Nutile T, Ruggiero D, Sung YJ, Hung YJ, Chen S, Liu F, Yang J, Kentistou KA, Gorski M, Brumat M, Meidtner K, Bielak LF, Smith JA, Hebbar P, Farmaki AE, Hofer E, Lin M, Xue C, Zhang J, Concas MP, Vaccargiu S, Van der Most PJ, Pitkänen N. 2021. The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**:675–679 DOI 10.1038/s41586-021-04064-3.

Hall JB, Dumitrescu L, Dilks HH, Crawford DC, Bush WS. 2014. Accuracy of administratively-assigned ancestry for diverse populations in an electronic medical record-linked biobank. *PLOS ONE* **9**:e99161 DOI 10.1371/journal.pone.0099161.

Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, Green ED. 2018. *Nature Reviews Genetics* **19**:175–185 DOI 10.1038/nrg.2017.89.

Hivert M-F, Jablonski KA, Perreault L, Saxena R, McAteer JB, Franks PW, Hamman RF, Kahn SE, Haffner S, DIAGRAM Consortium, Meigs JB, Altshuler D, Knowler

WC, Florez JC, Diabetes Prevention Program Research Group. 2011. Updated genetic score based on 34 confirmed type 2 diabetes loci associated with diabetes incidence and regression to normoglycemia in the diabetes prevention program. *Diabetes* **60**:1340–1348 DOI 10.2337/db10-1119.

Hu Y, Graff M, Haessler J, Buyske S, Bien SA, Tao R, Highland HM, Nishimura KK, Zubair N, Lu Y, Verbanck M, Hilliard AT, Klarin D, Damrauer SM, Ho Y-L, Program the VMV, Wilson PWF, Chang K-M, Tsao PS, Cho K, O'Donnell CJ, Assimes TL, Petty LE, Below JE, Dikilitas O, Schaid DJ, Kosel ML, Kullo IJ, Rasmussen-Torvik LJ, Jarvik GP, Feng Q, Wei W-Q, Larson EB, Mentch FD, Almoguera B, Sleiman PM, Raffield LM, Correa A, Martin LW, Daviglus M, Matise TC, Ambite JL, Carlson CS, Do R, Loos RJF, Wilkens LR, Marchand LL, Haiman C, Stram DO, Hindorff LA, North KE, Kooperberg C, Cheng I, Peters U. 2020. Minority-centric meta-analyses of blood lipid levels identify novel loci in the population architecture using genomics and epidemiology (PAGE) study. *PLOS Genetics* **16**:e1008684 DOI 10.1371/journal.pgen.1008684.

Igo Jr RP, Kinzy TG, Cooke Bailey JN. 2019. Genetic risk scores. *Current Protocols in Human Genetics* **104**:e95 DOI 10.1002/cphg.95.

Johnson L, Zhu J, Scott ER, Wineinger NE. 2015. An examination of the relationship between lipid levels and associated genetic markers across racial/ethnic populations in the multi-ethnic study of atherosclerosis. *PLOS ONE* **10**:e0126361 DOI 10.1371/journal.pone.0126361.

Kathiresan S, Melander O, Anevski D, Guiducci C, Burtt NP, Roos C, Hirschhorn JN, Berglund G, Hedblad B, Groop L, Altshuler DM, Newton-Cheh C, Orho-Melander M. 2008. *New England Journal of Medicine* **358**:1240–1249 DOI 10.1056/nejmoa0706728.

Kaur H, Crawford DC, Liang J, Benchek P, Consortium CB, Zhu X, Kallianpur AR, Bush WS. 2021. Replication of European hypertension associations in a case-control study of 9, 534 African Americans. *PLOS ONE* **16**:e0259962 DOI 10.1371/journal.pone.0259962.

Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, Denny JC, Peissig PL, Miller AW, Wei W-Q, Bielinski SJ, Chute CG, Leibson CL, Jarvik GP, Crosslin DR, Carlson CS, Newton KM, Wolf WA, Chisholm RL, Lowe WL. 2012. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association* **19**:212–218 DOI 10.1136/amiajnl-2011-000439.

Kuchenbaecker K, Telkar N, Reiker T, Walters RG, Eriksson ALinK, Gurdasani D, Gilly A, Southam L, Tsafantakis E, Karalefheri M, Seeley J, Kamali A, Asiki G, Millwood IY, Holmes M, Du H, Guo Y, Kumari M, Dedoussis G, l Li, Chen Z, Sandhu MS, Zeggini E, Understanding Society Scientific Group. 2019. The transferability of lipid loci across African, Asian and European cohorts. *Nature Communications* **10**:4330 DOI 10.1038/s41467-019-12026-7.

Kullo IJ, Lewis CM, Inouye M, Martin AR, Ripatti S, Chatterjee N. 2022. Polygenic scores in biomedical research. *Nature Reviews Genetics* **23**:524–532 DOI 10.1038/s41576-022-00470-z.

Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, McMahon A, Abraham G, Chapman M, Parkinson H, Danesh J, MacArthur JAL, Inouye M. 2021. The polygenic risk score catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics* **53**:420–425 DOI 10.1038/s41588-021-00783-5.

Levey AS, Stevens LA, Schmid CH, Zhang Y, Castro AF, Feldman HI, Kusek JW, Eggers P, Van Lente F, Greene T, Coresh J, for the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI). 2009. A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine* **150(9)**:604–612 DOI 10.7326/0003-4819-150-9-200905050-00006.

Li R, Chen Y, Ritchie MD, Moore JH. 2020. Electronic health records and polygenic risk scores for predicting disease risk. *Nature Reviews Genetics* **21**:493–502 DOI 10.1038/s41576-020-0224-1.

Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, Haiman CA, Heiss G, Kooperberg C, Marchand LL, Manolio TA, North KE, Peters U, Ritchie MD, Hindorff LA, Haines JL, for the PAGE Study. 2011. The next PAGE in understanding complex traits: design for the analysis of population architecture using genetics and epidemiology (PAGE) study. *American Journal of Epidemiology* **174**:849–859 DOI 10.1093/aje/kwr160.

Osterman MD, Kinzy TG, Cooke Bailey JN. 2021. Polygenic risk scores. *Current Protocols* **1**:e126 DOI 10.1002/cpz1.126.

Pendergrass SA, Brown-Gentry K, Dudek SM, Torstenson ES, Ambite JL, Avery CL, Buyske S, Cai C, Fesinmeyer MD, Haiman C, Heiss G, Hindorff LA, Hsu C-N, Jackson RD, Kooperberg C, Le Marchand L, Lin Y, Matise TC, Moreland L, Monroe K, Reiner AP, Wallace R, Wilkens LR, Crawford DC, Ritchie MD. 2011. The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genetic Epidemiology* **35**:410–422 DOI 10.1002/gepi.20589.

Pendergrass SA, Dudek SM, Crawford DC, Ritchie MD. 2010a. Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis. *BioData Mining* **3**:10 DOI 10.1186/1756-0381-3-10.

Pendergrass S, Dudek SM, Roden DM, Crawford DC, Ritchie MD. 2010b. Visual integration of results from a large dna biobank (biovu) using synthesis-view. *Pacific Symposium on Biocomputing* **2011**:265–275 DOI 10.1142/9789814335058_0028.

Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, Lam M, Iyegbe C, Strawbridge RJ, Brick L, Carey CE, Martin AR, Meyers JL, Su J, Chen J, Edwards AC, Kalungi A, Koen N, Majara L, Schwarz E, Smoller JW, Stahl EA, Sullivan PF, Vassos E, Mowry B, Prieto ML, Cuellar-Barboza A, Bigdeli TB, Edenberg HJ, Huang H, Duncan LE. 2019. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**:589–603 DOI 10.1016/j.cell.2019.08.051.

Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, Vilhjálmsson BJ. 2022. Portability of 245 polygenic scores when derived from the UK Biobank and

applied to 9 ancestry groups from the same cohort. *American Journal of Human Genetics* **109**(1):12–23 DOI 10.1016/j.ajhg.2021.11.008.

Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. 2010. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clinical and Translational Science* **3**:42–48 DOI 10.1111/j.1752-8062.2010.00175.x.

R Core Team. 2021. R: A language and environment for statistical computing. Version 4.1.0. Vienna: R Foundation for Statistical Computing. *Available at* https://www.r-project.org.

Roden D, Pulley J, Basford M, Bernard G, Clayton E, Balser J, Masys D. 2008. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics* **84**:362–369 DOI 10.1038/clpt.2008.89.

Sirugo G, Williams SM, Tishkoff SA. 2019. The missing diversity in human genetic studies. *Cell* **177**:1080 DOI 10.1016/j.cell.2019.04.032.

Tam CHT, Lim CKP, Luk AOY, Ng ACW, Lee H, Jiang G, Lau ESH, Fan B, Wan R, Kong APS, Tam W, Ozaki R, Chow EYK, Lee K, Siu S, Hui G, Tsang C, Lau K, Leung JYY, Tsang M, Kam G, Lau I, Li JKY, Yeung VTF, Lau E, Lo S, Fung S, Cheng Y, Chow C, Hu M, Yu W, Tsui SKW, Huang Y, Lan H, Szeto C, Tang NLS, Ng MCY, So W, Tomlinson B, Chan JCN, Ma RCW. 2021. Development of genome-wide polygenic risk scores for lipid traits and clinical applications for dyslipidemia, subclinical atherosclerosis, and diabetes cardiovascular complications among East Asians. *Genome Medicine* **13**:29 DOI 10.1186/s13073-021-00831-z.

Torkamani A, Wineiger NE, Topol EJ. 2018. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**:581–590 DOI 10.1038/s41576-018-0018-x.

Tromp TR, Cupido AJ, Reeskamp LF, Stroes ESG, Hovingh GK, Defesche JC, Schmidt AF, Zuurbier L. 2022. Assessment of practical applicability and clinical relevance of a commonly used LDL-C polygenic score in patients with severe hypercholesterolemia. *Atherosclerosis* **340**:61–67 DOI 10.1016/j.atherosclerosis.2021.10.015.

Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burtt NP, Fuchsberger C, Li Y, Erdmann J, Frayling TM, Heid IM, Jackson AU, Johnson T, Kilpeläinen TO, Lindgren CM, Morris AP, Prokopenko I, Randall JC, Saxena R, Soranzo N, Speliotes EK, Teslovich TM, Wheeler E, Maguire J, Parkin M, Potter S, Rayner NW, Robertson N, Stirrups K, Winckler W, Sanna S, Mulas A, Nagaraja R, Cucca F, Barroso I, Deloukas P, Loos RJF, Kathiresan S, Munroe PB, Newton-Cheh C, Pfeufer A, Samani NJ, Schunkert H, Hirschhorn JN, Altshuler D, McCarthy MI, Abecasis GR, Boehnke M. 2012. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLOS Genetics* **8**:e1002793 DOI 10.1371/journal.pgen.1002793.

Wand H, Knowles JW, Clarke SL. 2021. The need for polygenic score reporting standards in evidence-based practice: lipid genetics use case. *Current Opinion in Lipidology* **32**:89–95 DOI 10.1097/MOL.0000000000000733.

Wang Y, Tsuo K, Kanai M, Neale BM, Martin AR. 2022. Challenges and opportunities in developing more generalizable polygenic risk scores. *Annual Review of Biomedical Data Science* **5**:293–320 DOI 10.1146/annurev-biodatasci-111721-074830.

Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL, Belbin GM, Bien SA, Cheng I, Cullina S, Hodonsky CJ, Hu Y, Huckins LM, Jeff J, Justice AE, Kocarnik JM, Lim U, Lin BM, Lu Y, Nelson SC, Park SL, Poisner H, Preuss MH, Richard MA, Schurmann C, Setiawan VW, Sockell A, Vahi K, Verbanck M, Vishnu A, Walker RW, Young KL, Zubair N, Acuña Alonso V, Ambite JL, Barnes KC, Boerwinkle E, Bottinger EP, Bustamante CD, Caberto C, Canizales-Quinteros S, Conomos MP, Deelman E, Do R, Doheny K, Fernández-Rhodes L, Fornage M, Hailu B, Heiss G, Henn BM, Hindorff LA, Jackson RD, Laurie CA, Laurie CC, Li Y, Lin DY, Moreno-Estrada A, Nadkarni G, Norman PJ, Pooler LC, Reiner AP, Romm J, Sabatti C, Sandoval K, Sheng X, Stahl EA, Stram DO, Thornton TA, Wassel CL, Wilkens LR, Winkler CA, Yoneyama S, Buyske S, Haiman CA, Kooperberg C, Le Marchand L, Loos RJF, Matise TC, North KE, Peters U, Kenny EE, Carlson CS. 2019. Genetic analysis of diverse populations improves discovery for complex traits. *Nature* **570**:514–518 DOI 10.1038/s41586-019-1310-4.

Xie T, Wang B, Nolte IM, Van der Most P, Oldehinkel AJ, Hartman CA, Sieder H. 2020. Genetic risk scores for complex disease traits in youths. *Circulation: Genomic and Precision Medicine* **13**:e002775 DOI 10.1161/CIRCGEN.119.002775.