frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 01 August 2023
DOI 10.3389/fgene.2023.1181167

OPEN ACCESS

EDITED BY
Yuriy L. Orlov,
I. M. Sechenov First Moscow State
Medical University, Russia

REVIEWED BY
Eaaswarkhanth Muthukrishnan,
Georgia Institute of Technology,
United States
Nicholas B. Larson,
Mayo Clinic, United States

*CORRESPONDENCE
Eimear E. Kenny,
✉ eimear.kenny@mssm.edu

†PRESENT ADDRESS
Elena P. Sorokin,
Calico Life Sciences LLC, South San
Francisco, CA, United States
Noura S. Abul-Husn,
23andMe, Sunnyvale, CA, United States

RECEIVED 07 March 2023
ACCEPTED 10 July 2023
PUBLISHED 01 August 2023

CITATION
Cullina S, Wojcik GL, Shemirani R, Klarin D, Gorman BR, Sorokin EP, Gignoux CR, Belbin GM, Pyarajan S, Asgari S, Tsao PS, Damrauer SM, Abul-Husn NS and Kenny EE (2023), Admixture mapping of peripheral artery disease in a Dominican population reveals a putative risk locus on 2q35. *Front. Genet.* 14:1181167. doi: 10.3389/fgene.2023.1181167

Admixture mapping of peripheral artery disease in a Dominican population reveals a putative risk locus on 2q35

Sinead Cullina[1,2], Genevieve L. Wojcik[3], Ruhollah Shemirani[1], Derek Klarin[4,5], Bryan R. Gorman[6,7], Elena P. Sorokin[8†], Christopher R. Gignoux[9,10,11], Gillian M. Belbin[1,12], Saiju Pyarajan[6,13], Samira Asgari[1,2], Philip S. Tsao[4], Scott M. Damrauer[14,15,16], Noura S. Abul-Husn[1,17†] and Eimear E. Kenny[1,2,17,12*]

[1]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [3]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States, [4]VA Palo Alto Healthcare System, Palo Alto, CA, United States, [5]Division of Vascular Surgery, Stanford University School of Medicine, Palo Alto, CA, United States, [6]Center for Data and Computational Sciences (C-DACS), VA Boston Healthcare System, Boston, MA, United States, [7]Booz Allen Hamilton, McLean, VA, United States, [8]Department of Genetics, Stanford University, Stanford, CA, United States, [9]Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [10]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [11]Colorado Center for Personalized Medicine, Aurora, CO, United States, [12]Division of General Internal Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [13]Department of Medicine, Brigham Women's Hospital, Harvard Medical School, Boston, MA, United States, [14]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, United States, [15]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [16]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, United States, [17]Division of Genomic Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Peripheral artery disease (PAD) is a form of atherosclerotic cardiovascular disease, affecting ~8 million Americans, and is known to have racial and ethnic disparities. PAD has been reported to have a significantly higher prevalence in African Americans (AAs) compared to non-Hispanic European Americans (EAs). Hispanic/Latinos (HLs) have been reported to have lower or similar rates of PAD compared to EAs, despite having a paradoxically high burden of PAD risk factors; however, recent work suggests prevalence may differ between sub-groups. Here, we examined a large cohort of diverse adults in the Bio*Me* biobank in New York City. We observed the prevalence of PAD at 1.7% in EAs vs. 8.5% and 9.4% in AAs and HLs, respectively, and among HL sub-groups, the prevalence was found at 11.4% and 11.5% in Puerto Rican and Dominican populations, respectively. Follow-up analysis that adjusted for common risk factors demonstrated that Dominicans had the highest increased risk for PAD relative to EAs [OR = 3.15 (95% CI 2.33–4.25), $p < 6.44 \times 10^{-14}$]. To investigate whether genetic factors may explain this increased risk, we performed admixture mapping by testing the association between local ancestry and PAD in Dominican Bio*Me* participants (N = 1,813) separately from European, African, and Native American (NAT) continental ancestry tracts. The top association with PAD was an NAT ancestry tract at chromosome 2q35 [OR = 1.96 (SE = 0.16), $p < 2.75 \times 10^{-05}$) with 22.6% vs. 12.9% PAD prevalence in heterozygous NAT tract carriers versus non-carriers, respectively. Fine-mapping at this locus implicated tag SNP rs78529201 located within a long intergenic non-coding RNA (lincRNA) *LINC00607*, a gene expression regulator of key genes related to thrombosis and extracellular remodeling of endothelial cells, suggesting a putative link of the 2q35 locus to PAD etiology. Efforts to reproduce the signal in other Hispanic cohorts were unsuccessful. In summary, we showed how leveraging health system data helped understand nuances of PAD risk across HL sub-groups and admixture mapping approaches elucidated a putative risk locus in a Dominican population.

KEYWORDS

admixture mapping, peripheral artery disease, biobanks, Dominicans, genetic epidemiology

Introduction

Peripheral arterial disease (PAD) is a form of atherosclerotic disease leading to peripheral artery obstruction. PAD is characterized by classic symptoms of intermittent claudication of the lower extremities and ankle–brachial systolic pressure index (ABI) < 0.9 (Crawford et al., 2016). PAD is a common disease affecting ~8 million Americans (Leeper et al., 2012) with a combined annual healthcare cost exceeding $21 billion in the United States (US) (Mahoney et al., 2008). Symptomatic PAD negatively impacts the quality of life of patients, and severe outcomes include chronic ischemia and amputation (Elnady and Saeed, 2017; Klarin et al., 2021). Even when asymptomatic, PAD is associated with systemic vascular disease and increased rates of myocardial infarction, stroke, and death (American Diabetes Association, 2003). The incidence of PAD at a population level depends on a nuanced interplay between ancestral, social, and environmental factors (Allison et al., 2010; Wassel et al., 2011; Klarin et al., 2021). Non-genetic risk factors include diabetes, smoking, thrombosis, increased age, systolic blood pressure, C-reactive protein levels, and serum total cholesterol (Kannel and McGee, 1985; Wassel et al., 2011). Smoking and thrombosis are thought to have a larger contribution to PAD etiology compared to other circulatory diseases (Klarin et al., 2021). The heritability of PAD based on family studies is estimated to be 20%–30% (Klarin et al., 2021). Genome-wide association studies (GWAS) of PAD have provided further insights into the biological pathways contributing to this polygenic disease (Thorgeirsson et al., 2008; Kullo et al., 2014; Matsukura et al., 2015). The largest PAD GWAS conducted to date is in a multi-ancestry Million Veteran Program (MVP) cohort which identified 19 genome-wide significant (GWS) loci. Genes underlying PAD-associated loci are associated with biological processes linked to known PAD risk factors, including type 2 diabetes, smoking habits, hypertension, lipids, and thrombosis (Klarin et al., 2019; Klarin et al., 2021).

There are known racial and/or ethnic disparities in incidence rates of PAD (Allison et al., 2007; Shu and Santulli, 2018; Belbin et al., 2021; Hackler et al., 2021). African Americans (AAs) are reported to have the highest rates of PAD as well as PAD-associated amputation compared to other race and/or ethnicity groups in the United States (Shu and Santulli, 2018; Matsushita et al., 2019; Hackler et al., 2021). Studies have suggested that rates of PAD in Hispanic/Latinos (HLs) are lower than those in both non-Hispanic whites and AAs (Allison et al., 2006); however, there is also evidence that PAD rates may differ across HL sub-groups, with higher rates in HL groups with origins in the Caribbean (Daviglus et al., 2012; Allison et al., 2015). Previous phenome-wide association analysis from our group in the diverse Bio*Me* biobank in a large health system in New York City (NYC) observed increased odds of PAD in both AAs and in two HL sub-groups, Puerto Rican and Dominican. Of the three populations, the highest odds were observed in the Dominican population (Belbin et al., 2021).

Admixture mapping is a powerful approach for genomic discovery when it is suspected that the prevalence of disease and/or frequency of underlying causal variants may differ across populations. Unlike GWAS approaches, which treat population structure as a confounder, admixture mapping leverages genetic ancestry differences at a haplotypic level, usually in populations with recent ancestry from two or more continents, to test for correlation with health outcomes of interest (Winkler et al., 2010). The basic principle underlying admixture mapping is based on the observation that if a causal variant for a trait or disease is at different frequencies in the ancestral populations of admixed individuals and if the disease itself is more prevalent in that particular ancestral population, then it is possible to scan for regions where local ancestry is statistically enriched. This is performed by conducting association tests between local ancestries derived from a specific ancestral population (e.g., European) at each genomic location and the trait of interest in cases vs. controls. The objective is to identify genomic regions where the local ancestry differs significantly between individuals with different trait outcomes. Examples of admixture mapping discoveries include an association of *APOL1* and renal disease in AAs (Kao et al., 2008; Kopp et al., 2008) and numerous loci with Alzheimer's disease in HLs (Horimoto et al., 2021; Kizil et al., 2022). Notably, admixture mapping previously identified rs9665943 as being a risk locus for PAD (ankle–arm index) in AAs (Scherer et al., 2010). In this study, we examined the prevalence of PAD in self-reported race and/or ethnicity groups and population groups inferred using genetic ancestry, within the Bio*Me* biobank at the Mount Sinai health system in NYC. We found that Dominicans have the highest PAD risk. Consequently, we conducted genomic discovery analysis using admixture mapping in the Dominican population to identify regions associated with an elevated risk for PAD.

Materials and methods

Study population

The Bio*Me* biobank is an electronic health record-linked biorepository that has been enrolling participants from across the Mount Sinai health system in NYC since 2007. There are currently over 50,000 participants enrolled in the Bio*Me* biobank under an

TABLE 1 Characteristics of BioMe participants showing PAD prevalence across PAD risk factors, self-reported, and IBD community membership. *p*-values represent a chi-squared test for discrete variables and Mann–Whitney test for continuous variables.

Characteristics of the study participants				*p*-value
	N	PAD case	PAD control	
All Bio*Me* participants (N, %)	57,580	3,762 (6.5%)	53,818 (93.5%)	
Sex (N women, %)	57,776	2,049 (54.4%)	31,422 (58.2%)	6.95×10^{-06}
Age (median years, IQR)	57,671	73 (16)	59 (28)	2.22×10^{-308}
BMI (median kg/hˆ2, IQR)	59,340	28.2 (8.2)	26.7 (7.8)	1.60×10^{-51}
T2D (N cases, %)	51,532	2,130 (66.9%)	9327 (19.2%)	2.22×10^{-308}
Total cholesterol (median mg/dL, IQR)	42,372	166 (50)	181 (50)	6.17×10^{-88}
HDL-C (median mg/dL, IQR)	38,483	47 (18.5)	53 (22)	2.54×10^{-99}
Triglyceride (mg/dL, median (IQR))	42,170	120 (74)	107 (74)	9.11×10^{-43}
Ever smoked? (N yes, %)	32,983	945 (52.0%)	10,855 (34.8%)	5.26×10^{-50}
Age groups (N, %)				2.22×10^{-308}
Less than 40	10,608	35 (0.3%)	10,573 (99.7%)	
40–69	30,123	1,411 (4.7%)	28,712 (95.3%)	
Greater than or equal to 70	16,940	2,320 (13.7%)	14,620 (86.3%)	
Self-reported groups (N, %)				5.33×10^{-160}
African American	11,472	980 (8.5%)	10,492 (91.5%)	
East/South-East Asian	2051	35 (1.7%)	2016 (98.3%)	
European American	16,720	545 (3.3%)	16,175 (96.7%)	
Hispanic/Latino	19,574	1833 (9.4%)	17,741 (90.6%)	
Native American	80	12 (15%)	68 (85%)	
South Asian	1,484	48 (3.2%)	1,436 (96.8%)	
Other	1907	101 (5.3%)	1806 (94.7%)	
Multiple selected	2,205	101 (4.6%)	2,104 (95.4%)	
Not available	2087	107 (5.1%)	1980 (94.9%)	
Genetic ancestry groups (%)				3.63×10^{-100}
African American/African	7,191	634 (8.8%)	6,557 (91.2%)	
Ashkenazi Jewish	4,408	152 (3.4%)	4,256 (96.6%)	
Non-Ashkenazi Jewish European American	5,990	195 (3.3%)	5,795 (96.7%)	
Filipino and other Southeast Asian	614	13 (2.1%)	601 (97.9%)	
Dominican	1971	227 (11.5%)	1744 (88.5%)	
Ecuadorian	438	31 (7.1%)	407 (92.9%)	
Puerto Rican	5,343	608 (11.4%)	4,735 (88.6%)	
Other Central and South American	1,025	67 (6.5%)	958 (93.5%)	

Institutional Review Board (IRB)-approved study protocol and consent (IRB 07-0529). Recruitment occurs predominantly through ambulatory care practices, and participants consent to provide whole blood-derived germline DNA and plasma samples which are banked for future research. Participants also complete a questionnaire providing personal and family history as well as demographic and lifestyle information as has been previously described (Abul-Husn et al., 2021; Belbin et al., 2021). Bio*Me* participants represent the broad diversity of the New York metropolitan area, and more than 65% of participants represent minority populations in the US. All participants provided informed consent, and the study was approved by the Icahn School of Medicine at Mount Sinai's IRB (protocol number 07-0529).

Using self-reported and genetic ancestry information to define population groups

All participants were asked multiple choice questions at enrollment regarding their heritage and country of birth of self, parents, and grandparents, which have been previously described by Belbin et al. (2021). Participants' responses to the heritage question were mapped into eight single self-reported groups for this study, namely, AA, East/South-East Asian (EAsn), South Asian (SA), Native American (NA), EA, HL, other, and multiple selected. Those participants who selected either "Hispanic/Latino" alone or "Hispanic/Latino" in addition to one or more other categories were designated as HL. Participants who selected "White/Caucasian" and/or "Ashkenazi Jewish" were designated as EA. Participants who selected either "Mediterranean" or "other" or a combination of both were assigned Other, and those with any other combination of multiple race ethnicity labels were assigned Multiple Selected.

Genetically inferred ancestry information was used to designate sub-populations. In brief, array data [Illumina Omni-Express (OMNI) and Illumina Multi-Ethnic Global Array (MEGA)] were phased and used to infer pairwise shared haplotypes identical by descent (IBD). Unsupervised clustering methods based on population-level IBD sharing were used to define clusters or IBD communities of individuals sharing recent cryptic relatedness. Over 50% of participants responded to questions about their country of birth which were used to determine the positive predictive value (PPV) of IBD communities in detecting recent patterns of diaspora to NYC. For example, one IBD community (N = 2,075) had a high confidence of predicting (PPV > 0.9) individuals who were born or who had parents or grandparents born in the Dominican Republic and was designated the Dominican community; PPVs and definitions of all IBD communities used in the downstream analysis are described in detail in the work of Belbin et al. (2021) and reported in Table 1.

Phenotyping using electronic health records and survey data

Biological sex at birth and age was extracted from Bio*Me* questionnaire data. Electronic health record (EHR) data were accessed to extract relevant disease outcomes and biomarker data. Individuals were designated PAD cases based on having at least one instance of the PAD International Classification of Diseases 9th Revision (ICD-9) billing code 443.9 (from 2007 to 2018) or ICD-10 code I73.9 (from 2018 to 2020) within their EHR records from the Mount Sinai health system. Controls were defined as those with no record of either ICD-9 code 443.9 or ICD-10 code I73.9 in the EHR. Type 2 diabetes mellitus (T2D) cases and controls were defined using the Northwestern University Type 2 diabetes mellitus algorithm described by Jennifer Pacheco and Will Thompson (Pacheco and Thompson, 2012). Northwestern University. Type 2 Diabetes Mellitus., 2012. Triglyceride (TG; mg/dL), high-density lipoprotein (HDL; mg/dL), and total cholesterol (TC; mg/dL) laboratory values were extracted for all encounters for each participant (2007–2021). Laboratory values with invalid entry "999999" were removed. For each biomarker, the median value per participant was calculated. Median values per biomarker were plotted separately per sex and per self-reported population label. Outlier values were defined as individuals with log10-transformed median values greater than third quantile + (1.5*IQR) or first quantile–(1.5*IQR) (total N participants removed per biomarker: HDL = 506, TC = 562, and TG = 382). Biomarker values were then converted to z-scores for downstream analysis. The same filtering and normalization steps were applied per IBD community and per sex for IBD community-based PAD risk analysis (total N participants removed per biomarker: HDL = 269, TC = 296, and TG = 202). Body mass index (BMI) measures were extracted from the EHR (2007–2022), and the median value per participant was calculated with outliers removed in the same manner as described for biomarkers for both self-reported groups (N participants removed = 963) and genetically inferred sub-groups (N participants removed = 455) and converted to z-scores.

PAD incidence across population groups

Statistical tests for PAD odds across both self-reported groups and genetically inferred sub-groups were performed using a generalized linear model (GLM) in R. PAD risk was tested separately within each self-reported group relative to self-reported EAs. The same models were used for association testing within each genetically inferred sub-group, and this time, the aim was to test for risk relative to non-Jewish Europeans. The three models for regression analysis were defined as model 1: PAD ~ population group + age + sex; model 2: PAD ~ model 1 + BMI; and model 3: PAD ~ model 2 + T2D + TG + TC + HDL.

Global ancestry inference

OMNI and MEGA genotype data were used to calculate global ancestry proportions (Supplementary Table S1). Reference panels of 100 individuals each representing three continental populations, African (AFR), European (EUR), and Native American (NAT), were constructed using genotype array data from the 1000 Genomes Project (1KGP), Human Diversity Genome Project (HDGP), and Polygenic Architecture using Genetics and Epidemiology (PAGE) study (Supplementary Figure S2). Because Dominicans have ancestry primarily from three continental populations (European, African, and Native American), reference samples representing each of these populations were included in global ancestry calling. We randomly selected unrelated individuals from two European ancestry and two African ancestry reference populations in the 1KGP, Utah residents with Northern and Western European ancestry (CEU; N = 50), Iberian populations in Spain (IBS; N = 50), Yoruba in Ibadan, Nigeria (YRI; N = 50), and Luhya in Webuye, Kenya (LWK; N = 50) (1000 Genomes Project Consortium et al., 2015). We selected unrelated individuals with maximal NAT genetic ancestry from four NAT ancestry reference populations, an indigenous population from Oaxaca, Mexico (N = 25), in the HGDP (Cann et al., 2002), indigenous populations from

Honduras (N = 25) and Columbia (N = 25), and a Peruvian population (N = 25) in PAGE (Bien et al., 2016). MEGA, OMNI genotyping data were merged with the HGDP, 1KGP, and PAGE reference panels using PLINK (v1.9), leaving a total of n = 395,531 SNPs (Purcell et al., 2007). Sites were filtered to remove palindromes, and a minor allele frequency (MAF) threshold of 1% was applied. Linkage disequilibrium (LD) pruning was performed using PLINK according to the parameters --indep-pairwise 50 5 0.3. Regions known to be under recent selection were removed: the human leukocyte antigen region (chr6:27000000–35000000, hg37), the lactase gene (chr2:135000000–137000000), an inversion on chromosome 8 (chr8:6000000–16000000), a region of extended LD on chromosome 17 (chr17:40000000–45000000), the ectodysplasin A receptor gene (chr2:109000000–110000000), and the T-cell receptor beta variable 9 gene (chr7:142000000–142500000). Following these filtering steps, n = 155,702 SNPs and N = 2133 participants remained including N = 300 reference individuals with a total genotyping rate of 99%. ADMIXTURE software was used to calculate global ancestry proportions with 5-fold cross-validation, unsupervised, with K values set to 2, 3, 4, and 5 (Belbin et al., 2021). A nonparametric bootstrapping approach was used to calculate confidence intervals for each global ancestry proportion using the np.boot() function from the nptest package in R.

Local ancestry inference

Eagle v2.0 was used to phase the merged MEGA, OMNI, and reference panel dataset per chromosome using default parameters, and no phasing reference panel was used (Loh et al., 2016). The same filtering steps as before were used in this step, but no LD pruning or MAF filter was applied and regions defined as under recent selection were also not removed, leaving a total of n = 377,798 SNPs for analysis. Local ancestry (LA) calling was performed on this phased dataset using RFMix V1 software with the default parameters (Maples et al., 2013). LA depth was plotted for each SNP, and sites that deviated ±2 SDs from the median were removed along with any LA calls in the HLA region (n SNPs = 371,185). The haploid AFR, EUR, and NAT calls per individual were summed to obtain an LA inference-derived global proportion ancestry. These LA inference-derived global proportions were then compared to the proportion of global ancestry calls from the corresponding ancestry component at K = 3 in the ADMIXTURE analysis, and no outliers with discordance in ancestry proportion greater than 5% were observed (Supplementary Figures S4A–C).

Sample and site-level quality control for genomic discovery

The Dominican discovery cohort used in the GWAS analysis was genotyped using either the OMNI (N = 862) or MEGA (N = 803) arrays. Both MEGA and OMNI genotyped samples were imputed to the phase 3 1KGP reference panel using SHAPEIT2 for phasing and IMPUTE2 for imputation. MEGA imputation was carried out at the University of Washington Genetic Analysis Center, and quality control filtering of sites and imputation details are previously described by Wojcik et al. (2019). Imputation of OMNI genotype data, including details of genotype data quality control before phasing and imputation, is described in detail in the work of Belbin et al. (2017). In brief, samples with plate failures, call rates <98%, or deviances in rates of heterozygosity were removed. The samples with discordance between genetic and EHR recorded sex and duplicates were also removed. Sites with a call rate of <95% were filtered out along with sites significantly deviating from Hardy–Weinberg equilibrium ($p < 1 \times 10^{-5}$), calculated within ancestry groups separately. Further imputation-specific quality control steps included the removal of sites that failed the miss-hap test ($p < 1 \times 10^{-8}$) in PLINK and duplicated sites. The phased genotype data (n SNPs = 828,109 and N samples = 11,212) were imputed using IMPUTE2 in 5 MB chunks using the parameters "-Ne 20000 -buffer 250 -filt_rules_l 'ALL < 0.0002' 'ALL > 0.9998'."

Admixture mapping

A total of N = 1,813 (N = 245 cases and N = 1,568 controls) unrelated participants from the Dominican community, genotyped on either OMNI (N = 924) or MEGA (N = 889) arrays, were used to perform genome-wide admixture mapping. Separate LA haplotype call sets were constructed for each of the three ancestral groups, and haplotypes were represented as additive vectors (i.e., 0 = non-carriers, 1 = heterozygous, and 2 = homozygous). Each LA call set was used as a predictor in iterative GLMs according to the formula PAD~LA + Sex + Array Type (R version 3.2.0). The STEAM R package was used to calculate the admixture mapping significance threshold using the get_thresh_simstat() function with nreps = 10,000 (Grinde et al., 2019). The number of generations since the admixture (g) parameter was calculated using the correlation of local ancestry between pairs of pruned loci in the RFMix output. Loci were pruned to include one SNP per RFMix window (0.2 Cm). This correlation file was used with the get_g() function to calculate g = 9.729661.

Testing the generalizability of admixture mapping in Hispanic/Latino populations

Two additional cohorts of HL participants were identified to determine the generalizability of admixture mapping-associated loci. The first cohort was an independent dataset of self-reported, unrelated, HL participants in Bio*Me* (N = 6,801) that had not been included in the Dominican discovery cohort described previously. LA was inferred as described previously, and a GLM in R was used with the model PAD~LA + Sex + Array Type using NAT tracts only. The second cohort was an independent dataset of Hispanic ancestry participants (N = 3,675 cases, N = 29,558 controls) in the MVP (Gaziano et al., 2016). Array genotyping and quality control were described previously (Hunter-Zinck et al., 2020). LA inference was carried out using RFMIX version 2 (Maples et al., 2013). A local ancestry reference panel was constructed from the Genome Aggregate Database (GnomAD)1KGP and HGDP call set to version 3.1.2 (Karczewski et al., 2020). Non-admixed (>90% estimated ancestry) samples in EUR (N = 631), AFR (N = 695), and NAT (N = 78) populations were included in the LA inference reference panel. To ensure suitable phase quality, reference panel samples were phased in SHAPEIT4 using the TOPMed reference panel (n = 194,512 haplotypes) as a phasing reference. LA inference on MVP samples was performed using three rounds of expectation maximization using RFMIX. The number of inferred NAT haplotypes was then tested additively in a GLM model using PLINK 2.0, where the model included the covariates sex, age, and principal components (PCs) 1–5 (Chang et al., 2015).

Genome-wide association mapping

PLINK (v1.90) was used to carry out genome-wide association testing in the Dominicans using the --logistic function in both MEGA- and OMNI-imputed datasets separately. MAF thresholds of 1% were applied to MEGA- and OMNI-imputed data. GWAS analysis using the MEGA-imputed dataset (n = 14,216,111) consisted of N = 121 cases and N = 682 controls. The OMNI dataset (n = 13,717,849) included N = 118 cases and N = 744 controls. Quality control steps for calculating PCs are the same as those described in the LA inference, and a 1% MAF filter was applied (n = 281,666). PCs were calculated using the SMARTPCAv10210 software from the EIGENSOFTv5.0.1 (Price et al., 2006). Age, sex, and PCs 1–5 were used as covariates. METAL was used for the meta-analysis of OMNI and MEGA GWAS results (nsites = 11,382,663) (Willer et al., 2010).

Conditional analysis

Conditional logistic regressions were used to fine-map the admixture mapping signal on chromosome 2 (215–220 Mb) using R. Only participants with both local ancestry calls and 1KGP phase 3-imputed genotype data were included (N = 1,687). Genotype information in a 5 Mb region on chromosome 2: 215,041,532–219,954,973 was used, and NAT haplotypes at the admixture mapping peak (chr2:216,636,519) were converted in the form of an additive vector of 0 = non-carriers, 1 = heterozygous, and 2 = homozygous. Each SNP within the admixture mapping signal was tested as a predictor variable along with NAT tracts, sex and array type, and PAD as the outcome in iterative GLMs. The SNP which reduced the significance of the admixture mapping signal by the greatest amount was SNP rs78529201 (chr2:216,518,626). Additional regression analysis then tested NAT tracts for association with PAD with the rs78529201 genotype, sex, and chip as covariates. Following this, the top NAT haplotype signal (chr2:216,636,519) was included as a covariate along with the rs78529201 genotype, sex, and array type, and each SNP was iteratively tested in a GLM with PAD as the outcome to identify an additional tag SNP.

Plots

All plots were produced using R (version 3.4.2), and ggplot2. Karyotype plots were generated using chromoMap (Anand and Rodriguez Lopez, 2022). GWAS Manhattan plots and QQ plots were made using qqman and gap R packages, respectively (Zhao, 2008; Turner, 2014).

Results

Prevalence of PAD in the diverse Bio*Me* biobank

We assessed the prevalence of PAD and associated risk factors in the diverse Bio*Me* biobank in NYC (N = 57,580). Overall PAD prevalence was ~6.5%, which was within the 5.8%–7.2% estimate reported by Allison et al. (2007) in a study of US adults >40 years of age in the year 2000 (Table 1). The prevalence of PAD was lower in women (54% in cases vs. 58% in controls, $p < 6.95 \times 10^{-6}$), and cases were generally older (median 73 years in cases vs. 59 in controls; $p < 2.22 \times 10^{-308}$). PAD cases also had a slightly higher BMI (28.2 in cases vs. 26.7 in controls; $p < 1.6 \times 10^{-51}$), and the majority also had a T2D diagnosis (66.9% in cases vs. 19.2% in controls; $p < 2.22 \times 10^{-308}$). PAD cases had lower TC (166 mg/dL in cases vs. 181 mg/dL in controls; $p < 6.17 \times 10^{-88}$) and HDL levels (47 mg/dL in cases vs. 53 mg/dL in controls; $p < 2.54 \times 10^{-99}$), but higher triglyceride levels (120 mg/dL in cases vs. 107 mg/dL in controls; $p < 9.11 \times 10^{-43}$). A greater proportion of PAD cases had a history of smoking (52% in cases vs. 34.8% in controls; $p < 5.2610^{-50}$). As expected, PAD prevalence varied across age groups, ranging from 0.3% in participants under the age of 40, 4.7% in participants aged between 40 and 69, and 13.7% in participants over the age of 70 ($p < 2.22 \times 10^{-308}$) (Allison et al., 2007).

We next assessed the prevalence of PAD across nine self-reported groups and eight genetic ancestry sub-groups in Bio*Me*. Genetic ancestry sub-groups were determined by the community membership based on unsupervised clustering of pairwise shared IBD haplotypes that represents recent common ancestry as described in Belbin et al. (2021). PAD prevalence was significantly different across both self-reported groups ($p < 5.33 \times 10^{-160}$) and genetically inferred sub-groups ($p < 3.63 \times 10^{-100}$). Within self-reported groups, we observe NAs as having the highest proportion of PAD cases (15%), although sample sizes are small. The group with the second highest PAD prevalence was HLs (9.4%), followed closely by AAs (8.5%). EAsn had the lowest PAD prevalence (1.7%) consistent with reporting elsewhere (Hackler et al., 2021). A similar pattern emerged within genetically defined sub-groups, and we observed the highest PAD prevalence in Dominican (11.5%) and Puerto Rican (11.4%) sub-groups followed by AAs (8.8%), similar to what we observed in the work of Belbin et al. (2021).

To evaluate which factors could be driving differences in observed PAD prevalence across population groups, we performed a PAD risk analysis. We evaluated known PAD risk factors, including age, biological sex, BMI, T2D, and lipid levels HDL-C, TG, and TC. Smoking history is another well-known risk factor for PAD; however, due to a high degree of missing data, smoking was not included in this multivariable analysis. We compared PAD odds in non-European ancestry to European self-reported groups (Supplementary Figure S1; Supplementary Table S2) and in genetic ancestry sub-groups (Figure 1; Supplementary Table S3). In an approach adopted from the work of Allison et al. (2015), we assessed three models to test for PAD risk, where each model added covariates adjusting for common PAD risk factors. Model 1 adjusted for age and sex, model 2 adjusted for model 1 and BMI, and model 3 adjusted for model 2 and T2D, TG, TC, and HDL.

All self-reported populations are significantly enriched for PAD compared to the EA group across all models except for EAsn. NAs had the highest odds ratio (OR) for PAD for all three models, and the OR

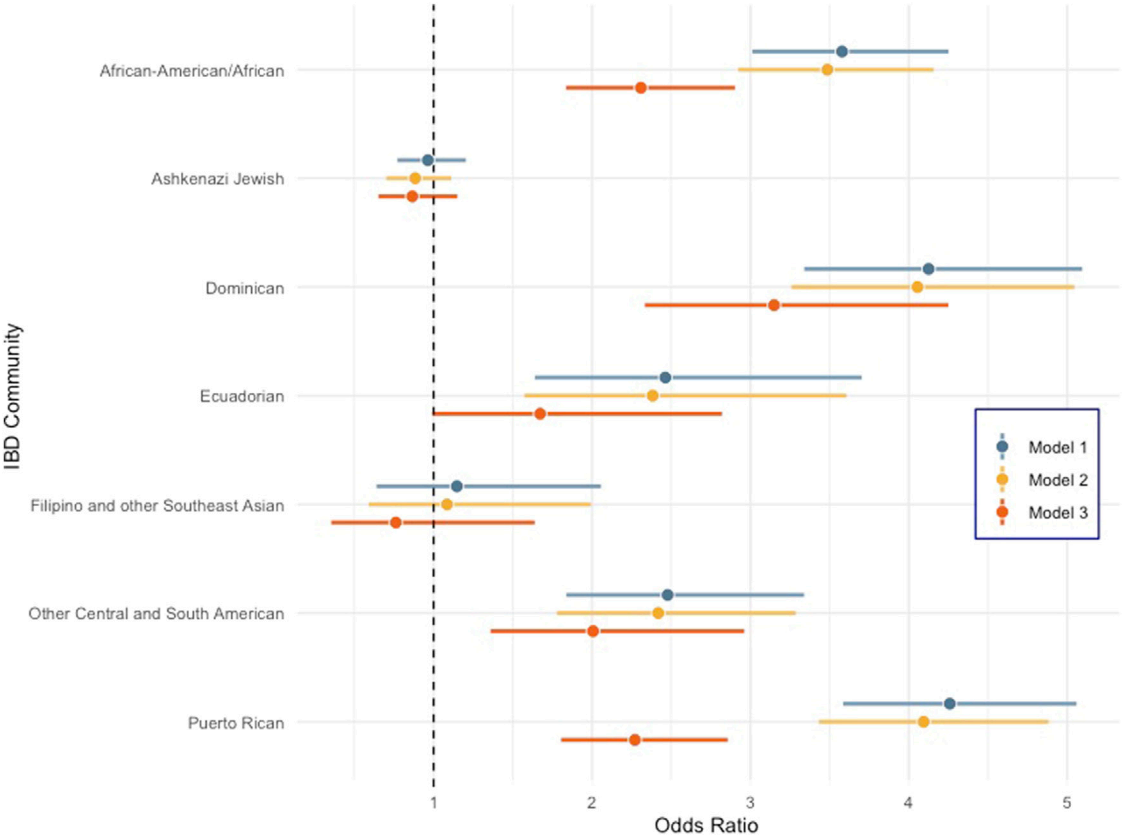

FIGURE 1
Forest plot comparing odds ratio of peripheral artery disease across diverse genetic ancestry groups in Bio*Me* compared to the non-Jewish European population. Error bars represent 95% confidence intervals. Model 1: PAD ~ population group + age + sex, model 2: PAD ~ model 1 + BMI, and model 3: PAD ~ model 2 + T2D + TG + TC + HDL.

in the model that included all covariates (model 3) was 6.14 (95% CI 2.75–13.69) with a $p < 9.33 \times 10^{-06}$ (Supplementary Figure S1). However, the 95% confidence interval of the OR overlaps with that observed for both HLs and AAs, so, it is not statistically different from these groups, likely due to the small sample size of NAs in Bio*Me*. Within IBD communities, all groups tested except "Filipino and other Southeast Asian" and "Ashkenazi Jewish" had increased odds of PAD cases relative to the "Non-Jewish Europeans." In model 3 when biomarkers and T2D status are included, we observed that Dominicans had the highest residual increase in PAD risk [OR = 3.15 (95% CI 2.33–4.25); $p = 6.44 \times 10^{-14}$] compared to any other genetic ancestry sub-group in Bio*Me* (Figure 1). We hypothesized that genetic risk factors common in Dominicans may be contributing to this increased PAD risk. Therefore, we decided to perform admixture mapping, which is optimally powered for genomic discovery when disease risk and/or disease-associated variants are enriched in a population.

Global and local genetic ancestry inference in Dominicans

The first step of admixture mapping is the accurate detection of genetic ancestry. First, we estimated individual-level or global genetic ancestry using ADMIXTURE, a model-based approach that when run in an unsupervised manner applies a pre-set number of putative ancestral populations to seek the best fit of ancestral clusters in the data. Hispanic populations, including Dominicans, primarily have genetic ancestry from European, African, and Native American populations, with minimal genetic contributions from Asian continental populations (Moreno-Estrada et al., 2013). Because of this, along with the limitations of local ancestry software in distinguishing East Asian and Native American haplotypes in Hispanics, mean Asian reference panels are not used for local ancestry inference in this study. ADMIXTURE analysis that was fit to three ancestral populations (K = 3) recapitulates three continental-level ancestral components corresponding to AFR, EUR, and NAT reference panels (Supplementary Figure S2). The median EUR and AFR genetic ancestry in the Dominican community were 56% [95% confidence intervals (95% CI) 55%–57%] and 37% (95% CI 36%–38%) at a population level, respectively. The majority of individuals harbored a median of 6% (95% CI 6.2%–6.5%) NAT genetic ancestry; however, a small minority of individuals (N = 126) harbored more than 10% (Supplementary Figure S3A.

To estimate haplotype-level genetic ancestry or LA, we used RFMix, a discriminative approach that estimates tracts of LA using conditional random fields parameterized with random forests

(Maples et al., 2013). We first phased the genotype data using Eagle v2.0 and then ran RFMix leveraging the same AFR, EUR, and NAT reference panels as used previously. The correlation between global ancestry estimates inferred using ADMIXTURE at K = 3 and local ancestry estimates inferred by RFMix was high for all three ancestral components (Pearson's correlation >0.97, Supplementary Figures S4A–C). Supplementary Figure S3B shows a karyogram painted with local ancestry tracts for one individual with ~56% EUR, 37% AFR, and 6% NAT genetic ancestry; however, patterns of LAI can differ substantially between Dominican individuals. Supplementary Figure S3C shows a plot of the tract length distribution for AFR, EUR, and NAT tracts. Previous work to estimate admixture timing has suggested a single pulse of NAT ancestry contributing to Dominican genetic ancestry that occurred at the time of European contact (Moreno-Estrada et al., 2013). This is consistent with what is seen in our analysis with NAT tracts being older and shorter with a median track length of 6.6 cM. EUR ancestry tracks were the longest on average (15.35 cM) followed by African tracts (11.31 cM).

Admixture mapping of peripheral artery disease in Dominicans

We next performed case-control admixture mapping to test whether regions of EUR, AFR, or NAT ancestry are associated with PAD risk. Cases for PAD (N = 245) were defined as having one or more billing codes for PAD (ICD-9 443.9), and controls (N = 1,568) were defined as having no PAD billing codes. Admixture mapping was performed using a GLM for each ancestry group separately, including sex and array type as covariates. To estimate the genome-wide significance threshold, we needed to account for the long-range correlation in local-ancestry linkage disequilibrium across the genome in the admixed Dominican community. We used the STEAM algorithm to estimate admixture proportions, generations since admixture, and genetic distances between loci in order to calculate the asymptotic joint distribution of the test statistic (Grinde et al., 2019). We estimated the genome-wide significance for admixture mapping in this population to be 5.282×10^{-6}. The top admixture mapping signal was an association between PAD and NAT ancestry at the chromosome 2q35 locus (chr2:216636519–216811790, build 37; $p < 2.75 \times 10^{-05}$, OR = 1.96, SE = 0.16; Figure 2), just below genome-wide significance. To determine whether this suggestive association might have been detected using a traditional GWAS approach, we performed GWAS in the same participants, but no GWS association was found (Figure 3) (Supplementary Figure S5 for Manhattan plots and Supplementary Figure S6 for the QQ plot).

Characterizing the 2q35 PAD-associated locus in Hispanic/Latino populations

The gold standard to validate suggestive genetic associations is to independently replicate them in other cohorts. However, we were unable to obtain an independent Dominican cohort with a sample size sufficiently powered to attempt replication of the 2q35 PAD-associated locus discovered in Bio*Me*. To test if the admixture mapping signal replicated broadly across HL groups, we repeated the same admixture mapping approach in self-report HL Bio*Me* participants (N = 6,801) and excluded the Dominican sub-group used in the discovery analysis. The top admixture mapping signal in the 2q35 region was significantly associated, however, in the opposite direction of effect, with NAT ancestry in the HL cohort being protective against PAD (OR = 0.7, SE = 0.069, $p < 5.87 \times 10^{-07}$). We also tested for the association of NAT tracts at the 2q35 locus in the HL participants in the MVP cohort (Gaziano et al., 2016; Hunter-Zinck et al., 2020). We found no significant admixture mapping association in this region (OR = 0.99, SE = 0.03, $p < 0.73$). Notably, the predominant HL sub-group in the Bio*Me* independent cohort is of Puerto Rican descent, whereas in MVP, the predominant sub-group is of Mexican descent. Therefore, it is possible that these findings support a role for the 2q35 PAD-associated locus in HL populations with ancestry from the Caribbean and not from other parts of the Americas.

Fine-mapping the 2q35 locus

Iterative conditional analysis within a 5 Mb window in this region was used for fine-mapping to identify a tag SNP within the NAT haplotype. Participants in the Dominican discovery cohort with both NAT haplotypes and imputed genotype calls were included (N = 1,687) in the analysis. A series of GLM were carried out with PAD as the outcome variable and local ancestry status at position 2:216636519 (build 37) (0 = no NAT ancestry, 1 = heterozygous NAT, and 2 = homozygous NAT), test tag SNP, genotype chip, and sex as predictor variables. The inclusion of the tag SNP in this conditional logistic regression with the top local ancestry association signal causes the ancestry haplotype significance to increase toward 0. Using this approach, we identified tag SNP rs78529201 (chr2:216518626, build GRCh 37), which attenuates the *p*-value of the local ancestry (from $P < 4 \times 10^{-05}$ to $p < 0.013$; Figure 4A). Though this SNP was able to explain much of the signal, it was not sufficient to fully explain the PAD-associated signal at 2q35. Therefore, the windowed conditional analysis was carried out again to find a likely second contributory tag SNP in the region. When the SNP rs77979649 (chr2:218477448, build 37) upstream of the primary tag SNP was included in a model with rs78529201, the admixture mapping signal was fully attenuated ($p < 0.4$). Neither SNP met genome-wide significance using the traditional GWAS approach, rs78529201 (OR = 3.66, SE = 0.4, $p < 1.36 \times 10^{-03}$) and rs77979649 (OR = 2.78, SE = 0.39, $p < 9.4 \times 10^{-03}$). Finally, we examined the frequency of both SNPs in the gnomAD browser v3.2.1 (Karczewski et al., 2020). They are both common (MAF > 5%) in the Latino/admixed American population, and rs77979649 is also common in the East Asian population and present in low frequency (MAF < 5%) or absent in other populations in the gnomAD. In the gnomAD Latino/admixed American group, they are at high frequency (MAF>25%) on Amerindigenous LA tracts and rare (MAF < 0.1%) on AFR and EUR LA tracts. This suggests that both tag SNPs are highly differentiated on NAT tracts in Bio*Me* Dominicans and the Amerindigenous LA tracts in the gnomAD Latino/admixed Americans population; however, neither SNP perfectly tags the NAT signal at 2q35 in Bio*Me* Dominicans.

The top tag SNP, rs78529201, falls within an intron of the lincRNA, *LINC00607*. Bulk tissue gene expression results from the Genotype-

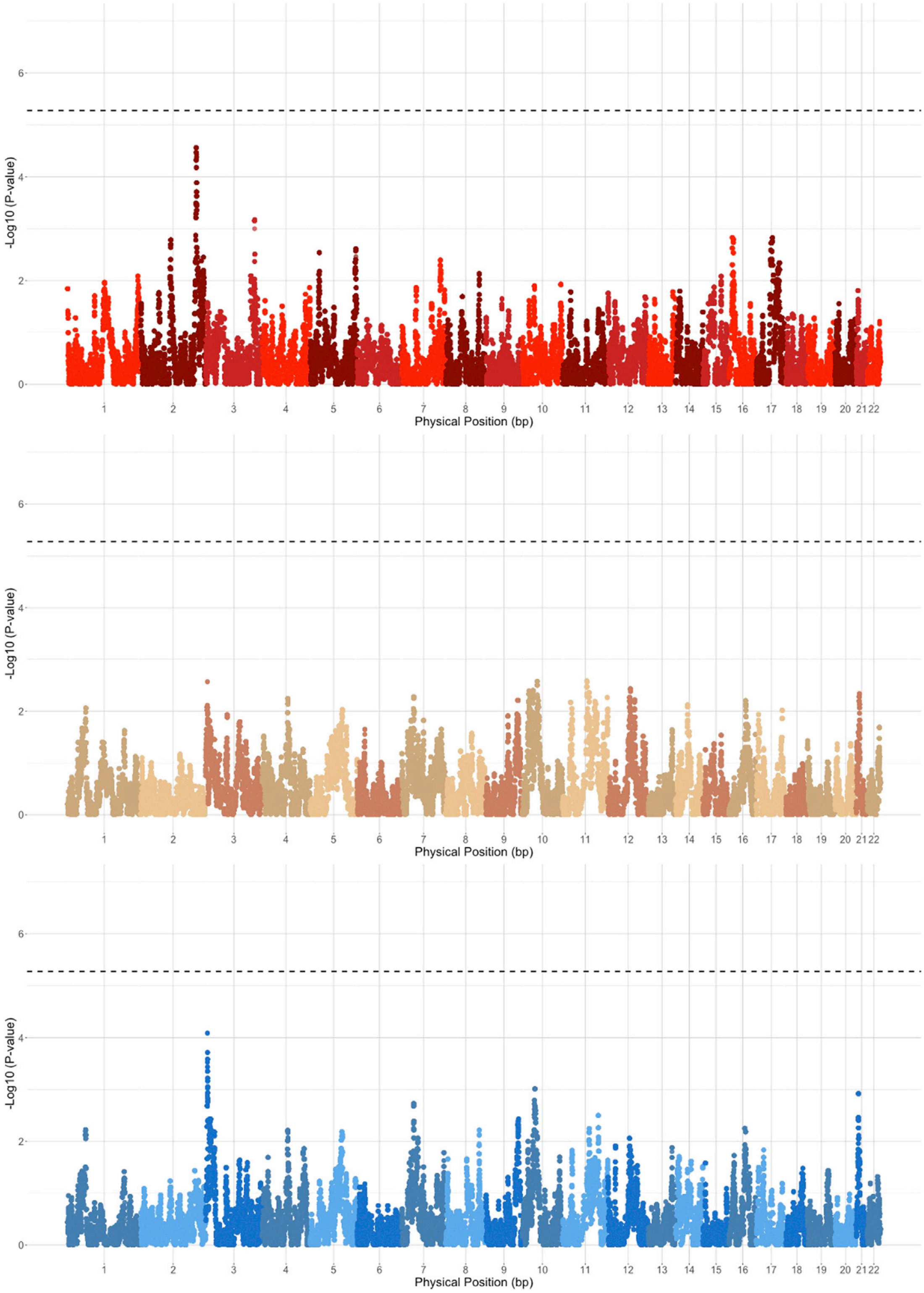

FIGURE 2
Admixture mapping results stratified based on EUR (blue), AFR (brown), and NAT (red) haplotypes. Genome-wide significance threshold is indicated by the dashed black line.

Tissue Expression portal (GTEx Portal, 2022) show the tissues with the highest expression of *LINC00607* are arteries (Sriram et al., 2022), and another recent study demonstrated that it is highly enriched in endothelial cells (Boos et al., 2022). Previous work has shown a key role of *LINC00607* in the control of cellular processes underlying inflammatory responses, angiogenesis, collagen catabolism, and

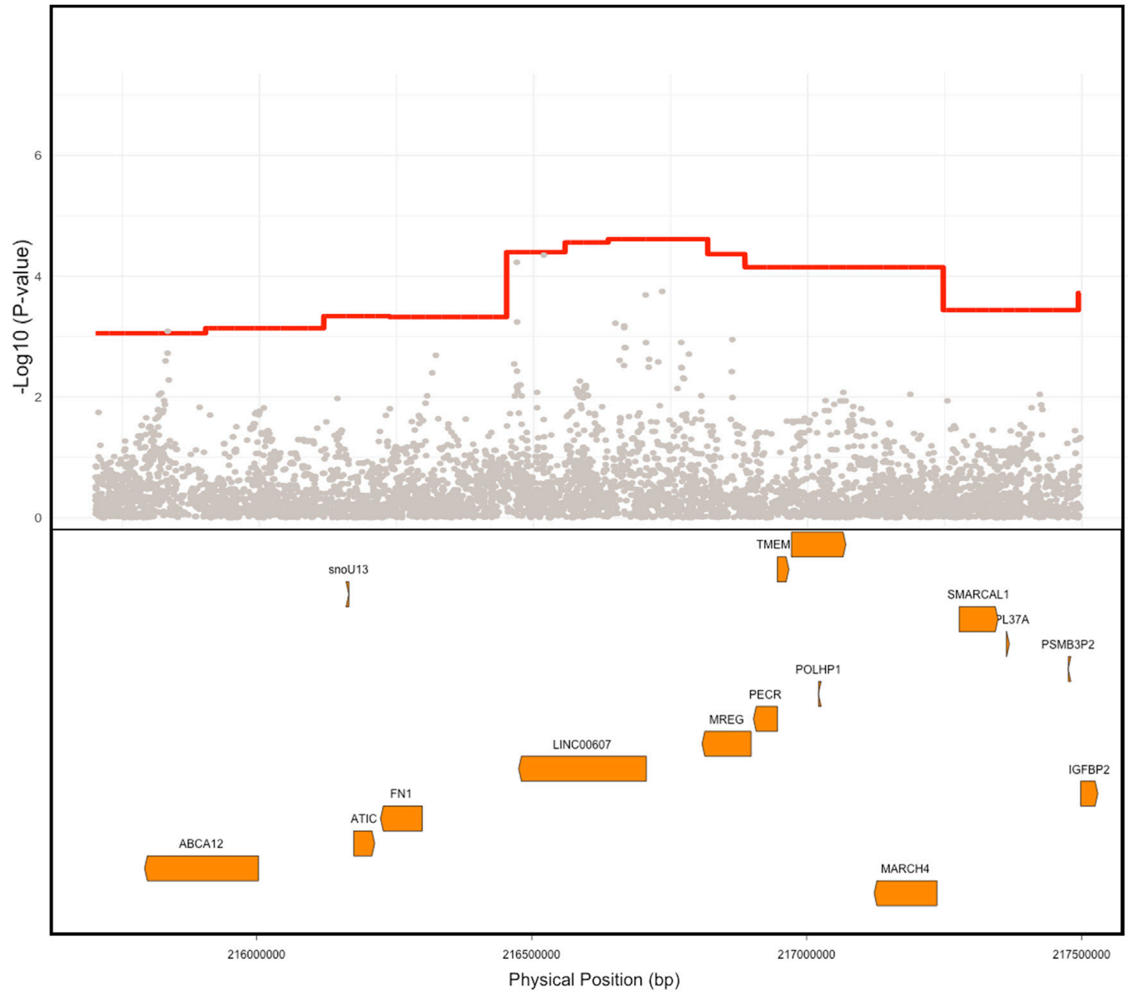

FIGURE 3
Comparison of ancestry signal Vs. SNP level signal. Native American (NAT) admixture mapping signal is depicted in red showing the PAD association signal compared to the individual SNP associations in gray. The location of genes within this region is depicted in orange in the lower box. Physical positions are build GRCh 37.

extracellular matrix organization in endothelial cells via its inactivation of *PAI-1*, an inhibitor of plasminogen activators (Calandrelli et al., 2019; Sriram et al., 2022). To assess the impact of harboring a risk allele on PAD risk, we showed that Dominican individuals who are heterozygous for SNP rs78529201 show a 39% risk of the disease compared to 13.5% of non-carriers (OR = 4.08, CI 95% 2.00–8.1, $p < 6.05 \times 10^{-5}$). No homozygous carriers are present in the Bio*Me* biobank for this variant. We also show that heterozygous carriers of an NAT haplotype in the 2q35 region carry a 22.6% risk of PAD compared to 12.9% for non-carriers (OR = 1.98, 95% CI 1.34–2.88, $p < 0.00045$) (Figure 4B). There are only 11 homozygous carriers of an NAT haplotype at rs76984916, prohibiting statistical comparison with non-carriers.

Discussion

In this study, we assessed clinical, demographic, and genetic factors underlying PAD risk in the diverse Bio*Me* biobank in NYC. Among genetic ancestry-defined sub-groups, Dominicans had the highest odds of PAD (3-fold compared to European ancestry Bio*Me* participants), when accounting for known clinical and demographic PAD risk factors. Local ancestry inference delineated haplotypes of recent AFR, EUR, and NAT continental ancestry in the Dominican group, and admixture mapping revealed a suggestive signal of association of NAT ancestry tracts on chromosome 2q35 linked to an almost 2-fold increase in PAD risk. Individuals who are heterozygous for NAT at the 2q35 locus have a 22.6% incidence of PAD in Bio*Me* compared to 12.9% in those with no NAT at that locus. Fine-mapping revealed a top tag SNP, rs78529201, which falls within an intron of the lincRNA, *LINC00607*. *LINC00607* has been previously shown to play a key role in the control of cellular processes underlying angiogenesis, extracellular matrix organization, and other vascular-related processes in endothelial cells. These findings highlight a previously under-appreciated risk for PAD and a putative genetic driver of increased PAD risk at chromosome 2q35 in Dominican populations in NYC.

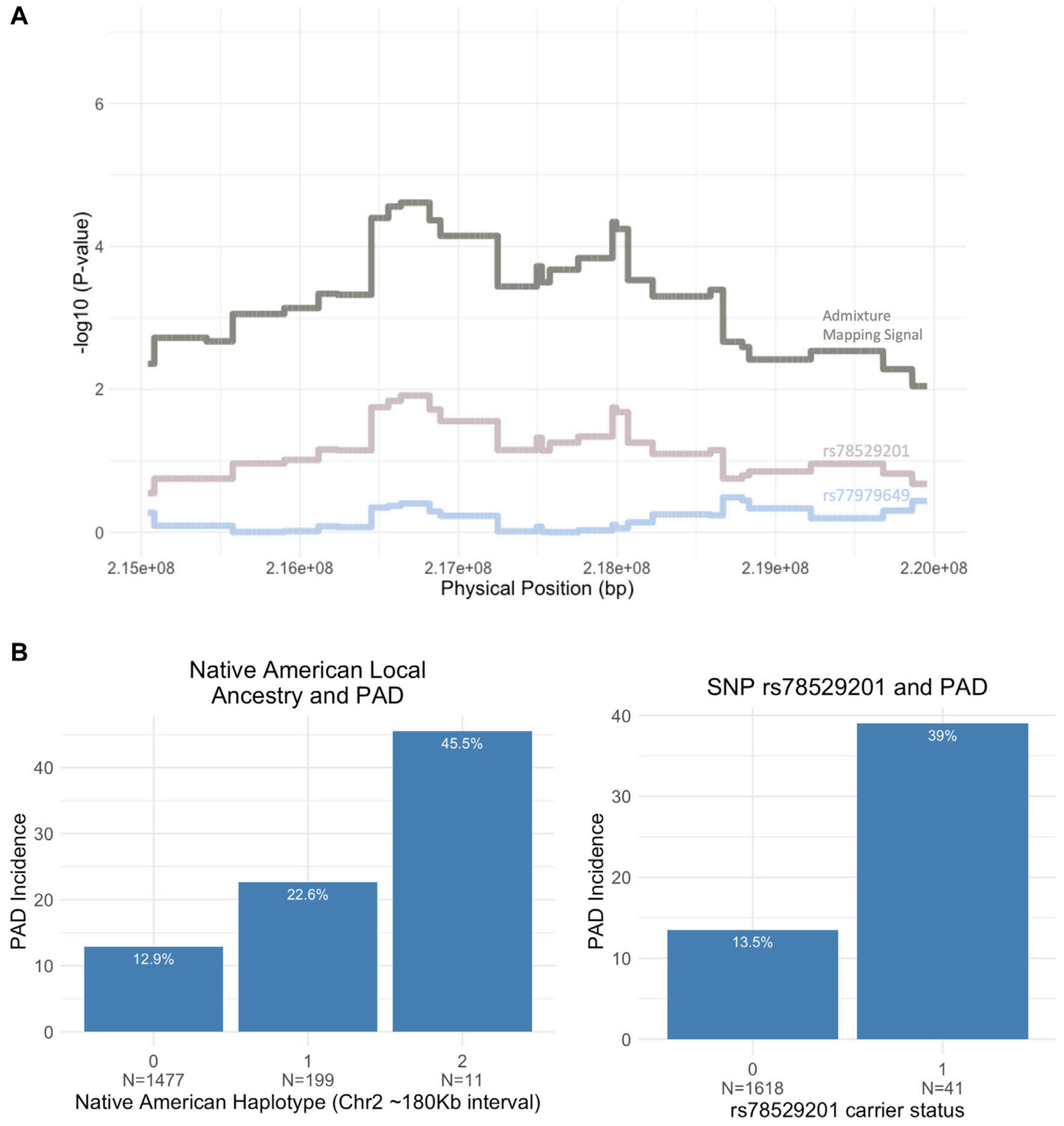

FIGURE 4
(A) Fine-mapping the 5 Mb region on chromosome 2 (215,000,000–220,000,000) using conditional analysis. Two tag SNPs, rs78529201 and rs77979649, increase the NAT admixture mapping signal toward 0 when included in the association model. **(B)** Bar plot showing the breakdown of PAD incidence by Native American (NAT) ancestry at admixture mapping peak (2:216,636,519) and rs78529201 carrier status (N = sample size).

Within self-reported populations in our study, we found that HLs and AAs had similar odds (~2.6-fold) of PAD compared to EAs. This finding differs from some previous studies which have reported AAs as having the greatest PAD risk among race and/or ethnicity groups in the US (Allison et al., 2006). HLs have been reported to have lower or similar rates of PAD compared to EAs, despite having a paradoxically high burden of PAD risk factors; however, the picture appears to be nuanced (Criqui et al., 2005; Forbang et al., 2014; Allison et al., 2015). Allison et al. reported Puerto Ricans and Dominicans have higher rates of PAD compared to Mexican Americans. Using genetic ancestry to identify HL sub-groups in NYC, we were able to explore the incidence of PAD in four sub-groups; three were significantly enriched compared to the European ancestry population, Puerto Ricans (2.3-fold), Dominicans (3.2-fold), and Central and South Americans groups (2-fold). Factors not measured in this study, including structural racism, socioeconomic status, disparities in healthcare access, detailed smoking habits, and statin use, discussed further by Hackler et al. (2021), are all likely contributing to differences in PAD incidence both at local and national levels. However, this study, along with work from the Hispanic Community Health Study/Study of Hispanics (Lavange et al., 2010) and the Multi-Ethnic Study of Atherosclerosis (Bild et al., 2002), demonstrates that HLs are culturally, socioeconomically, and genetically heterogeneous, and disease incidence rates may differ across HL sub-groups. We note that, although a small sample size, the self-reported NA group in

Bio*Me* had the highest PAD odds (6-fold), warranting further study of PAD risk factors in this population.

Admixture mapping of PAD in a Dominican population from NYC identified a suggestively associated region on chromosome 2q35. The risk locus spanned a 0.2 MB region on chromosome 2, where NAT genetic ancestry was associated with an increased risk for PAD. An association with NAT genetic ancestry at the same locus, but in the opposite direction, was demonstrated in the same biobank in an independent cohort of self-report HLs of predominantly Puerto Rican descent, and no association was observed in the HLs in the MVP, who are predominantly of Mexican descent (Gaziano et al., 2016). It is possible that the observed differences in PAD association at the 2q35 locus may be tied to patterns in NAT population structure in admixed populations from the Americas. Previous work has demonstrated the population substructure in the NAT ancestry components in HL populations with origins from Mexico (Moreno-Estrada et al., 2014) and the Caribbean (Moreno-Estrada et al., 2013) and genetic divergence between ancestral NAT populations (Reich et al., 2012). The interaction of the 2q35 with social or environmental factors impacting PAD may also modulate its impact across HL groups.

Fine-mapping of the associated locus revealed a top tag SNP residing in the intronic region of a lincRNA *LINC00607*. LincRNAs are known to play an important role in gene expression, usually in a cell type-specific manner, and have recently been shown to play an important role in complex diseases (Deniz and Erman, 2017). Recent work to characterize *LINC00607* expression and gene regulatory networks in vascular smooth muscle cells and endothelial cells found it to be an essential regulator of vascular cell function. Temporal changes to gene expression in endothelial cells mimicking diabetic conditions induced an intrachromosomal interaction between *LINC00607* and a super-enhancer overlapping the SERPINE1/PAI-1 gene (Calandrelli et al., 2020). *LINC00607* knockout reduced *SERPINE1* expression along with numerous other genes including *FN1*, *TRIO*, and *COL4* via a super-enhancer network, promoting endothelial cell dysfunction (Calandrelli et al., 2019). *LINC00607* is also implicated as a vital epigenetic regulator in arterial tissue and a potential target for CVD therapies (Sriram et al., 2022). Taken together, this supports a role for *LINC00607* as a gene expression regulator of key genes related to extracellular matrix organization, inflammation, angiogenesis, and endotheliopathy, suggesting a putative link of the 2q35 locus to PAD etiology.

There are several limitations to this study. First, cohorts with origins from the Dominican Republic are vastly underrepresented in genomic research databases (Estrada-Veras et al., 2016), and we could not access an independent cohort of Dominican ancestry of sufficient size and with the relevant phenotype to replicate our finding. Because this association did not surpass genome-wide significance, and without independent replication, we cannot rule out association due to biases (Kraft et al., 2009) or effect size estimate confounding due to winners' curses (Zou et al., 2022). Additional genetic studies of PAD in Dominican populations are needed to confirm the association, examine differences in LD and effect sizes, and assess any effect modifiers across studies. Second, deriving phenotype information from health systems data can result in selection bias and confounding that can affect case ascertainment and prevalence estimates (Haneuse and Daniels, 2016), although we note that the prevalence estimates reported here are in line with previously published estimates from large cohort studies. Third, we were not able to assess the impact of some known risk factors for PAD, such as smoking. Previous studies have demonstrated the odds ratio for symptomatic PAD in smokers is 2.3 (Willigendael et al., 2004). Ideally, the smoking status would have been included as a covariate in our models; however, smoking is not well captured in Bio*Me* questionnaire responses with >40% missing data and was excluded from the multivariable analysis due to low power. Finally, small sample sizes for individuals self-identifying as NA limited our ability to estimate prevalence and PAD relative odds in this group with strong confidence. Ongoing efforts, to enhance ethical genomic research with indigenous communities (Claw et al., 2018), studies focused on underrepresented populations like Human Heredity and Health in Africa (H3Africa Consortium et al., 2014), and large biobank initiatives that enrich for admixed and diverse populations, such as Bio*Me* (Belbin et al., 2021) and the All of Us Research Project (All of Us Research Program Investigators et al., 2019), will provide data that may improve power for discovery efforts in the future.

This work demonstrates how genomic discovery pipelines that leverage recent patterns of demography can be better powered to elucidate disease risk variants over traditional approaches. Furthermore, extensive population diversity often encountered in urban healthcare systems offers the potential to examine the generalizability of risk variants across populations and the interplay with clinical and social determinants of health and to identify populations at higher risk due to these factors. This is particularly important for complex conditions that are historically underdiagnosed. In the case of PAD, only 10%–30% of patients show intermittent claudication which is an important sequela of PAD. The Peripheral Arterial Disease Awareness Risk and Treatment: New Resources for Survival Study found that only 45% of a cohort with PAD had been diagnosed (Hirsch et al., 2001). Therefore, a better understanding of the genetics of the disease in diverse populations could help identify groups at higher risk for follow-up care and prevention and ensure equity and implementation of precision medicine globally.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by the IRB 07-0529. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

SC, GW, ES, CG, GB, NA-H, and EK participated in the study design. SC, GL, RS, DK, PT, BG, SP, GB, and EK participated in statistical analysis. SC, SMD, NA-H, and EK participated in phenotyping. SC and EK drafted the manuscript. GW, GB, SMD, SA, BG, ES, NA-H, and EK gave critical edits to the manuscript. All authors contributed to the article and approved the submitted version.

Funding

EK, NA-H, RS, and SC were supported by U01 HG009080 from the National Institutes of Health (NIH), the National Human Genome Research Institute (NHGRI), and the National Institute for Minority Heath and Health Disparities of the National Institutes of Health. EK, GB, and CG were supported by R01 HG011345 from the NHGRI. CG was also supported by R01 HG010297. SA was supported by the Sinsheimer Fund. DK was supported by the Department of Veterans Affairs Office of Research and Development, grant IK2BX005759-01. This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration, and was supported by award BX003362 and the Million Veteran Program-MVP000. SMD was supported by the US Department of Veterans Affairs Clinical Research and Development award IK2-CX001780. This publication does not represent the views of the Department of Veteran Affairs or the United States Government.

Acknowledgments

The authors are grateful to the patients and their families who contributed to this study.

Conflict of interest

EK received personal fees from Illumina, 23andMe, Allelica, and Regeneron Pharmaceuticals, received research funding from Allelica, and serves as a scientific advisory board member for Encompass Bio, Foresite Labs, and Galateo Bio. NA-H is an employee and equity holder of 23andMe; serves as a scientific advisory board member for Allelica; received personal fees from Genentech, Allelica, and 23andMe; received research funding from Akcea; and was previously employed by Regeneron Pharmaceuticals. ES is an employee of Calico Life Sciences LLC. SA received consulting fees from Variant Bio and Biogen Inc. CG owns stock in 23andMe and serves as an advisory board member for Encompass Bio. DK serve as scientific advisor to Bitterroot Bio, Inc. SMD receives research support from RenalytixAI and Novo Nordisk, outside the scope of the current research.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1181167/full#supplementary-material

References

Abul-Husn, N. S., Soper, E. R., Braganza, G. T., Rodriguez, J. E., Zeid, N., Cullina, S., et al. (2021). Implementing genomic screening in diverse populations. *Genome Med.* 13, 17. doi:10.1186/s13073-021-00832-y

Allison, M. A., Criqui, M. H., McClelland, R. L., Scott, J. M., McDermott, M. M., Liu, K., et al. (2006). The effect of novel cardiovascular risk factors on the ethnic-specific odds for peripheral arterial disease in the Multi-Ethnic Study of Atherosclerosis (MESA). *J. Am. Coll. Cardiol.* 48, 1190–1197. doi:10.1016/j.jacc.2006.05.049

Allison, M. A., Gonzalez, F., 2nd, Raij, L., Kaplan, R., Ostfeld, R. J., Pattany, M. S., et al. (2015). Cuban Americans have the highest rates of peripheral arterial disease in diverse Hispanic/Latino communities. *J. Vasc. Surg.* 62, 665–672. doi:10.1016/j.jvs.2015.03.065

Allison, M. A., Ho, E., Denenberg, J. O., Langer, R. D., Newman, A. B., Fabsitz, R. R., et al. (2007). Ethnic-specific prevalence of peripheral arterial disease in the United States. *Am. J. Prev. Med.* 32, 328–333. doi:10.1016/j.amepre.2006.12.010

Allison, M. A., Peralta, C. A., Wassel, C. L., Aboyans, V., Arnett, D. K., Cushman, M., et al. (2010). Genetic ancestry and lower extremity peripheral artery disease in the Multi-Ethnic Study of Atherosclerosis. *Vasc. Med.* 15, 351–359. doi:10.1177/1358863X10375586

All of Us Research Program InvestigatorsDenny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., et al. (2019). The "all of us" research Program. *N. Engl. J. Med.* 381, 668–676. doi:10.1056/NEJMsr1809937

American Diabetes Association (2003). Peripheral arterial disease in people with diabetes. *Diabetes Care* 26, 3333–3341. doi:10.2337/diacare.26.12.3333

Anand, L., and Rodriguez Lopez, C. M. (2022). ChromoMap: an R package for interactive visualization of multi-omics data and annotation of chromosomes. *BMC Bioinforma.* 23, 33. doi:10.1186/s12859-021-04556-z

1000 Genomes Project Consortium et al., 2015 1000 Genomes Project ConsortiumAuton A., Brooks L. D., Durbin R. M., Garrison E. P., Kang H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi:10.1038/nature15393

Belbin, G. M., Cullina, S., Wenric, S., Soper, E. R., Glicksberg, B. S., Torre, D., et al. (2021). Toward a fine-scale population health monitoring system. *Cell* 184, 2068–2083.e11. doi:10.1016/j.cell.2021.03.034

Belbin, G. M., Odgis, J., Sorokin, E. P., Yee, M.-C., Kohli, S., Glicksberg, B. S., et al. (2017). Genetic identification of a common collagen disease in puerto ricans via identity-by-descent mapping in a health system. *Elife* 6, e25060. doi:10.7554/eLife.25060

Bien, S. A., Wojcik, G. L., Zubair, N., Gignoux, C. R., Martin, A. R., Kocarnik, J. M., et al. (2016). Strategies for enriching variant coverage in candidate disease loci on a multiethnic genotyping array. *PLoS One* 11, e0167758. doi:10.1371/journal.pone.0167758

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., et al. (2002). Multi-ethnic study of atherosclerosis: Objectives and design. *Am. J. Epidemiol.* 156, 871–881. doi:10.1093/aje/kwf113

Boos, F., Oo, J. A., Warwick, T., Günther, S., Ponce, J. I., Buchmann, G., et al. (2022). The endothelial-specific LINC00607 mediates endothelial angiogenic function. *bioRxiv*. doi:10.1101/2022.05.09.491127

Calandrelli, R., Xu, L., Luo, Y., Wu, W., Fan, X., Nguyen, T., et al. (2019). Dynamic changes in RNA-chromatin interactome promote endothelial dysfunction. bioRxiv, 712950. doi:10.1101/712950

Calandrelli, R., Xu, L., Luo, Y., Wu, W., Fan, X., Nguyen, T., et al. (2020). Stress-induced RNA–chromatin interactions promote endothelial dysfunction. *Nat. Commun.* 11, 5211. doi:10.1038/s41467-020-18957-w

Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., et al. (2002). A human genome diversity cell line panel. *Science* 296, 261–262. doi:10.1126/science.296.5566.261b

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi:10.1186/s13742-015-0047-8

Claw, K. G., Anderson, M. Z., Begay, R. L., Tsosie, K. S., Fox, K., Garrison, N. A., et al. (2018). A framework for enhancing ethical genomic research with Indigenous communities. *Nature* 9, 2957. doi:10.1038/s41467-018-05188-3

Crawford, F., Welch, K., Andras, A., and Chappell, F. M. (2016). Ankle brachial index for the diagnosis of lower limb peripheral arterial disease. *Cochrane Database Syst. Rev.* 9, 10680. doi:10.1002/14651858.CD010680.pub2

Criqui, M. H., Vargas, V., Denenberg, J. O., Ho, E., Allison, M., Langer, R. D., et al. (2005). Ethnicity and peripheral arterial disease: The san diego population study. *Circulation* 112, 2703–2707. doi:10.1161/CIRCULATIONAHA.105.546507

Daviglus, M. L., Talavera, G. A., Avilés-Santa, M. L., Allison, M., Cai, J., Criqui, M. H., et al. (2012). Prevalence of major cardiovascular risk factors and cardiovascular diseases among Hispanic/Latino individuals of diverse backgrounds in the United States. *JAMA* 308, 1775–1784. doi:10.1001/jama.2012.14517

Deniz, E., and Erman, B. (2017). Long noncoding RNA (lincRNA), a new paradigm in gene expression control. *Funct. Integr. Genomics* 17, 135–143. doi:10.1007/s10142-016-0524-x

Elnady, B. M., and Saeed, A. (2017). Peripheral vascular disease: The beneficial effect of exercise in peripheral vascular diseases based on clinical trials. *Adv. Exp. Med. Biol.* 1000, 173–183. doi:10.1007/978-981-10-4304-8_11

Estrada-Veras, J. I., Cabrera-Peña, G. A., and Pérez-Estrella de Ferrán, C. (2016). Medical genetics and genomic medicine in the Dominican republic: Challenges and opportunities. *Mol. Genet. Genomic Med.* 4, 243–256. doi:10.1002/mgg3.224

Forbang, N. I., Hughes-Austin, J. M., Allison, M. A., and Criqui, M. H. (2014). Peripheral artery disease and non-coronary atherosclerosis in Hispanics: Another paradox? *Prog. Cardiovasc. Dis.* 57, 237–243. doi:10.1016/j.pcad.2014.07.008

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., et al. (2016). Million veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* 70, 214–223. doi:10.1016/j.jclinepi.2015.09.016

Grinde, K. E., Brown, L. A., Reiner, A. P., Thornton, T. A., and Browning, S. R. (2019). Genome-wide significance thresholds for admixture mapping studies. *Am. J. Hum. Genet.* 104, 454–465. doi:10.1016/j.ajhg.2019.01.008

GTEx Portal (2022). *GTEx portal*. Available at: https://gtexportal.org/home/ (Accessed December 8 2022).

H3Africa ConsortiumRotimi, C., Abayomi, A., Abimiku, A. 'le, Adabayeri, V. M., Adebamowo, C., et al. (2014). Research capacity. Enabling the genomic revolution in Africa. *Science* 344, 1346–1348. doi:10.1126/science.1251546

Hackler, E. L., 3rd, Hamburg, N. M., and White Solaru, K. T. (2021). Racial and ethnic disparities in peripheral artery disease. *Circ. Res.* 128, 1913–1926. doi:10.1161/CIRCRESAHA.121.318243

Haneuse, S., and Daniels, M. (2016). A general framework for considering selection bias in EHR-based studies: What data are observed and why? *EGEMS (Wash DC)* 4, 1203. doi:10.13063/2327-9214.1203

Hirsch, A. T., Criqui, M. H., Treat-Jacobson, D., Regensteiner, J. G., Creager, M. A., Olin, J. W., et al. (2001). Peripheral arterial disease detection, awareness, and treatment in primary care. *JAMA* 286, 1317–1324. doi:10.1001/jama.286.11.1317

Horimoto, A. R. V. R., Xue, D., Thornton, T. A., and Blue, E. E. (2021). Admixture mapping reveals the association between Native American ancestry at 3q13.11 and reduced risk of Alzheimer's disease in Caribbean Hispanics. *Alzheimers. Res. Ther.* 13, 122–214. doi:10.1186/s13195-021-00866-9

Hunter-Zinck, H., Shi, Y., Li, M., Gorman, B. R., Ji, S.-G., Sun, N., et al. (2020). Genotyping array design and data quality control in the million veteran Program. *Am. J. Hum. Genet.* 106, 535–548. doi:10.1016/j.ajhg.2020.03.004

Kannel, W. B., and McGee, D. L. (1985). Update on some epidemiologic features of intermittent claudication: The framingham study. *J. Am. Geriatr. Soc.* 33, 13–18. doi:10.1111/j.1532-5415.1985.tb02853.x

Kao, W. H. L., Klag, M. J., Meoni, L. A., Reich, D., Berthier-Schaad, Y., Li, M., et al. (2008). MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nat. Genet.* 40, 1185–1192. doi:10.1038/ng.232

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi:10.1038/s41586-020-2308-7

Kizil, C., Sariya, S., Kim, Y. A., Rajabli, F., Martin, E., Reyes-Dumeyer, D., et al. (2022). Admixture Mapping of Alzheimer's disease in Caribbean Hispanics identifies a new locus on 22q13.1. *Mol. Psychiatry* 27, 2813–2820. doi:10.1038/s41380-022-01526-6

Klarin, D., Lynch, J., Aragam, K., Chaffin, M., Assimes, T. L., Huang, J., et al. (2019). Genome-wide association study of peripheral artery disease in the Million Veteran Program. *Nat. Med.* 25, 1274–1279. doi:10.1038/s41591-019-0492-5

Klarin, D., Tsao, P. S., and Damrauer, S. M. (2021). Genetic determinants of peripheral artery disease. *Circ. Res.* 128, 1805–1817. doi:10.1161/CIRCRESAHA.121.318327

Kopp, J. B., Smith, M. W., Nelson, G. W., Johnson, R. C., Freedman, B. I., Bowden, D. W., et al. (2008). MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nat. Genet.* 40, 1175–1184. doi:10.1038/ng.226

Kraft, P., Zeggini, E., and Ioannidis, J. P. A. (2009). Replication in genome-wide association studies. *Stat. Sci.* 24, 561–573. doi:10.1214/09-STS290

Kullo, I. J., Shameer, K., Jouni, H., Lesnick, T. G., Pathak, J., Chute, C. G., et al. (2014). The ATXN2-SH2B3 locus is associated with peripheral arterial disease: An electronic medical record-based genome-wide association study. *Front. Genet.* 5, 166. doi:10.3389/fgene.2014.00166

Lavange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., et al. (2010). Sample design and cohort selection in the hispanic community health study/study of latinos. *Ann. Epidemiol.* 20, 642–649. doi:10.1016/j.annepidem.2010.05.006

Leeper, N. J., Kullo, I. J., and Cooke, J. P. (2012). Genetics of peripheral artery disease. *Circulation* 125, 3220–3228. doi:10.1161/CIRCULATIONAHA.111.033878

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., et al. (2016). Reference-based phasing using the haplotype reference Consortium panel. *Nat. Genet.* 48, 1443–1448. doi:10.1038/ng.3679

Mahoney, E. M., Wang, K., Cohen, D. J., Hirsch, A. T., Alberts, M. J., Eagle, K., et al. (2008). One-year costs in patients with a history of or at risk for atherothrombosis in the United States. *Circ. Cardiovasc. Qual. Outcomes* 1, 38–45. doi:10.1161/CIRCOUTCOMES.108.775247

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93, 278–288. doi:10.1016/j.ajhg.2013.06.020

Matsukura, M., Ozaki, K., Takahashi, A., Onouchi, Y., Morizono, T., Komai, H., et al. (2015). Genome-wide association study of peripheral arterial disease in a Japanese population. *PLoS One* 10, e0139262. doi:10.1371/journal.pone.0139262

Matsushita, K., Sang, Y., Ning, H., Ballew, S. H., Chow, E. K., Grams, M. E., et al. (2019). Lifetime risk of lower-extremity peripheral artery disease defined by ankle-brachial index in the United States. *J. Am. Heart Assoc.* 8, e012177. doi:10.1161/JAHA.119.012177

Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Sikora, M., Contreras, A. V., et al. (2014). Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* 344, 1280–1285. doi:10.1126/science.1251688

Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J. L., Byrnes, J. K., Gignoux, C. R., et al. (2013). Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* 9, e1003925. doi:10.1371/journal.pgen.1003925

Pacheco, J., and Thompson, W. (2012). *Type 2 diabetes mellitus*. USA: Northwestern University. PheKB.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi:10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., et al. (2012). Reconstructing Native American population history. *Nature* 488, 370–374. doi:10.1038/nature11258

Scherer, M. L., Nalls, M. A., Pawlikowska, L., Ziv, E., Mitchell, G., Huntsman, S., et al. (2010). Admixture mapping of ankle-arm index: Identification of a candidate locus associated with peripheral arterial disease. *J. Med. Genet.* 47, 1–7. doi:10.1136/jmg.2008.064808

Shu, J., and Santulli, G. (2018). Update on peripheral artery disease: Epidemiology and evidence-based facts. *Atherosclerosis* 275, 379–381. doi:10.1016/j.atherosclerosis.2018.05.033

Sriram, K., Luo, Y., Yuan, D., Malhi, N. K., Tapia, A., Samara, V. A., et al. (2022). Vascular regulation by super enhancer-derived LINC00607. *Front. Cardiovasc Med.* 9, 881916. doi:10.3389/fcvm.2022.881916

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 452, 638–642. doi:10.1038/nature06846

Turner, S. D. (2014). *qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots*. bioRxiv. doi:10.1101/005165005165

Wassel, C. L., Loomba, R., Ix, J. H., Allison, M. A., Denenberg, J. O., and Criqui, M. H. (2011). Family history of peripheral artery disease is associated with prevalence and severity of peripheral artery disease: The san diego population study. *J. Am. Coll. Cardiol.* 58, 1386–1392. doi:10.1016/j.jacc.2011.06.023

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi:10.1093/bioinformatics/btq340

Willigendael, E. M., Teijink, J. A. W., Bartelink, M.-L., Kuiken, B. W., Boiten, J., Moll, F. L., et al. (2004). Influence of smoking on incidence and prevalence of peripheral arterial disease. *J. Vasc. Surg.* 40, 1158–1165. doi:10.1016/j.jvs.2004.08.049

Winkler, C. A., Nelson, G. W., and Smith, M. W. (2010). Admixture mapping comes of age. *Annu. Rev. Genomics Hum. Genet.* 11, 65–89. doi:10.1146/annurev-genom-082509-141523

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 570, 514–518. doi:10.1038/s41586-019-1310-4

Zhao, J. H. (2008). gap: Genetic analysis package. *J. Stat. Softw.* 23, 1–18. doi:10.18637/jss.v023.i08

Zou, J., Zhou, J., Faller, S., Brown, R. P., Sankararaman, S. S., and Eskin, E. (2022). Accurate modeling of replication rates in genome-wide association studies by accounting for Winner's Curse and study-specific heterogeneity. *G3* 12, jkac261. doi:10.1093/g3journal/jkac261

Glossary

PAD	Peripheral artery disease
AA	African American
EA	European American
HL	Hispanic/Latino
NYC	New York City
US	United States
LA	Local ancestry
EUR	European
AFR	African
NAT	Native American
GWAS	Genome-wide association study
MVP	Million Veteran Program
GWS	Genome-wide significant
SNP	Single-nucleotide polymorphism
lincRNA	Long intergenic non-coding RNA
EAsn	East/South-East Asian
SA	South Asian
NA	Native American
OMNI	Omni-Express Array
MEGA	Multi-Ethnic Genotype Array
IBD	Identical by descent
PPV	Positive predictive value
EHR	Electronic Health Record
ICD	International Classification of Diseases
T2D	Type 2 diabetes
TG	Triglyceride
HDL	High-density lipoprotein
TC	Total cholesterol
BMI	Body mass index
GLM	Generalized linear model
1KGP	1000 Genomes Project
HDGP	Human Diversity Genome Project
PAGE	Polygenic Architecture using Genetics and Epidemiology
CEU	Utah residents with Northern and Western European ancestry
IBS	Iberian populations in Spain
YRI	Yoruba in Ibadan, Nigeria
LWK	Luhya in Webuye, Kenya
MAF	Minor allele frequency
LD	Linkage disequilibrium
GnomAD	Genome Aggregate Database
PCs	Principal components
OR	Odds ratio