Journal of the American Heart Association

ORIGINAL RESEARCH

Unveiling the Genetic Landscape of Coronary Artery Disease Through Common and Rare Structural Variants

Kruthika R. Iyer, ScD; Shoa L. Clarke, MD, PhD; Rodrigo Guarischi-Sousa, PhD; Ketrin Gjoni, BS; Adam S. Heath, MS; Erica P. Young, MD; Nathan O. Stitziel, MD, PhD; Cecelia Laurie, PhD; Jai G. Broome, BA; Alyna T. Khan, BA; Joshua P. Lewis, PhD; Huichun Xu, MD, PhD; May E. Montasser, PhD, MS; Kellan E. Ashley, MD; Natalie R. Hasbani, MPH; Eric Boerwinkle, PhD; Alanna C. Morrison, PhD; Nathalie Chami, PhD; Ron Do, PhD; Ghislain Rocheleau, PhD; Donald M. Lloyd-Jones, MD; Rozenn N. Lemaitre, PhD; Joshua C. Bis, PhD; James S. Floyd, MD, MS; Gregory L. Kinney, PhD, MPH; Donald W. Bowden, PhD; Nicholette D. Palmer, PhD; Emelia J. Benjamin, MD, ScM; Matthew Nayor, MD, MPH; Lisa R. Yanek, MPH; Brian G. Kral, MD, MPH; Lewis C. Becker, MD; Sharon L. R. Kardia, PhD; Jennifer A. Smith, PhD, MPH; Lawrence F. Bielak, DDS, MPH; Arnita F. Norwood, PhD; Yuan-I Min, PhD; April P. Carson, PhD; Wendy S. Post, MD, MS; Stephen S. Rich, PhD; David Herrington, MD; Xiuqing Guo, PhD; Kent D. Taylor, PhD; JoAnn E. Manson, MD, DrPH; Nora Franceschini, MD, MPH; Katherine S. Pollard, PhD; Braxton D. Mitchell, PhD, MPH; Ruth J. F. Loos, PhD; Myriam Fornage, PhD; Lifang Hou, PhD; Bruce M. Psaty, MD, PhD; Kendra A. Young, PhD, MSPH; Elizabeth A. Regan, PhD; Barry I. Freedman, MD; Ramachandran S. Vasan, MD; Daniel Levy, MD; Rasika A. Mathias, ScD; Patricia A. Peyser, PhD; Laura M. Raffield, PhD; Charles Kooperberg, PhD; Alex P. Reiner, MD, MSc; Jerome I. Rotter, MD; Goo Jun, PhD; Paul S. de Vries, PhD; Themistocles L. Assimes, MD, PhD

BACKGROUND: Genome-wide association studies have identified several hundred susceptibility single nucleotide variants for coronary artery disease (CAD). Despite single nucleotide variant-based genome-wide association studies improving our understanding of the genetics of CAD, the contribution of structural variants (SVs) to the risk of CAD remains largely unclear.

METHOD AND RESULTS: We leveraged SVs detected from high-coverage whole genome sequencing data in a diverse group of participants from the National Heart Lung and Blood Institute's Trans-Omics for Precision Medicine program. Single variant tests were performed on 58 706 SVs in a study sample of 11 556 CAD cases and 42 907 controls. Additionally, aggregate tests using sliding windows were performed to examine rare SVs. One genome-wide significant association was identified for a common biallelic intergenic duplication on chromosome 6q21 (P=1.54E-09, odds ratio=1.34). The sliding window-based aggregate tests found 1 region on chromosome 17q25.3, overlapping *USP36*, to be significantly associated with coronary artery disease (P=1.03E-10). *USP36* is highly expressed in arterial and adipose tissues while broadly affecting several cardiometabolic traits.

CONCLUSIONS: Our results suggest that SVs, both common and rare, may influence the risk of coronary artery disease.

Key Words: association testing ■ coronary artery disease ■ genetics ■ structural variants ■ whole-genome sequencing

See Editorial by Chen and Stewart

Correspondence to: Themistocles L. Assimes, MD, PhD, Stanford University School of Medicine, 3801 Miranda Ave., Palo Alto, CA 94304. Email: tassimes@stanford.edu

This article was sent to Barry London, MD, PhD, Senior Guest Editor, for review by expert referees, editorial decision, and final disposition.

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.124.036499

For Sources of Funding and Disclosures, see page 13.

JAHA is available at: www.ahajournals.org/journal/jaha

RESEARCH PERSPECTIVE

What Is New?

- This study represents the first large-scale multi-ancestry association study between structural variants identified through whole genome sequencing and coronary artery disease.
- The study identifies candidate common and rare structural variants associated with coronary artery disease in an ancestrally diverse cohort some of which include genes implicated in cardiometabolic risk factors for coronary artery disease.

What Question Should Be Addressed Next?

- Larger studies with enhanced calling of structural variants using either short- or long-read sequencing technologies are needed to replicate and extend these findings.

Nonstandard Abbreviations and Acronyms

CNV	copy number variants
PCs	principal components
SV	structural variants
TOPMed	Trans-Omics for Precision Medicine program
WGS	whole-genome sequencing

Coronary artery disease (CAD), characterized by the accumulation of atherosclerotic plaque in blood vessels, is the leading cause of mortality worldwide.[1] A complex interplay between genetic and lifestyle factors influences the development of CAD. Genome-wide studies have uncovered the genetic basis of CAD using single nucleotide variants (SNVs), identifying >300 loci associated with coronary atherosclerosis.[2–5] Despite heritability estimates ranging from 40% to 60%, the identified loci explain only a fraction of the overall disease heritability.[2,4–6] The missing heritability may be at least partially attributed to another class of genetic variation known as structural variants (SVs).[7]

SVs refer to large-scale alterations in chromosome structure, ranging in length from approximately 50 base pairs to multiple megabases, making them the most extensive class of genomic variation.[8] They include various subclasses such as unbalanced copy number variants (CNVs) involving deletions and duplications of genetic material, as well as balanced rearrangements like inversions.[8] The impact of SVs extends to gene dosage alterations, thereby influencing phenotypic variations.[9,10] Nonetheless, their contributions to human diseases to date have been predominantly studied in the context of congenital anomalies[11] and severe neurological[12–14] or psychiatric disorders.[15] The potential of SVs to act as modulators of genetic susceptibility to other complex disease, like CAD, remains inadequately investigated. This can be partly attributed to the rarity of randomly occurring SVs in the general population[8] and the complexity involved in identifying them accurately.[16] Very few studies in the past 15 years have examined the association between atherosclerotic phenotypes and structural variants.[17–22] Early studies with limited sample sizes and low-resolution array platforms, which could detect only large CNVs (typically >100 kb), failed to identify SVs associated with CAD.[17–21] More recent array-based CNV association studies performed by biobanks have identified genome-wide significant CNV loci for CAD, but given that only large CNVs could be detected, these studies primarily focused on a small fraction of SVs, leaving the majority unexplored.[22]

Using sequencing data, SVs as small as 50 base pairs can be accurately called, and this can be inclusive of all categories of SVs without limiting to only CNVs.[16,23] We aimed to build on previous efforts by conducting an SV association study with CAD using deep whole-genome sequencing (WGS) data from 54 463 individuals of diverse ancestries participating in the National Heart Lung and Blood Institute's Trans-Omics for Precision Medicine (TOPMed) program.[24]

METHODS

Data Availability

An overview of the TOPMed participant consent and data access procedures is provided at https://topmed.nhlbi.nih.gov/topmed-data-access-scientific-community. Participant-level genotype and phenotype data are available to approved investigators via the database of Genotypes and Phenotypes. The database accession numbers for all TOPMed studies referenced in this paper are listed in Table S1. Additionally, genomic summary results pertaining to this study are available at phs001974, as detailed at https://topmed.nhlbi.nih.gov/topmed-genomic-summary-results-public.

TOPMed Analytical Samples

The National Heart Lung and Blood Institute-sponsored TOPMed program, comprising of several preexisting studies that ascertained participants on heart, lung, blood, and sleep disorders, represents a substantial and expanding resource, currently encompassing 217 936 participants. However, only a fraction of these

contributing studies have CAD data. We included 54 463 participants from 13 TOPMed studies: Genetics of Cardiometabolic Health in the Amish (n=1088), ARIC (Atherosclerosis Risk in Communities Study, n=3789), BioMe (Mount Sinai BioMe Biobank, n=8718), CARDIA (Coronary Artery Risk Development in Young Adults, n=3044), CHS (Cardiovascular Health Study, n=2936), COPDGene (Genetic Epidemiology of Chronic Obstructive Pulmonary Disease Study, n=9908), DHS (Diabetes Heart Study, n=334), FHS (Framingham Heart Study, n=3912), GeneSTAR (Genetic Studies of Atherosclerosis Risk, n=1365), GENOA (Genetic Epidemiology Network of Arteriopathy, n=1211), JHS (Jackson Heart Study, n=3270), MESA (Multi-Ethnic Study of Atherosclerosis, n=4466), and WHI (Women's Health Initiative, n=10 422). Further descriptions of the design of the participating TOPMed cohorts and the sampling of individuals within each cohort for TOPMed WGS are provided elsewhere.[24] All studies were approved by the appropriate institutional review boards and informed consent was obtained from all participants (Data S1, "Brief Information about the TOPMed studies included in this paper").

Phenotype Harmonization

Incident and prevalent cases were defined as individuals with acute myocardial infarction, coronary revascularization, or coronary heart disease death as adjudicated by the respective cohorts. Controls were defined as noncases who additionally did not have angina or "possible" coronary heart disease death. Harmonization of outcomes across cohorts was conducted by the TOPMed data coordinating center in consultation with 2 coauthors trained in adult cardiology (S.L.C., T.L.A.). Complete details of this harmonization process have been published elsewhere[25] and are also available at https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes.

TOPMed Whole-Genome Sequencing

WGS was performed to an average depth of 38× using DNA isolated from blood, polymerase chain reaction-free library construction, and Illumina HiSeq X-Ten technology with a 150 bp paired end reads. Sequencing was performed across seven centers (Broad Genomics, University of Washington Northwest Genomics Center, New York Genome Center, Illumina Genomic Services, Macrogen Corp., Baylor College of Medicine Human Genome Sequencing Center, and McDonnell Genome Institute) (Table S1) where the reads were aligned to the build 38 human reference genome (GRCh38). Details for variant calling and quality control are described on the TOPMed website.[26] Briefly, SNVs/small indels discovery and genotype calling were performed jointly across all the available TOPMed Freeze 9 studies using the GotCloud pipeline[27,28] resulting in a single multistudy genotype call set, and variant filtering was done using support vector machine classifier.

Structural Variants in TOPMed

SVs in TOPMed were available through SV Freeze 1.1 on 138 134 diverse human genomes across 44 studies. SV calling and genotyping have been described elsewhere.[29] Briefly, SV calling was done by deploying the Parliament2 pipeline in which SVs were identified using a multitool ensemble of SV callers like Manta,[30] Delly,[31] Lumpy,[32] BreakSeq2,[33] and CNVnator.[34] SV calls were then integrated using SURVIVOR,[35] filtered using SVTyper,[36] and genotyped using muCNV.[37] The TOPMed SV Freeze 1.1 call set consisted of a total of 466 455 SV sites on 22 autosomes including 231 817 deletions, 197 412 duplications and 37 226 inversions, and the genotypes of each of these SVs were represented in a bi-allelic genotype format (Genotype=0/0, 0/1, 1/1), similar to SNVs and small indels generated from the same WGS data

STATISTICAL ANALYSIS

Single SV Test for Association

We performed genome-wide tests for association of SV using a mixed effects model approach implemented in the R Bioconductor package GENESIS[38] with CAD status as the outcome of interest. First, a null model was fit adjusting for several fixed effects including age at baseline, sex, TOPMed study, and a variable indicating sequencing center. Adjustments for the first 10 ancestry-informative principal components (PCs) were also made. PCs for our analytic sample were derived using the WGS SNV genotypes on PC-AiR.[39] Subsequently, K-means clustering was applied to the PCs data to categorize individuals into European, African, Admixed American, East Asian, and South Asian genetic ancestry groups (Figure S1). To account for technical sources of variability arising from the sequencing process, we additionally included the first 10 read depth-based principal components (batch-PCs) in our model. For the estimation of batch-PCs, average sequencing depth for every 1000 bp region ("bin") across the genome was computed using Mosdepth.[40] We removed bins known to be problematic including those containing highly repetitive DNA sequence with poor mapping (mappability<1.0 based on 50 bp k-mers in GEMTools v1.75945).[41] To avoid overcorrecting for sex, bins were limited to autosomes. After normalizing ~150 000 remaining bins, we performed Randomized Singular Value Decomposition,[42] a scalable alternative to PC analysis, to generate batch-PCs. To account for genetic relatedness among participants, a

random effect with covariance matrix proportional to the fourth degree sparse empirical kinship matrix computed using PC-Relate[39] was included in the model (Data S1, "TOPMed Freeze 9 Relatedness Analysis"). The output of the null model was then used to perform genome-wide Score tests of genetic association for all SVs that passed the TOPMed SV quality filters, had a site missingness of <50%, and had a minor allele count≥10 (Table S6). Significance was evaluated after estimating Bonferroni threshold (P<8.52E-07) and a secondary threshold based on false discovery rate (P<1.03E-06), accounting for the number of independent SVs (N=58 706) across the genome.

Chromatin Interaction Map Prediction

We used Akita,[43,44] a deep learning model based on neural networks, to predict 3-dimensional genome contact matrices solely from DNA sequence data, enabling the evaluation of chromatin perturbations induced by SVs. Both the reference sequence and the alternate sequence, encompassing a 1 Mb region centered around the SV events, were input into the model. Subsequently, contact maps were generated for both the reference and alternate sequences. Additionally, a subtraction matrix was constructed to depict alterations in chromatin disruption between the 2 sequences.

Aggregate Tests for Rare SVs

We conducted a series of aggregate tests using a sliding-window approach to investigate the association between rare SVs and CAD with the R Bioconductor GENESIS package.[38] The baseline test was a burden test with null weights, which assumes the same directionality of effects within each window and assigns equal weight to all variants. We also applied (1) the burden test with Madsen-Browning[45] weights, where the weighting scheme emphasizes on rarer SVs by assigning weights inversely proportional to their allele frequency; (2) Sequence Kernel Association Test[46] with Wu weights, which is designed to handle variants with opposing effect directions within windows. The Wu weights are derived from the beta distribution, giving higher weightage to rare SVs; and finally (3) the Variant Set Mixed Model Association Test[47], a robust hybrid of the Sequence Kernel Association Test and the burden test, allowing it to accommodate both homogeneous and heterogeneous variant effect directions within a window. The Variant Set Mixed Model Association Test was also implemented with Wu weights to prioritize rare SVs. To decrease the number of tests and to avoid spurious associations, we included only SVs that passed the TOPMed SV quality filters and had a site missingness of <50%. To expand the scope of rare SV analyzed, we stipulated a minor allele frequency of <=1% while also ensuring a minor allele count>=5. We defined a 100 kb sliding window with a bin size of 30 kb across the entire genome. The null model from single variant test was used to perform aggregate tests incorporating the Score statistic, Q statistic, and a combined approach that integrates both the Score and Q statistic. Significance was evaluated after a Bonferroni correction for 54 523 windows was made (P<0.05/54523=9.17E-07).

SV Look-Up

For thorough investigation of the hits we obtained from our single SV association tests and rare SV aggregate tests, we queried various databases including the Genome-Wide Association Studies (GWAS) catalog,[48] the Common Metabolic Diseases Knowledge Portal[49] for phenome-wide association studies, TOP-LD,[50] AnnotSV,[51] ClinVar,[52] deCode,[53] ENCODE,[54] GTex[55] and the 1000 Genomes Project Phase 3 structural variant data set.[56]

RESULTS

We conducted a comprehensive examination of 54 463 subjects enrolled in 13 TOPMed studies. In this cohort, approximately one third were identified as cases of CAD; these cases exhibited an approximate age difference of 11 years compared with our controls (Table 1). In this group of cases, 59% self-identified themselves as White, 19% as Black, 14% as Hispanic, and the remaining as Asian.

We examined a total of 58 706 SVs in these individuals with the smallest variant measuring 11 bp, and an average SV length of 18.9 kb (Figure S2). Approximately 75% were rare (minor allele frequency<1%) whereas only 12.5% of the variants were categorized as common. Notably, over half of these variants were identified as deletions with minor allele frequency<1% (Table 2).

Table 1. Sample Characteristics

	Cases	Controls
Total (N=54 463)	11 556	42 907
Incident	5 184	
Prevalent	6 372	
% male sex	35	44
Age,y, mean±SD	66±13	55±15
Self-reported race or ethnicity		
White	6 842	25 303
Black	2 175	12 978
Hispanic/Latino	1 580	2 788
East Asian	101	857
South Asian	121	96

Table 2. Structural Variants Distribution by Minor Allele Frequency and Type

	MAF<1%	1%<=MAF<5%	MAF>=5%	Total
Deletion	31 170 (53.10%)	6 454 (10.99%)	6 512 (11.09%)	44 136 (75.18%)
Duplication	7 924 (13.5%)	603 (1.03%)	646 (1.10%)	9 173 (15.63%)
Inversion	4 923 (8.39%)	327 (0.56%)	147 (0.25%)	5 397 (9.19%)
Total	44 017 (74.98%)	7 384 (12.58%)	7 305 (12.44%)	58 706

MAF indicates minor allele frequency.

The QQ plot and genomic inflation factor of our genome-wide association study between SVs and CAD suggested a well-calibrated analysis with little evidence of inflation from population stratification (λ=1.034) (Figure S3). Subsequent stratification of the QQ plots based on allele frequency also revealed minimal to no evidence of inflation (λ ranging from 1.031 to 1.053; Figure S4).

The single variant association test identified one significant biallelic SV, chr6:112 405 701-112 406 900, surpassing both the false discovery rate and Bonferroni thresholds (*P*=1.54E-09, Figure 1). This specific duplication variation occurred on chromosome 6q21, spanning a genomic region of 1200 bp, with an observed allele frequency of 5.9%. Notably, this variant was more common among individuals with genetically inferred ancestries of South Asian and Admixed American (Table 3). In the pooled ancestry analysis, each additional copy of this duplication was associated with a 1.34-fold increased odds ratio for CAD and explained 0.06% of CAD phenotype variance.

We examined the genomic landscape surrounding the duplication region, extending up to 1 Mb, for associations with CAD (Figure 2). Interestingly, we identified 4 SNVs and 1 indel at a distance of over 500 kb from the duplication event, demonstrating a significant association with CAD. However, these variants possessed an exceptionally low frequency in our samples (minor allele frequency <0.00005) and were not in LD with our significant duplication SV. We also identified some common intergenic SNVs located within the duplication on chromosome 6. However, these SNVs that were not found to be in LD with each other, were only nominally associated with cardio-metabolic traits in the phenome-wide association studies catalog (Table S2). We did not identify other SV and SNVs/indels in LD with our duplication using TOP-LD. Furthermore, we looked up the GWAS catalog and found no known GWAS hits

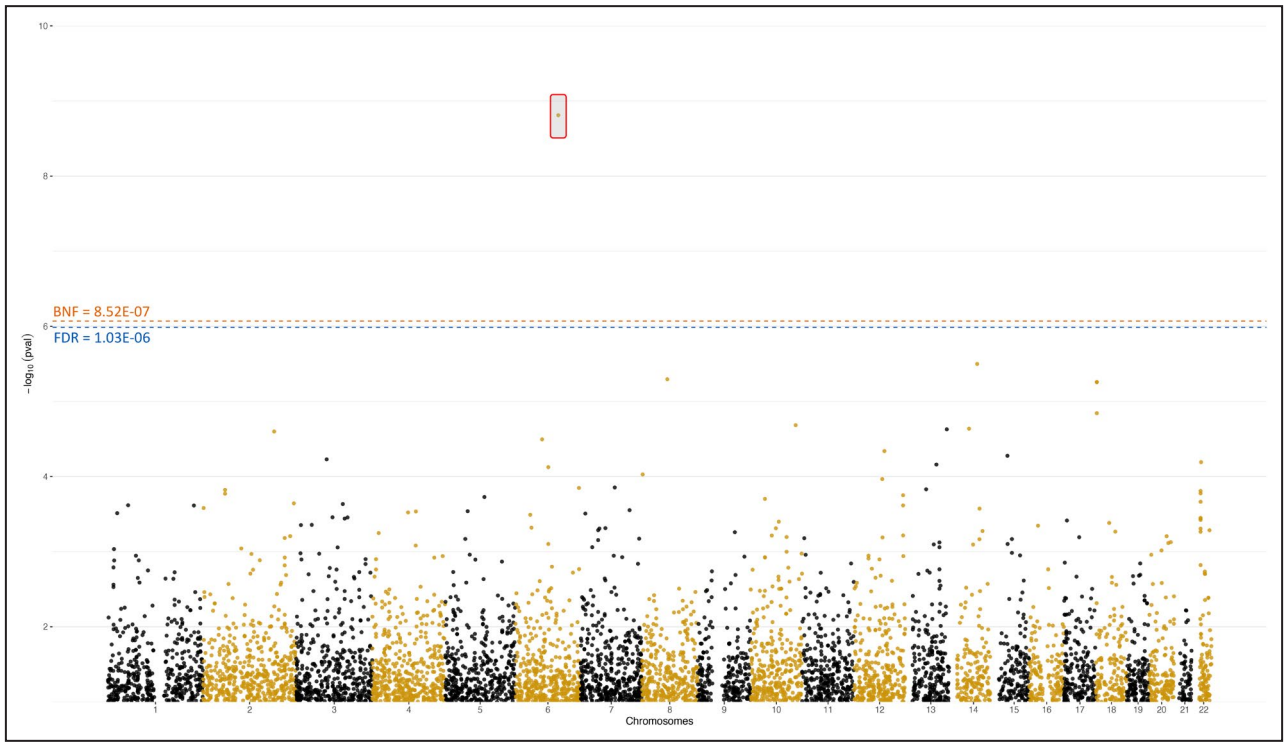

Figure 1. Manhattan plot for genome-wide single variant association test of SVs with CAD.
BNF indicates Bonferroni threshold; CAD, coronary artery disease; FDR, false discovery rate; and SV, structural variant.

Table 3. Minor Allele Frequency and Genotype Distribution of the Duplication Variant Stratified By Genetically Inferred Ancestry and CAD Case/Control Status

	N	N (Case:Control)	MAF (Case:Control)	Genotype 0/0 or CN 2 (Case:Control)	Genotype 0/1 or CN 3 (Case:Control)	Genotype 1/1 or CN 4 (Case:Control)
European ancestry	28 806	5 933: 22 873	9.6%: 5%	3 537: 17 008	843: 1 892	02: 15
African ancestry	16 036	2 593: 13 443	10%: 4.7%	1 600: 10 923	399: 1 090	05: 26
Admixed American ancestry	8 272	2 720: 5 552	15.2%: 3.2%	1 362: 4 534	579: 358	08: 07
East Asian ancestry	1 025	117: 908	5%: 1.6%	89: 826	10:26	00: 01
South Asian ancestry	324	193: 131	16.5%: 7.5%	77: 80	38:12	00: 01

CAD indicates coronary artery disease; CN, copy number; and MAF, minor allele frequency.

within 1 Mb of the duplication event. Interestingly, the duplication overlapped with large known SVs that were previously identified in studies focusing on developmental delay. Most of these variants were classified as pathogenic by ClinVar (Table S3).

Based on the deCode database, the duplication event was localized within a genomic segment demonstrating minimal recombination activity. However, it does not align with regions identified by TOPMed as being densely or sparsely populated with SVs.[29] Functional annotation using AnnotSV indicated that this duplication belonged to the category of unknown significance, was located within an intergenic region of the genome, and did not overlap any short or long ncRNAs. The 2 closest genes to the duplication were *RFPL4B*, positioned ~55 kb upstream, and *MARCKS*, located ~1.45 mb downstream of the event. Additionally, the identified duplication overlapped with an ENCODE predicted distal enhancer characterized by elevated levels of H3K27ac and H3K4me3 histone marks and high CTCF (CCCTC-binding factor) binding sites (Figure 3). Upon comparing the chromatin contact matrices derived from the reference sequence and the alternate sequence featuring the duplication event, Akita predicted modifications in the contact map. The alteration was characterized by an increase in contact between the *LAMA4* gene and the intergenic region situated downstream of the duplication event (Figure 4).

All 4 sliding-window-based aggregate tests identified a significant region on 17q25.3 (P=1.03E-10, Figures S5 and S6). This region overlapped *USP36* gene, and harbored a 45 bp deletion, chr17:78 792 636-78 792 679. This intronic deletion had an observed allele frequency of 0.00006 and a minor allele count of

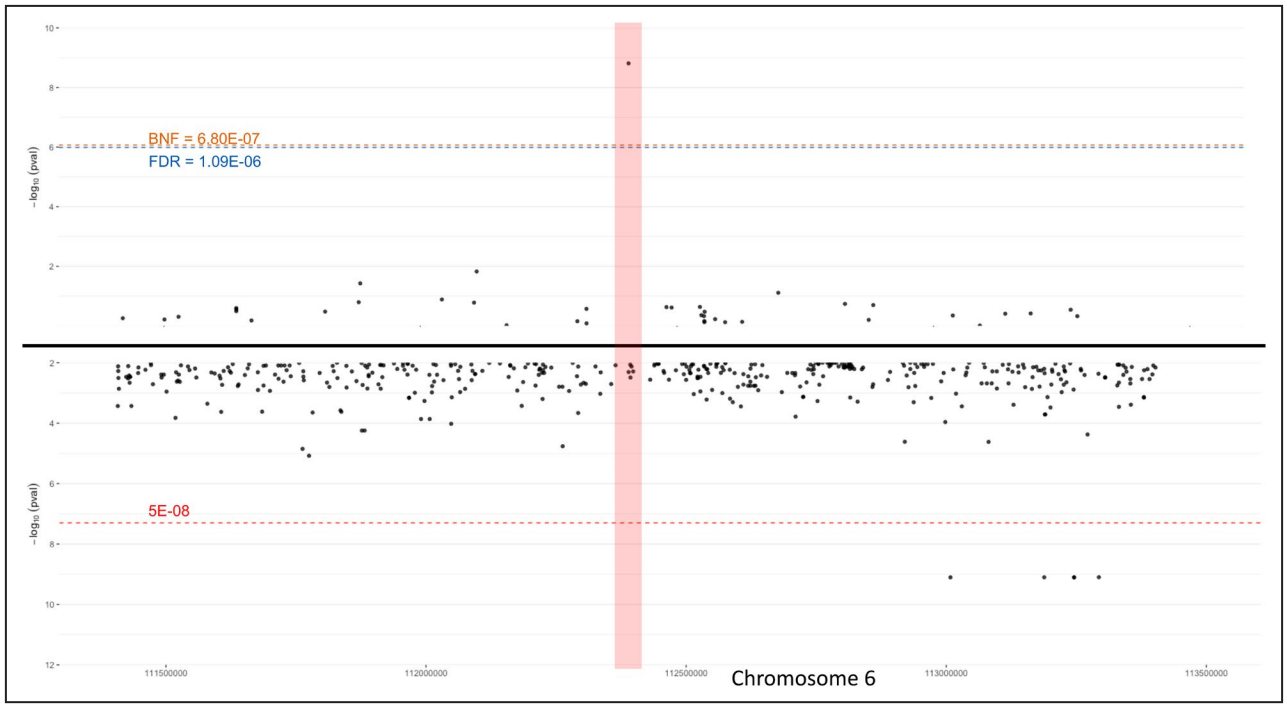

Figure 2. Miami plot for regional associations of SVs (top), SNVs/indels (bottom) with CAD on 6q21.
BNF indicates Bonferroni threshold; CAD, coronary artery disease; FDR, false discovery rate; SNV, single nucleotide variant; and SV, structural variant.

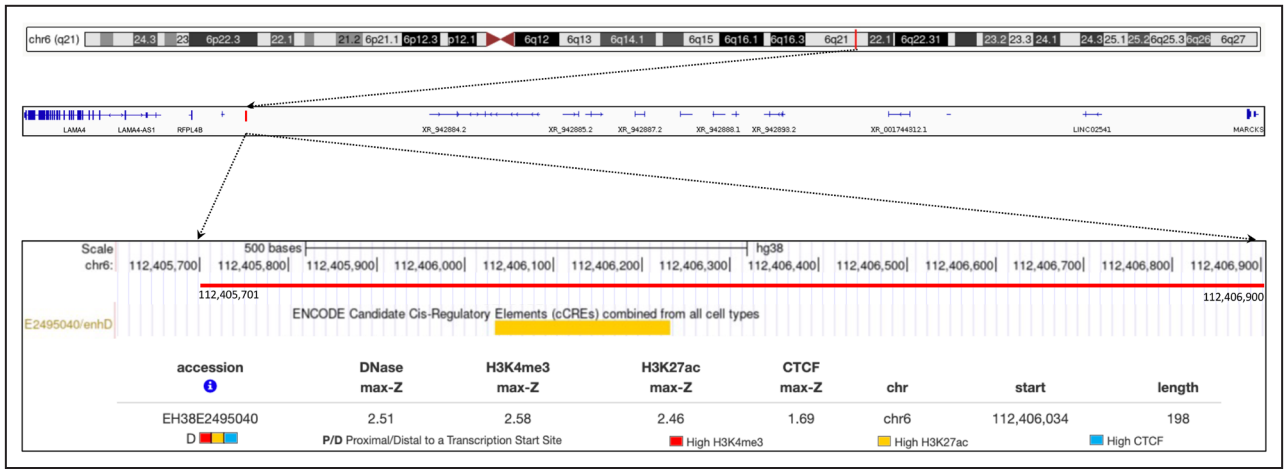

Figure 3. Location and functional annotation of the duplication variant.
CTCF indicates CCCTC-binding factor; DNase, deoxyribonuclease; LINC, long intergenic noncoding RNA; max-Z, maximum *Z* score; and XR noncoding RNA.

6 split equally between cases and controls of African ancestry. GTEx data revealed that *USP36* was well expressed in adipose tissues including both visceral and subcutaneous (median transcripts per million: 31.15; 27.05 respectively), aorta (median transcripts per million: 24.31) and coronary artery (median transcripts per million: 23.77).

We also conducted a more in-depth analysis on regions with a suggestive significance level of $P<1E-4$. Our investigation revealed 24 additional regions, 6 identified by all 4 tests, distributed across various chromosomes (Table 4, Figures S5 through S7). Notably, structural variants were identified in regions on chromosomes 4, 3, 12, 14, and 15, coinciding with the genomic loci of *WDR19*, *ROBO1*, *SOX5*, *DLGAP5*, and *SH3GL3* genes. *WDR19* and *ROBO1* genes were found to be well expressed within aortic tissue. The Variant Set Mixed Model Association Test also identified a region with SVs that overlapped a long noncoding RNA (lncRNA) gene *LINC00570* on chromosome 2; this lncRNA gene's expression was upregulated in myocardial tissues. Interestingly, the inversion

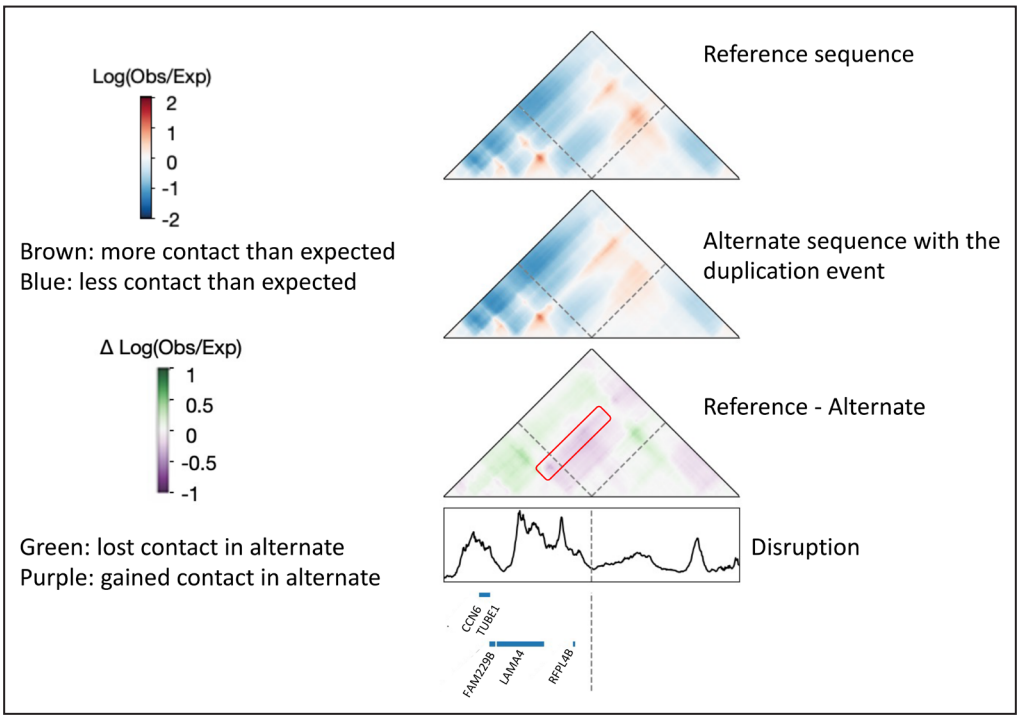

Figure 4. Comparing the chromatin contact matrices with and without the duplication variant.
Obs/Exp indicates observed over expected.

Table 4. Regions That Crossed the Suggestive Significance Threshold *P*<1E-4 During Aggregate Analyses

Chromosome	Region start position	Region end position	# SVs in the region	Identified by RV methods	*P* value burden null weights	*P* value burden Madsen–Browning weights	*P* value SKAT Wu weights	*P* value SMMAT Wu weights	Coordinates of SVs within the region	SV type	SV length in base pairs	MAF of SV	MAC of SV	Does SV overlap a gene?	Does the SV or gene overlap GWAS catalog hits associated with CAD or cardiometabolic traits?
17	78720001	78820000	1	Burden, SKAT, SMMAT	1.04E-10	1.04E-10	1.04E-10	1.04E-10							
									78792634–78792679	Deletion	45	6.04E-05	6	*USP36*	Yes
15	83490001	83590000	2	SMMAT				1.48E-10							
									83522868–83539869	Deletion	17002	1.08917E-04	11	*SH3GL3*	Yes
									83523234–83539870	Deletion	16637	1.10333E-04	12	*SH3GL3*	Yes
15	83520001	83620000	2	SMMAT				1.48E-10							
									83522868–83539869	Deletion	17002	1.08917E-04	11		Yes
									83523234–83539870	Deletion	16637	1.10333E-04	12		Yes
17	78750001	78850000	2	SMMAT				5.89E-10							
									78792634–78792679	Deletion	46	6.04E-05	6	*USP36*	
									78832121–78832845	Deletion	725	6.26E-05	6	*USP36*	
10	90270001	90370000	1	Burden, SKAT, SMMAT	1.04E-06	1.04E-06	1.04E-06	1.04E-06							
									90319449–90326856	Deletion	7407	5.52E-05	6		
4	39180001	39280000	1	Burden, SKAT, SMMAT	1.40E-06	1.40E-06	1.40E-06	1.40E-06							
									39261918–39264046	Deletion	2128	4.62E-05	5	*WDR19*	
3	79200001	79300000	2	Burden, SKAT, SMMAT	2.03E-06	4.63E-06	7.75E-06	1.30E-05							
									79226816–79231573	Deletion	4757	1.38E-04	15	*ROBO1*	Yes
									79253090–79253698	Deletion	608	7.58E-05	8	*ROBO1*	Yes
10	90300001	90400000	2	SMMAT				8.46E-06							
									90319449–90326856	Deletion	7408	5.52E-05	6		
									90372818–90372907	Duplication	90	2.59E-04	23		
17	61680001	61780000	1	Burden, SKAT, SMMAT	9.61E-06	9.61E-06	9.61E-06	9.61E-06							
									61727428–61727987	Deletion	559	6.10E-05	6	*BRIP1*	
10	90240001	90340000	2	SMMAT				1.29E-05							

(*Continued*)

Table 4. Continued

Chromosome	Region start position	Region end position	# SVs in the region	Identified by RV methods	*P* value burden null weights	*P* value burden Madsen–Browning weights	*P* value SKAT Wu weights	*P* value SMMAT Wu weights	Coordinates of SVs within the region	SV type	SV length in base pairs	MAF of SV	MAC of SV	Does SV overlap a gene?	Does the SV or gene overlap GWAS catalog hits associated with CAD or cardiometabolic traits?
									90 246 575–90 246 622	Deletion	48	2.67E-03	267		
									90 319 449–90 326 856	Deletion	7 408	5.52E-05	6		
3	79 170 001	79 270 000	3	SMMAT				1.45E-05							
									79 199 233–79 199 704	Duplication	472	1.58E-03	153	*ROBO1*	Yes
									79 226 816–79 231 573	Deletion	4 758	1.38E-04	15	*ROBO1*	Yes
									79 253 090–79 253 698	Deletion	609	7.58E-05	8	*ROBO1*	Yes
14	51 660 001	51 760 000	2	Burden, SKAT	1.82E-05		2.49E-05								
									51 743 649–51 744 657	Inversion	1 008	9.50E-03	964		
									51 743 714–51 744 622	Deletion	908	3.82E-04	38		
14	51 720 001	51 820 000	3	Burden, SKAT	2.64E-05		2.49E-05								
									51 743 649–51 744 657	Inversion	1008	9.50E-03	964		
									51 743 714–51 744 622	Deletion	908	3.82E-04	38		
									51 815 195–51 816 198	Deletion	1 003	9.28E-05	10	ENSG00000258535	
6	86 430 001	86 530 000	4	SKAT			3.99E-05								
									86 480 910–86 485 951	Deletion	5 042	5.52E-05	6		
									86 494 924–86 498 898	Deletion	3 975	1.10E-04	12		
									86 513 301–86 518 500	Deletion	5 200	1.80E-04	18		
									86 513 383–86 519 209	Deletion	5 827	6.99E-04	70		
2	11 370 001	11 470 000	2	SMMAT				4.53E-05							
									11 377 194–11 382 355	Deletion	5162	5.56E-05	6		
									11 402 930–11 406 923	Deletion	3994	1.84E-04	20	LINC00570	Yes
14	55 080 001	55 180 000	3	SMMAT				5.60E-05							
									55 106 128–55 111 969	Deletion	5 842	4.60E-05	5		Yes
									55 113 947–55 114 087	Duplication	141	5.44E-05	5		
									55 164 816–55 169 005	Deletion	4 190	2.08E-04	21	*DLGAP5*	
2	5 430 001	5 530 000	1	Burden, SKAT, SMMAT	5.82E-05	5.82E-05	5.82E-05	5.82E-05							
									5 488 416–5 489 716	Deletion	1 300	4.64E-05	5		
2	56 880 001	56 980 000	4	Burden	6.26E-05	6.67E-05									
									56 947 546–57 087 116	Inversion	139 570	7.43E-05	8		Yes
									56 957 001–56 960 700	Duplication	3 699	4.78E-05	5		

(*Continued*)

Table 4. Continued

Chromosome	Region start position	Region end position	# SVs in the region	Identified by RV methods	*P* value burden null weights	*P* value burden Madsen–Browning weights	*P* value SKAT Wu weights	*P* value SMMAT Wu weights	Coordinates of SVs within the region	SV type	SV length in base pairs	MAF of SV	MAC of SV	Does SV overlap a gene?	Does the SV or gene overlap GWAS catalog hits associated with CAD or cardiometabolic traits?
									56973801–56986200	Deletion	12399	1.14E-03	114		
									56974993–56986266	Deletion	11273	1.22E-03	132		
2	56940001	57040000	4	Burden	6.26E-05	6.67E-05									
									56947546–57087116	Inversion	139570	7.43E-05	8		Yes
									56957001–56960700	Duplication	3699	4.78E-05	5		
									56973801–56986200	Deletion	12399	1.14E-03	114		
									56974993–56986266	Deletion	11273	1.22E-03	132		
18	2340001	2440000	7	Burden, SKAT, SMMAT	7.90E-05		4.24E-05	8.91E-05							
									2372223–2376973	Deletion	4750	6.09E-03	658	ENSG00000263745	
									2382701–2398800	Deletion	16099	6.42E-04	64	ENSG00000263745	
									2383501–2398700	Deletion	15199	1.01E-03	101	ENSG00000263745	
									2384165–2397062	Deletion	12897	1.25E-03	136	ENSG00000263745	
									2384201–2397000	Deletion	12799	1.25E-03	136	ENSG00000263745	
									2388105–2391779	Deletion	3674	4.59E-05	5	ENSG00000263745	
									2425383–2433546	Deletion	8163	4.60E-05	5	ENSG00000263745	
18	2370001	2470000	8	Burden, SKAT, SMMAT	8.08E-05		4.26E-05	9.18E-05							
									2372223–2376973	Deletion	4750	6.09E-03	658	ENSG00000263745	
									2382701–2398800	Deletion	16099	6.42E-04	64	ENSG00000263745	
									2383501–2398700	Deletion	15199	1.01E-03	101	ENSG00000263745	
									2384165–2397062	Deletion	12897	1.25E-03	136	ENSG00000263745	
									2384201–2397000	Deletion	12799	1.25E-03	136	ENSG00000263745	
									2388105–2391779	Deletion	3674	4.59E-05	5	ENSG00000263745	
									2425383–2433546	Deletion	8163	4.60E-05	5	ENSG00000263745	
									2447858–2448033	Duplication	175	6.52E-05	6	ENSG00000263745	
12	23880001	23980000	1	Burden, SKAT, SMMAT	8.87E-05	8.87E-05	8.87E-05	8.87E-05							
									23951238–23998804	Deletion	47566	4.60E-05	5	*SOX5*	Yes
18	2310001	2410000	7	Burden, SKAT, SMMAT	8.88E-05		4.23E-05	9.88E-05							
									2312870–2312923	Deletion	53	5.85E-05	6	ENSG00000263745	

(*Continued*)

Table 4. Continued

Chromosome	Region start position	Region end position	# SVs in the region	Identified by RV methods	*P* value burden null weights	*P* value burden Madsen–Browning weights	*P* value SKAT Wu weights	*P* value SMMAT Wu weights	Coordinates of SVs within the region	SV type	SV length in base pairs	MAF of SV	MAC of SV	Does SV overlap a gene?	Does the SV or gene overlap GWAS catalog hits associated with CAD or cardiometabolic traits?
									2 372 223–2 376 973	Deletion	4 750	6.09E-03	658	ENSG00000263745	
									2 382 701–2 398 800	Deletion	16 099	6.42E-04	64	ENSG00000263745	
									2 383 501–2 398 700	Deletion	15 199	1.01E-03	101	ENSG00000263745	
									2 384 165–2 397 062	Deletion	12 897	1.25E-03	136	ENSG00000263745	
									2 384 201–2 397 000	Deletion	12 799	1.25E-03	136	ENSG00000263745	
									2 388 105–2 391 779	Deletion	3 674	4.59E-05	5	ENSG00000263745	
17	61 650 001	61 750 000	2	SMMAT				9.43E-05							
									61 674 501–61 678 000	Duplication	3 500	6.15E-04	65		
									61 727 428–61 727 987	Deletion	560	6.10E-05	6	*BRIP1*	
7	138 210 001	138 310 000	4	Burden	9.97E-05										
									138 215 122–138 228 635	Deletion	13 513	1.10E-04	12		
									138 244 214–138 244 234	Deletion	20	3.60E-04	34		
									138 256 941–138 257 120	Deletion	179	4.54E-04	37		
									138 304 601–138 306 000	Deletion	1 399	5.27E-04	53		

CAD indicates coronary artery disease; GWAS, genome-wide association studies; MAC, minor allele count; MAF, minor allele frequency; RV, rare variant; SKAT, sequence kernel association test; SMMAT variant set mixed model association test; and SV, structural variant.

on chromosome 2 and the deletion on chromosome 12 exhibited overlaps with genomic loci implicated in type-2 diabetes and ischemic stroke traits, as indicated by hits in the GWAS catalog.

DISCUSSION

GWAS have identified several SNVs and small indels that contribute to CAD. However, the impact of SVs on atherosclerotic phenotypes, particularly in individuals of non-European ancestry, has not been adequately assessed. In this study, we evaluated the influence of SVs on CAD in a diverse group of participants from the TOPMed program. Through analyses using single SV association tests, we discovered a duplication on chromosome 6 that exhibited statistically significant associations with CAD. The odds ratio for our duplication variant exhibited a large effect size, especially when considering the small magnitude of the SNV effects profiled from prior CAD association studies. The rare SV aggregate tests helped us uncover a deletion on chromosome 17 that overlapped a gene exhibiting significant expression levels in arterial and adipose tissues.

We cross-examined our SV hits in the recently made available 1000 Genomes Project Phase 3 structural variant data set where SVs were detected by high coverage sequenced data using Illumina short-reads on 3202 samples. However, none of the variants that we found to be associated in our data were present in the 1000 Genomes Project data. This could be attributed to the smaller sample size of the 1000 Genomes Project data set compared with TOPMed.

The modeling of 3D genome structure identified heightened spatial interactions, notably originating from the promoter/enhancer region of *LAMA4* situated on the complementary strand. *LAMA4*, characterized by high expression in aortic and cardiac muscles, though not previously linked to atherosclerotic diseases, has been implicated in human conditions such as arrhythmogenic right ventricular cardiomyopathy/dysplasia,[57] hereditary electrocardiogram abnormalities,[57] and dilated cardiomyopathy.[58,59] These conditions share biological pathways with CAD involving inflammation, dyslipidemia, metabolic syndrome, and cardiac remodeling. We speculate that *LAMA4* could be a candidate gene contributing to these pathways, and our duplication event, due to the spatial orientation, may be in close proximity to the regulatory elements of *LAMA4*, potentially influencing its transcriptional activity, thereby affecting endothelial cells and their function.

One of the neighboring genes near the duplication event, *MARCKS*, exhibits heightened expression levels across various tissues, except for arterial tissue. Previous studies have indicated that silencing of the *MARCKS* gene might carry significant beneficial implications for the biological dynamics of vascular smooth muscle cells situated within blood vessels.[60,61] We postulate that the duplication event may influence the silencing mechanism of *MARCKS*, given the elevated levels of H3K27ac and H3K4me3, the presence of active enhancer elements, and high CTCF binding sites within the duplicated genomic region. The SV, through CTCF proteins—known for mediating long-range chromatin interactions, could bring the enhancer into close proximity with the *MARCKS* promoter region. This, coupled with histone modifications in the region, could signal an upregulation of *MARCKS* expression, thereby increasing vascular smooth muscle cell proliferation and contributing to the progression of atherosclerosis. The potential role of the lncRNA gene *LINC02541*, located between *MARCKS* and the SV cannot be ignored. Prior studies have shown that this lncRNA can modulate gene expression in the context of type 2 diabetes[62] and metabolically unhealthy obesity[63]—both of which fall within the same disease spectrum as CAD. Thus, we conjecture that the SV may upregulate *LINC02541*, which in turn could mediate the upregulation of *MARCKS'* expression. Subsequent investigations are warranted to elucidate the potential causal role of these candidate genes, *LAMA4* and *MARCKS*, in the development of CAD.

Our agnostic rare SV analysis, while combing through the entire genome, identified several hits associated with CAD. Within the genomic vicinity of the identified deletion event on chromosome 17 (17:78792634–78 792 679), we used the GWAS catalog to identify a cluster of 10 SNVs that spanned 1 to 53 kb from the duplication. These SNVs exhibited statistically significant associations with various traits encompassing systolic blood pressure, body mass index, ischemic stroke, and insulin resistance. In addition to *USP36*, several SNVs from *ROBO1*, *SOX5*, and *DLGAP5* were found to be associated with various cardiometabolic traits. A subset of these genes (*USP36*, *ROBO1*, and *WDR19)* were also highly expressed in adipose and aortic tissues, whereas the product of lncRNA gene *LINC00570* was found to be upregulated in myocardial tissues of patients with hypertrophic cardiomyopathy.[64] This is significant because hypertrophic cardiomyopathy shares several biological pathways with CAD, including pathways involved in inflammation, dyslipidemia, metabolic syndrome, and cardiac remodeling, all of which are central to the pathophysiology of both conditions. Furthermore, *SH3GL3*, another gene identified in this analysis, has been implicated as a regulator of blood vessel lumen maintenance in zebrafish.[65] These findings prompt contemplation regarding the potential influence of these rare variants on these genes and their contributions to CAD.

To the best of our knowledge, only 2 GWAS of CAD focusing on CNVs have been reported to date.[20,22]

Among these studies, only 1 investigation involving more than 470 000 individuals from the UK Biobank identified significant associations with CAD,[22] specifically the common deletion at 9p23 (~189 kb), common duplication at 4p16.3 (~63 kb), and aggregated rare deletions at *LDLRAD3*. In our study, we examined these 2 previously implicated common regions and identified 27 rare overlapping deletions (ranging in size from 26 bp to >400 kb) within the 9p23 locus, but none demonstrated statistical significance (Table S4). Similarly, within the 4p16.3 locus, we observed 1 deletion and 1 duplication event, both of which were rare but not even nominally significant (Table S4). Our aggregate analysis did not identify any significant signals across 5 windows on chromosome 11 that overlapped *LDLRAD3* gene region (Table S4). We also investigated the nature of association of our significant findings in the UK Biobank study; we did not find any overlapping in CNVs but identified genes like *ROBO1, SH3GL3*, and *RFPL4B* through their gene-based burden tests that exhibited nominal significance (Table S5). Lack of replication may be attributed to (1) limitations with respect to the sample size of our study; (2) differences in the SV calling methods, (3) our use of whole-genome sequenced data versus the UK Biobank's use of genotyping arrays; (4) methodological differences in assessing rare SV, with their approach centering on genes as biological units for aggregation excluding elements from noncoding genome while we canvassed the whole genome employing the sliding window approach; and lastly (5) differences in the sample composition where UK Biobank participants were healthy and of European ancestry compared with the multiancestry participants ascertained for CAD that we leveraged within the TOPMed program.

We acknowledge several study constraints despite the innovative nature of our study: (1) our analyses were confined to 3 specific categories of autosomal SVs—deletions, duplications, and inversions; (2) solely an additive framework was employed for single variant genome-wide discovery; (3) due to sample size limitations, ancestry-specific assessments and replication were unfeasible; (4) given the low number and frequency of SVs in comparison to SNVs, many sliding windows encompassed only 1 SV; and (5) lack of replication given the unavailability of SV data sets that have been called through sequenced data. Additionally, it is worth noting that ambiguity arises in SV identification through short reads; inaccuracies in the contig assembly process using short reads can result in unreliable variant breakpoints.[66] The resolution of these constraints could be achieved by integrating long-read data into expansive multiancestry SV association studies within the context of CAD. Such endeavors will enhance our comprehension of the contribution of SVs to susceptibility to highly heritable complex traits like CAD.

CONCLUSIONS

In conclusion, our investigation has brought to light novel variants dispersed across multiple chromosomes to help elucidate the previously largely unexplored interplay between SVs and CAD. Although our findings are fewer compared with SNV-based GWAS studies, and the identified SVs appear to play a limited role in CAD, this may be due to the suboptimal sample size, the inherent complexity of SVs, and their calling algorithms. Nevertheless, as in many complex diseases, SVs likely contribute to CAD pathogenesis. And for this, larger WGS studies are essential to replicate our findings and to identify additional associations. Such studies should ideally leverage the fidelity of recently developed long-read sequencing technologies. Addressing the current knowledge gap in understanding the role of SVs in CAD across diverse ancestral groups represents a crucial opportunity to unravel the biological foundations underlying CAD.

ARTICLE INFORMATION

Received July 29, 2024; accepted October 21, 2024.

Affiliations

Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA (K.R.I., K.G., K.S.P.); Department of Medicine, Division of Cardiovascular Medicine (K.R.I., S.L.C., R.G., T.L.A.) and Department of Medicine, Stanford Prevention Research Center (S.L.C.), Stanford University School of Medicine, Stanford, CA; Department of Epidemiology and Biostatistics, University of California, San Francisco, CA (K.G., K.S.P.); Department of Epidemiology, Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX (A.S.H., N.R.H., E.B., A.C.M., G.J., P.S.d.V.); Department of Medicine, Division of Cardiology, Washington University School of Medicine, Saint Louis, MO (E.P.Y., N.O.S.); McDonnell Genome Institute, Washington University School of Medicine, Saint Louis, MO (E.P.Y., N.O.S.); Department of Genetics, Washington University School of Medicine, Saint Louis, MO (N.O.S.); Department of Biostatistics (C.L., J.G.B., A.T.K.) and Department of Medicine, Division of Internal Medicine (J.G.B.), University of Washington, Seattle, WA; Department of Medicine, University of Maryland School of Medicine, Baltimore, MD (J.P.L., H.X., M.E.M., B.D.M.); Department of Medicine, University of Mississippi Medical Center, Jackson, MS (K.E.A., A.F.N., Y.-I.M., A.P.C.); Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX (E.B.); The Charles Bronfman Institute for Personalized Medicine (N.C., R.D., G.R., R.J.L.) and Department of Genetics and Genomic Sciences (G.R.), Icahn School of Medicine at Mount Sinai, New York, NY; Department of Preventive Medicine, Northwestern University, Chicago, IL (D.M.L., L.H.); Department of Medicine, Cardiovascular Health Research Unit (R.N.L., J.C.B., J.S.F., B.M.P.) and Department of Epidemiology (J.S.F., B.M.P.), University of Washington, Seattle, WA; Department of Epidemiology, Colorado School of Public Health, Aurora, CO (G.L.K., K.A.Y.); Department of Biochemistry, Wake Forest University School of Medicine, Winston-Salem, NC (D.W.B., N.D.P.); Department of Medicine, Cardiovascular Medicine, Boston Medical Center, Boston University Chobanian & Avedisian School of Medicine, Boston, MA (E.J.B.); Department of Epidemiology, Boston University School of Public Health, Boston, MA (E.J.B.); Department of Medicine, Cardiovascular Medicine (M.N.) and Department of Medicine, Preventive Medicine & Epidemiology (M.N.), Boston University Chobanian & Avedisian School of Medicine, Boston, MA; Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD (L.R.Y., B.G.K., L.C.B., R.A.M.); Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI (S.L.K., J.A.S., L.F.B., P.A.P.); Institute for Social Research, Survey Research Center, University of Michigan, Ann Arbor, MI (J.A.S.); Department of Medicine, Division of Cardiology, Johns Hopkins University, Baltimore, MD (W.S.P.); Department of Genome Sciences, University of Virginia School of Medicine, Charlottesville, VA (S.S.R.); Department of

Medicine, Wake Forest University School of Medicine, Winston-Salem, NC (D.H.); Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA (J.E.M.); Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC (N.F.); Geriatric Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD (B.D.M.); Faculty of Health and Medical Sciences, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark (R.J.L.); Brown Foundation Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, TX (M.F.); Department of Health Systems and Population Health, University of Washington, Seattle, WA (B.M.P.); Division of Rheumatology, National Jewish Health, Denver, CO (E.A.R.); Department of Internal Medicine, Section on Nephrology, Wake Forest University School of Medicine, Winston-Salem, NC (B.I.F.); University of Texas Health Sciences Center, San Antonio, TX (R.S.V.); Division of Intramural Research, Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD (D.L.); Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC (L.M.R.); Division of Public Health, Fred Hutchinson Cancer Center, Seattle, WA (C.K., A.P.R.); Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA (X.G., K.D.T., J.I.R.); Chan Zuckerberg Biohub, San Francisco, CA (K.S.P.); and VA Palo Alto Healthcare System, Palo Alto, CA (T.L.A.).

Sources of Funding

The principal funding for this work was provided by the National Heart, Lung, and Blood Institute (R01HL146860). Additional study/cohort-specific sources of funding are listed in Data S1, "TOPMed study-specific acknowledgments".

Disclosures

Laura M. Raffield is a consultant for the TOPMed Administrative Coordinating Center (through Westat). The remaining authors have no disclosures to report.

Supplemental Material

Data S1
Tables S1–S6

REFERENCES

1. Martin SS, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, Baker-Smith CM, Barone Gibbs B, Beaton AZ, Boehme AK, et al. 2024 Heart disease and stroke statistics: a report of US and global data from the American Heart Association. *Circulation*. 2024;149:e347–e913.
2. Khera AV, Kathiresan S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat Rev Genet*. 2017;18:331–344.
3. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, Lee KM, Fang H, Chen F, Lu Y, et al. Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat Med*. 2022;28:1679–1692.
4. Aragam KG, Jiang T, Goel A, Kanoni S, Wolford BN, Atri DS, Weeks EM, Wang M, Hindy G, Zhou W, et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat Genet*. 2022;54:1803–1815.
5. van der Harst P, Verweij N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ Res*. 2018;122:433–443.
6. Zhu H, Wang X, Gutin B, Davis CL, Keeton D, Thomas J, Stallmann-Jorgensen I, Mooken G, Bundy V, Snieder H, et al. Leukocyte telomere length in healthy Caucasian and African-American adolescents: relationships with race, sex, adiposity, adipokines, and physical activity. *J Pediatr*. 2011;158:215–220.
7. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747–753.
8. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al. An integrated map of structural variation in 2504 human genomes. *Nature*. 2015;526:75–81.
9. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genet*. 2009;84:524–533.
10. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. A structural variation reference for medical and population genetics. *Nature*. 2020;581:444–451.
11. Priest JR, Osoegawa K, Mohammed N, Nanda V, Kundu R, Schultz K, Lammer EJ, Girirajan S, Scheetz T, Waggott D, et al. De novo and rare variants at multiple loci support the oligogenic origins of atrioventricular septal heart defects. *PLoS Genet*. 2016;12:e1005963.
12. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. Strong association of de novo copy number mutations with autism. *Science*. 2007;316:445–449.
13. Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, Baker C, Franke A, Malafosse A, Genton P, Thomas P, et al. Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS Genet*. 2010;6:e1000962.
14. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. A copy number variation morbidity map of developmental delay. *Nat Genet*. 2011;43:838–846.
15. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*. 2008;320:539–543.
16. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. *Genome Biol*. 2019;20:246.
17. Matarin M, Simon-Sanchez J, Fung HC, Scholz S, Gibbs JR, Hernandez DG, Crews C, Britton A, De Vrieze FW, Brott TG, et al. Structural genomic variation in ischemic stroke. *Neurogenetics*. 2008;9:101–108.
18. Shia WC, Ku TH, Tsao YM, Hsia CH, Chang YM, Huang CH, Chung YC, Hsu SL, Liang KW, Hsu FR. Genetic copy number variants in myocardial infarction patients with hyperlipidemia. *BMC Genomics*. 2011;12(Suppl 3):S23.
19. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*. 2009;41:334–341.
20. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3000 shared controls. *Nature*. 2010;464:713–720.
21. Wu Z, Sheng H, Chen Y, Tang J, Liu Y, Chen Q, Lu L, Jin W. Copy number variation of the lipoprotein(a) (LPA) gene is associated with coronary artery disease in a southern Han Chinese population. *Int J Clin Exp Med*. 2014;7:3669–3677.
22. Aguirre M, Rivas MA, Priest J. Phenome-wide burden of copy-number variation in the UK biobank. *Am J Hum Genet*. 2019;105:373–383.
23. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet*. 2020;21:171–189.
24. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*. 2021;590:290–299.
25. Stilp AM, Emery LS, Broome JG, Buth EJ, Khan AT, Laurie CA, Wang FF, Wong Q, Chen D, D'Augustine CM, et al. A system for phenotype harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) program. *Am J Epidemiol*. 2021;190:1977–1992.
26. TOPMed Whole Genome Sequencing Methods: Freeze 9. Accessed 29 November, 2023. https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-9.
27. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res*. 2015;25:918–925.
28. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. *Bioinformatics*. 2015;31:2202–2204.
29. Jun G, English AC, Metcalf GA, Yang J, Chaisson MJ, Pankratz N, Menon VK, Salerno WJ, Krasheninina O, Smith AV, et al. Structural variation across 138,134 samples in the TOPMed consortium. *bioRxiv*2023.
30. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32:1220–1222.
31. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28:i333–i339.

32. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15:R84.
33. Abyzov A, Li S, Kim DR, Mohiyuddin M, Stütz AM, Parrish NF, Mu XJ, Clark W, Chen K, Hurles M, et al. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun*. 2015;6:7256.
34. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 2011;21:974–984.
35. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun*. 2017;8:14061.
36. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods*. 2015;12:966–968.
37. Jun G, Sedlazeck F, Zhu Q, English A, Metcalf G, Kang HM, Lee C, Gibbs R, Boerwinkle E; (HGSVC) HGSVC. muCNV: genotyping structural variants for population-level sequencing. *Bioinformatics*. 2021;37:2055–2057.
38. Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, Thornton TA, Rice KM, Conomos MP. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*. 2019;35:5346–5348.
39. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free estimation of recent genetic relatedness. *Am J Hum Genet*. 2016;98:127–148.
40. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*. 2018;34:867–868.
41. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. Fast computation and applications of genome mappability. *PLoS One*. 2012;7:e30377.
42. NaMPGaTJA H. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev*. 2011;53:217–288.
43. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods*. 2020;17:1111–1117.
44. Ketrin G, Katherine SP. SuPreMo: a computational tool for streamlining in silico perturbation using sequence-based predictive models. bioRxiv2023:2023.11.03.565556.
45. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009;5:e1000384.
46. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89:82–93.
47. Chen H, Huffman JE, Brody JA, Wang C, Lee S, Li Z, Gogarten SM, Sofer T, Bielak LF, Bis JC, et al. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *Am J Hum Genet*. 2019;104:260–274.
48. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res*. 2023;51:D977–D985.
49. Common metabolic diseases knowledge portal. https://hugeamp.org/.
50. Huang L, Rosen JD, Sun Q, Chen J, Wheeler MM, Zhou Y, Min YI, Kooperberg C, Conomos MP, Stilp AM, et al. TOP-LD: a tool to explore linkage disequilibrium with TOPMed whole-genome sequence data. *Am J Hum Genet*. 2022;109:1175–1181.
51. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, Muller J. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*. 2018;34:3572–3574.
52. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42:D980–D985.
53. Lui TW, Tsui NB, Chan LW, Wong CS, Siu PM, Yung BY. DECODE: an integrated differential co-expression and differential expression analysis of gene expression data. *BMC Bioinformatics*. 2015;16:182.
54. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
55. Consortium G. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45:580–585.
56. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*. 2022;185:3426–3440.e19
57. Chen J, Ma Y, Li H, Lin Z, Yang Z, Zhang Q, Wang F, Lin Y, Ye Z. Rare and potential pathogenic mutations of LMNA and LAMA4 associated with familial arrhythmogenic right ventricular cardiomyopathy/dysplasia with right ventricular heart failure, cerebral thromboembolism and hereditary electrocardiogram abnormality. *Orphanet J Rare Dis*. 2022;17:183.
58. Abdallah AM, Carlus SJ, Al-Mazroea AH, Alluqmani M, Almohammadi Y, Bhuiyan ZA, Al-Harbi KM. Digenic inheritance of LAMA4 and MYH7 mutations in patient with infantile dilated cardiomyopathy. *Medicina (Kaunas)*. 2019;55:17.
59. Knöll R, Postel R, Wang J, Krätzner R, Hennecke G, Vacaru AM, Vakeel P, Schubert C, Murthy K, Rana BK, et al. Laminin-alpha4 and integrin-linked kinase mutations cause human cardiomyopathy via simultaneous defects in cardiomyocytes and endothelial cells. *Circulation*. 2007;116:515–525.
60. Monahan TS, Andersen ND, Martin MC, Malek JY, Shrikhande GV, Pradhan L, Ferran C, LoGerfo FW. MARCKS silencing differentially affects human vascular smooth muscle and endothelial cell phenotypes to inhibit neointimal hyperplasia in saphenous vein. *FASEB J*. 2009;23:557–564.
61. Yu D, Makkar G, Dong T, Strickland DK, Sarkar R, Monahan TS. MARCKS signaling differentially regulates vascular smooth muscle and endothelial cell proliferation through a KIS-, p27kip1- dependent mechanism. *PLoS One*. 2015;10:e0141397.
62. Distefano R, Ilieva M, Madsen JH, Ishii H, Aikawa M, Rennie S, Uchida S. T2DB: a web database for long non-coding RNA genes in type II diabetes. *Noncoding RNA*. 2023;9:30.
63. Prashanth G, Vastrad B, Tengli A, Vastrad C, Kotturshetti I. Investigation of candidate genes and mechanisms underlying obesity associated type 2 diabetes mellitus using bioinformatics analysis and screening of small drug molecules. *BMC Endocr Disord*. 2021;21:80.
64. Cao J, Yuan L. Identification of key genes for hypertrophic cardiomyopathy using integrated network analysis of differential lncRNA and gene expression. *Front Cardiovasc Med*. 2022;9:946229.
65. Zhao Y, Lin S. Essential role of SH3-domain GRB2-like 3 for vascular lumen maintenance in zebrafish. *Arterioscler Thromb Vasc Biol*. 2013;33:1280–1286.
66. Gong T, Hayes VM, Chan EKF. Detection of somatic structural variants from short-read next-generation sequencing data. *Brief Bioinform*. 2021;22:1–15.