Polygenic risk scores for CARDINAL study

Clement A. Adebamowo[1,2], **Adebowale Adeyemo**[3], **Adeyinka Ashaye**[4], **Onoja M. Akpa**[5], **Tinashe Chikowore**[6], **Ananyo Choudhury**[6], **Yasmina J. Fakim**[7], **Segun Fatumo**[8,9], **Neil Hanchard**[3], **Michael Hauser**[10], **Braxton Mitchell**[11], **Nicola Mulder**[12], **Solomon F. Ofori-Acquah**[13,14], **Mayowa Owolabi**[15], **Michèle Ramsay**[6], **Bamidele Tayo**[16], **Archana Bhavani VasanthKumar**[1], **Yuji Zhang**[2,17], **Sally N. Adebamowo**[1,2,✉]

[1]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA.

[2]University of Maryland Marlene and Stewart Greenbaum Comprehensive Cancer Center, University of Maryland School of Medicine, Baltimore, MD, USA.

[3]National Human Genome Research Institute, Rockville, MD, USA.

[4]Department of Ophthalmology, University of Ibadan, Ibadan, Nigeria.

[5]Department of Epidemiology and Medical Statistics, University of Ibadan, Ibadan, Nigeria.

[6]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[7]Department of Biotechnology, University of Mauritius, Reduit, MU, Mauritius.

[8]The African Computational Genomics (TACG) Research Group, MRC/UVRI and LSHTM (Uganda Research Unit), Entebbe, Uganda.

[9]London School of Hygiene and Tropical Medicine, London, UK.

[10]Department of Medicine, Duke University, Durham, NC, USA.

[11]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA.

[12]Computational Biology Division, IDM, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa.

[13]West African Genetic Medicine Centre (WAGMC), University of Ghana, Accra, Ghana.

[14]Center for Translational and International Hematology, Vascular Medicine Institute, University of Pittsburgh, Pittsburgh, PA, USA.

[15]Department of Medicine and Center for Genomic and Precision Medicine, University of Ibadan, Ibadan, Nigeria.

✉ sadebamowo@som.umaryland.edu .

[16]Department of Public Health Sciences, Loyola University Chicago, Maywood, IL, USA.

[17]Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA.

Abstract

The Cardiometabolic Disorders in African-Ancestry Populations (CARDINAL) study site is a well-powered, first-of-its-kind resource for developing, refining and validating methods for research into polygenic risk scores that accounts for local ancestry, to improve risk prediction in diverse populations.

Cardiometabolic disorders (CMDs), particularly heart disease and stroke, are major contributors to the global burden of disease, disability and mortality[1]. These complex disorders, which disproportionately affect people of African ancestry, are caused by several interacting, environmental and genetic factors[2]. Although considerable progress has been made in the identification of the sociocultural, demographic and lifestyle risk factors for CMDs, genetic factors that underlie individual susceptibility to these diseases remain largely unknown[2]. Progress in genomic methodologies has enabled the systematic characterization of genome-wide diversity in health and disease in European-ancestry populations, but the conduct of genetic studies in African-ancestry populations has been underwhelming until recently.

Over the past decade, genome-wide association studies (GWASs) have revealed numerous genetic susceptibility loci for CMDs[3–6], generating interest in the use of polygenic risk scores (PRSs) to improve the prediction of adverse CMD events. However, most genomics studies so far have been conducted in European-ancestry populations and increasingly in East Asian populations, with few studies in other populations[7,8]. The predictive value of PRSs for multiple traits from large-scale GWASs in different populations has consistently shown that PRSs predict individual risk far more accurately in European ancestry than in non-European-ancestry populations[9–13]. This is a consequence of the underrepresentation of non-European-ancestry populations in genomics research globally[14]. Most (approximately 76%) of the participants in all GWASs performed so far are of European ancestry[15] (Fig. 1), even though they constitute only 16% of the global population. Despite this bias away from non-European-ancestry populations, people of African, Hispanic or Latin American ancestry contribute a disproportionately great number of genotype–phenotype associations[15,16], which suggests that analyses that include these groups will be effective in identifying new associations[17].

Genetically, European-ancestry populations are relatively homogeneous compared to African-ancestry populations[18,19]. Although prediction accuracy of PRS decays across all non-European-ancestry populations when using European-derived summary statistics, the prediction accuracy of PRSs is lowest in African-ancestry populations on average, compared to other populations[20,21]. In addition, several studies have demonstrated that the prediction accuracy of PRSs is consistently higher when GWAS summary statistics and allele frequencies for the corresponding ancestry are used[12,22]. Consequently, there is a need for well-powered African ancestry GWAS summary statistics, to improve the

predictive accuracy of PRSs in African-ancestry populations. To address the lack of diversity and underrepresentation of African-ancestry participants in GWAS and PRS studies, and evaluate the relationship between the genetic diversity and PRS performance, we have established the PRS for Cardiometabolic Disorders in African-Ancestry Populations (CARDINAL) study, by integrating genomics datasets consisting of people of African ancestry, including continental Africans, African Americans, and African Jamaicans. The goal of this project is to develop a resource for generating PRSs, improving genomic risk prediction and prioritizing individuals for subsequent clinical surveillance and treatment, in diverse populations.

Development of the CARDINAL study site

CARDINAL is a study site within the Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium. The CARDINAL study site aims to integrate phenotype and genomic datasets from over 50,000 people of African ancestry, to reduce heterogeneity, discrepancies, misclassification and measurement error in the study data. We will evaluate the current approaches for constructing PRSs by comparing their predictivity for diseases versus disease subtypes, African ancestry versus non-African-ancestry populations and in diverse ethnic groups; and to develop a method for the estimation of polygenic score — ancestry partitioned polygenic scores (APPS) — which accounts for ancestry-specific effect size differences and allelic heterogeneity in polygenic score estimation.

Study population

The CARDINAL study site leverages data from several core cohorts, including the African Collaborative Center for Microbiome and Genomics Research (ACCME); Africa Wits-Indepth Partnership for Genomic Studies (AWI-Gen); H3Africa Kidney Disease Research Network (H3A-KDRN); Loyola University Cohorts (LUC); the General Population Cohort, MRC/UVRI and LSHTM Uganda (GPC/MRC); Stroke Investigative Research and Education Network (SIREN); Eyes of Africa: the Genetics of Blindness; and Africa America Diabetes Mellitus (AADM) study. The characteristics of the study cohorts are summarized in Table 1 and the geographical representation of the cohorts is shown in Fig. 2.

The CARDINAL study site also includes over 100,000 participants from replication cohorts, with a wide range of cardiovascular, metabolic, cancer and other complex traits and diseases. These studies include the University of Maryland Amish Complex Genetic Disease Research Program ($n = 7,000$), which examined the genetic determinants of cardiometabolic health; the Stroke Genetics Network, which includes more than 30,000 stroke cases across Europe, the US and Australia and more than 50,000 controls; the African Female Breast Cancer Study ($n = 3,000$) and The Nigerian Integrative Epidemiology of Breast Cancer study ($n = 1,000$), which are evaluating the genomic and epidemiological risk of breast cancer and its molecular subtypes; and the Sickle Cell Disease Genomics of Africa Network[23] ($n = 7,000$), which is identifying genetic markers that influence the expression of hemolysis cytoprotective proteins and the risk of echo cardiovascular dysfunctions.

Data harmonization

To standardize phenotypes, we will identify common data elements collected and create a core data model. We will create a data repository and data dictionary for variables of interest and develop scripts for automated and standardized data transformation. Across the PRIMED consortium, phenotypes may be harmonized according to established principles and guidelines in three main stages, collation, comparison and standardization.

To perform the genomic harmonization, standard quality control parameters will be implemented. We have evaluated different reference panels on African genotype data. We expect the African Genome Resource (AGR) and TOPMed panels to have higher imputation quality across the datasets as AGR comprises a larger number of continental African genomes, while TOPMed has greater overall diversity. Owing to the large genetic and geographical diversity across the cohorts in our study site, we will conduct principal component analysis of the pooled data to reveal how all the populations cluster or overlap.

Evaluate existing PRS models in diverse populations

The core cohorts will be used for discovery, and African ancestry data from the replication cohorts will be used as the target dataset. We will construct PRSs for cardiometabolic traits and diseases (Table 2), with currently available PRS methods. We will use models adjusted for age, sex and population structure, to determine the PRSs with the best predictivity, compare PRSs for selected diseases and subtypes of interest, and compare PRSs in different ancestry groups.

Ancestry-partitioned polygenic scores

The PRS method will account for ancestry-specific differences in effect size and allelic heterogeneity in estimation of partial scores as previously described[24]. The APPS approach can be applied to populations with two or more admixed ancestries if independent summary statistics are available for each ancestry. First, local ancestry deconvolution will be performed for each individual in the dataset. Estimates of local ancestry deconvolution have been shown to rely heavily on the size and diversity of the ancestral proxy panel used. As none of the publicly available datasets represent African genetic and haplotype diversity comprehensively, we will generate an extensive African-ancestry panel using randomly selected individuals from our pan-African dataset. This panel will improve the identification of African haplotypes and thus improve local ancestry partitioning. Next, a partial polygenic score will be calculated for each ancestry based on matched-ancestry summary statistics, weighted, and integrated into a combined score for each individual. The APPS approach will be validated by estimating additional trait heterogeneity attributable to the PRS, additional discriminatory power attributable to the partial score, and P value[25], using datasets from our replication cohorts.

Conclusions and future directions

PRSs developed using large-scale genomic data from epidemiological studies are rapidly becoming ubiquitous and linked to health outcomes in European-ancestry populations.

However, it is not currently possible to apply this approach to diverse, non-European-ancestry populations because of the lack of large, well-characterized datasets and resources for these populations. Therefore, there is an urgent, unmet need to develop a platform for managing and sharing PRS-related data from them. Although the CARDINAL study site represents a first-of-its-kind resource for research among non-European-ancestry populations with direct significant application to individuals of African ancestry globally, we do recognize that the complete genetic diversities across sub-Saharan Africa are not represented in the current sample. Thus, we seek inclusion of datasets from other African-ancestry genomic projects, in regions that lack sufficient representation, to enable us to explore region-based analyses.

The APPS PRS method is primarily designed for analysis of African populations admixed with other ancestry populations. Nevertheless, this approach can be applied to any population with two or more ancestries, in which independent summary statistics are available for each ancestry. By being based on multiple instead of single ancestry-based summary statistics, the APPS method improves on existing approaches. Although it is unknown if the APPS method will have uniform high predictive performance across all diverse ancestries, it will provide a framework for incorporating ancestry effects into PRS estimation and use in multi-ethnic populations globally. We will work with the PRIMED consortium to improve the method by leveraging expertise and datasets from other study sites within and beyond the consortium. We expect to develop an online visualization dashboard for research scientists to access the refined PRS and models created by members of the CARDINAL study site.

Acknowledgements

References

1. Diseases, G. B. D. & Injuries, C. Lancet 396, 1204–1222 (2020). [PubMed: 33069326]

2. Adebamowo SN, Tekola-Ayele F, Adeyemo AA & Rotimi CN Public Health Genomics 20, 9–26 (2017). [PubMed: 28482349]

3. Dikilitas O et al. Am. J. Hum. Genet 106, 707–716 (2020). [PubMed: 32386537]

4. Wuttke M et al. Nat. Genet 51, 957–972 (2019). [PubMed: 31152163]

5. Levy D et al. Nat. Genet 41, 677–687 (2009). [PubMed: 19430479]

6. Hu Y et al. Stroke 53, 875–885 (2022). [PubMed: 34727735]

7. Popejoy AB & Fullerton SM Nature 538, 161–164 (2016). [PubMed: 27734877]

8. Mogil LS et al. PLoS Genet. 14, e1007586 (2018). [PubMed: 30096133]

9. Vilhjalmsson BJ et al. Am. J. Hum. Genet 97, 576–592 (2015). [PubMed: 26430803]

10. Schizophrenia Working Group of the Psychiatric Genomics. Nature 511, 421–427 (2014). [PubMed: 25056061]

11. Belsky DW et al. Biodemogr. Soc. Biol 59, 85–100 (2013).

12. Akiyama M et al. Nat. Genet 49, 1458–1467 (2017). [PubMed: 28892062]

13. Lee JJ et al. Nat. Genet 50, 1112–1121 (2018). [PubMed: 30038396]

14. Martin AR et al. Nat. Genet 51, 584–591 (2019). [PubMed: 30926966]

15. Buniello A et al. Nucleic Acids Res. 47, D1005–D1012 (2019). [PubMed: 30445434]

16. Morales J et al. Genome Biol. 19, 21 (2018). [PubMed: 29448949]

17. Hindorff LA et al. Nat. Rev. Genet 19, 175–185 (2018). [PubMed: 29151588]

18. Choudhury A et al. Nature 586, 741–748 (2020). [PubMed: 33116287]

19. Lazaridis I et al. Nature 536, 419–424 (2016). [PubMed: 27459054]

20. Khera AV et al. N. Engl. J. Med 375, 2349–2358 (2016). [PubMed: 27959714]

21. Prive F et al. Am. J. Hum. Genet 109, 12–23 (2022). [PubMed: 34995502]

22. Kanai M et al. Nat. Genet 50, 390–400 (2018). [PubMed: 29403010]

23. Ofori-Acquah SF & The SickleGenAfrica Network. Lancet Glob Health 8, e1255–e1256 (2020). [PubMed: 32971049]

24. Marnetto D et al. Nat. Commun 11, 628 (2020). [PubMed: 32005863]

25. Choi SW, Mak TS & O'Reilly PF Nat. Protoc 15, 2759–2772 (2020). [PubMed: 32709988]
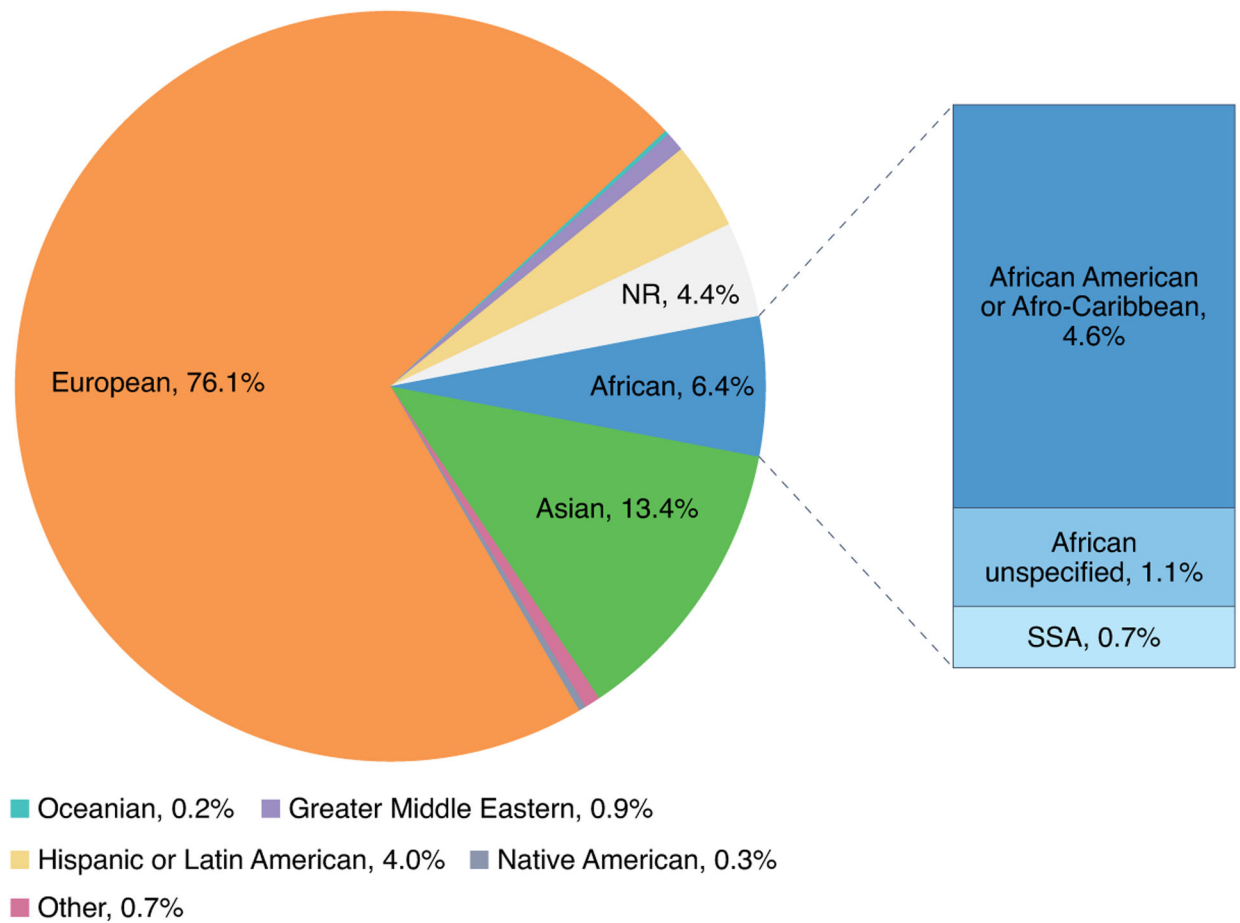
Fig. 1 |. Ancestry distribution in GWAS.

This figure summarizes the distribution of ancestry data contained in the GWAS catalog for studies published from March 2005 to January 2022. It shows the distribution of ancestry categories in percentages, of individuals over time from studies indicating one broad ancestry group. NR, not reported; SSA, sub-Saharan Africa. The figure was created with data obtained from the GWAS catalog.
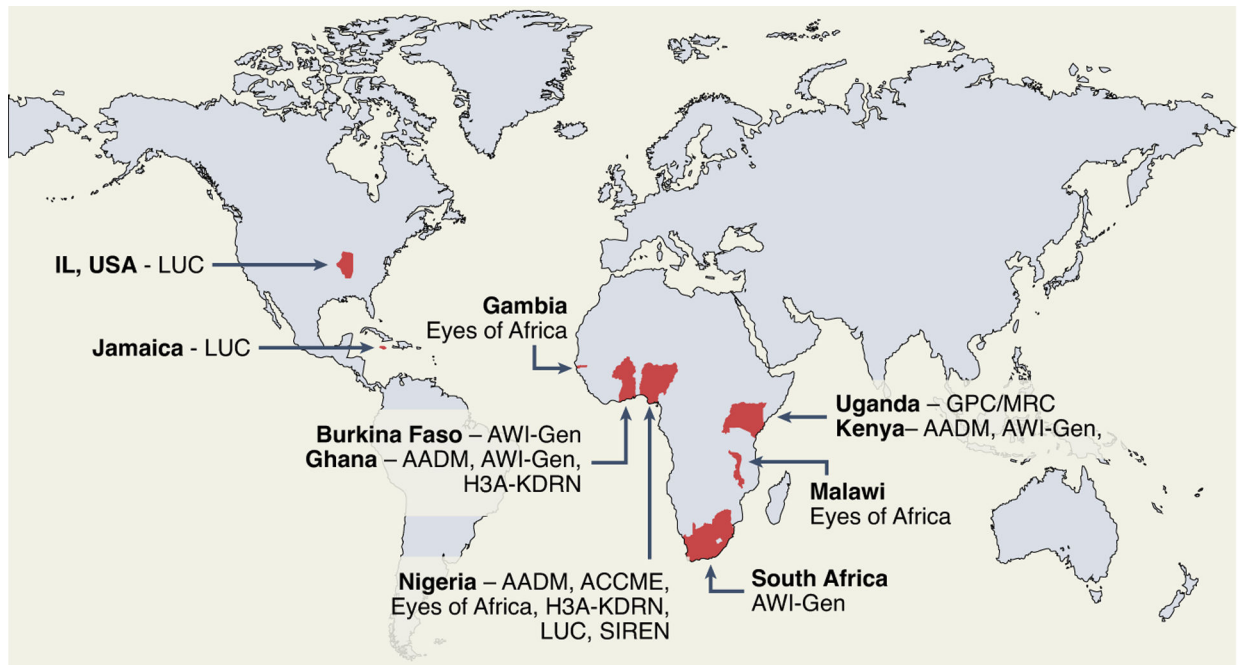
Fig. 2 |. Geographic representation of the core cohorts.
Red indicates the regions where study participants were recruited into the core cohorts in CARDINAL. The figure was created using ArcGIS Pro software by Esri (https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview).

Table 1 |

Characteristics of the core cohorts in the CARDINAL study

Cohorts	ACCME	AWI-Gen	H3A-KDRN & LUC	GPC/MRC	AADM	SIREN	Eyes of Africa
Study design	Prospective longitudinal cohort study	Cross-sectional study	Case-control and cross-sectional studies	Cross-sectional study	Cross-sectional and cohort studies	Case-control study	Cross-sectional study
Participants' country of enrollment	Nigeria	Ghana Burkina Faso Kenya South Africa	Nigeria Ghana USA Jamaica	Uganda	Ghana Nigeria Kenya	Nigeria	Nigeria Gambia South Africa Malawi
Number enrolled	11,711	11,057	8,779	6,867	5,231	6,000	8,000
Age in years, mean (±s.d.)	39 (±10)	50.7 (±6.6)	45.8 (±15.3)	34.4 (±5.1)	50 (±13.0)	58.5 (±13.9)	63.2 (±11.19)
Male/female (%)	0/100	45.2/54.8	42.2/57.8	45/55	36.6/63.4	54.9/45.1	44.7/55.3
No. genotyped after quality control	11,337	10,676	10,355	6,407	5,231	3,421	-
Genotyping array	Illumina H3Africa Array	Illumina H3Africa Array	Illumina H3Africa Array, MetaboChip, BeadChip; Affymetrix 6.0	Illumina Human Omni2.5–8 array	Illumina Multiethnic Genotyping Array (MEGA); Affymetrix PanAFR Array	Illumina H3Africa Array	Illumina H3Africa Array

Table 2 |

Cardiometabolic traits, diseases and related covariates in the CARDINAL study cohorts

Cardiometabolic endpoints and related traits	Potential confounders and effect modifiers
Anthropometry	**Demography**
Height and weight	Age
Waist and hip circumference	Gender
Biomarkers	Ethnicity
Blood glucose, insulin and HbA1c	Education
Creatinine (serum and urine)	Marital Status
Lipids (TC/LDL/TG/HDL)	**Behaviors**
Urinary protein and albumin	Smoking
Platelets, red and white blood cell indices	Alcohol
Liver function parameters	Physical activity
Disease endpoints	Diet
Hypertension	**Health history**
Dyslipidemia	Medical history
Diabetes	Family health history
Stroke	
Kidney disease	

Liver function parameters: are alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase and gamma glutamyl transferase. HbA1c, glycated hemoglobin A1c; HDL, high-density lipoprotein; LDL, low-density lipoprotein; TC, total cholesterol; TG, triglycerides.