nature medicine

BRIEF COMMUNICATION

https://doi.org/10.1038/s41591-022-01835-x

Check for updates

OPEN

Transferability of genetic risk scores in African populations

Abram B. Kamiza[1,2,3], Sounkou M. Toure[1,4], Marijana Vujkovic[5], Tafadzwa Machipisa[6,7,8], Opeyemi S. Soremekun[1], Christopher Kintu[1], Manuel Corpas[9,10,11], Fraser Pirie[12], Elizabeth Young[13], Dipender Gill[14,15], Manjinder S. Sandhu[14], Pontiano Kaleebu[16], Moffat Nyirenda[16], Ayesha A. Motala[12], Tinashe Chikowore[3,17,20] ✉ and Segun Fatumo[1,16,18,19,20] ✉

The poor transferability of genetic risk scores (GRSs) derived from European ancestry data in diverse populations is a cause of concern. We set out to evaluate whether GRSs derived from data of African American individuals and multiancestry data perform better in sub-Saharan Africa (SSA) compared to European ancestry-derived scores. Using summary statistics from the Million Veteran Program (MVP), we showed that GRSs derived from data of African American individuals enhance polygenic prediction of lipid traits in SSA compared to European and multiancestry scores. However, our GRS prediction varied greatly within SSA between the South African Zulu (low-density lipoprotein cholesterol (LDL-C), $R^2 = 8.14\%$) and Ugandan cohorts (LDL-C, $R^2 = 0.026\%$). We postulate that differences in the genetic and environmental factors between these population groups might lead to the poor transferability of GRSs within SSA. More effort is required to optimize polygenic prediction in Africa.

Genome-wide association studies (GWASs) have successfully identified and characterized genetic variants associated with lipid traits[1–3]. To date, roughly 700 single-nucleotide polymorphisms (SNPs) are associated with various lipid traits[3–9]. These discoveries are now beginning to unravel the biology of dyslipidemia and aid prediction for precision medicine. To date, polygenic risk across the genome can be aggregated by generating genome-wide weighted scores to predict the risk of a disease in an independent population[10,11]. However, most lipid trait discoveries have been made in European or Asian ancestries[4–9]. Genetic risk scores (GRSs) derived from European ancestry tend to perform poorly in genetically diverse populations, including Africans[10], probably due to unique differences in linkage disequilibrium (LD) patterns, allele frequencies and environmental exposures[12] between different populations. Lack of precise GRSs in Africans hinders risk stratification and targeted treatments essential for precision medicine and may exacerbate health disparities.

Recent studies have indicated that using multiancestry summary statistics enhance GRS performance in diverse populations[13]. Moreover, previous studies suggested that using summary statistics from African Americans may improve GRS performance in sub-Saharan Africans[14]. We, therefore, undertook a study to determine the best approach for lipid traits polygenic risk prediction, including low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TGs) and total cholesterol (TC) in sub-Saharan Africans using publicly available GWAS summary statistics. This study assessed the performance, portability and predictivity of GRSs derived from data of African Americans, Europeans and multiancestry (African American, European and Hispanic American) individuals in Ugandan and South African Zulu cohorts.

We computed GRSs using PRSice-2. Of the many GRSs computed at various *P*-value thresholds that ranged from 1 to 5×10^{-8}, the GRS that explained the highest proportion of variance (R^2) in any trait for the African, European and multiancestry populations (Methods) was selected as the best-performing one (Extended Data Table 1 and Extended Data Fig. 1). In the South African Zulu cohort, the best-performing GRSs for LDL-C was African American ($R^2 = 8.14\%$, *P*-value threshold (P_T) $< 5 \times 10^{-8}$), followed by the multiancestry approach (derived from individuals of African ancestry, European ancestry and Hispanic American) ($R^2 = 6.32\%$,

[1]The African Computational Genomics (TACG) Research Group, MRC/UVRI and LSHTM, Entebbe, Uganda. [2]Malawi Epidemiology and Intervention Research Unit, Lilongwe, Karonga, Malawi. [3]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [4]African Centre of Excellence in Bioinformatics, University of Science and Technologies of Bamako, Bamako, Mali. [5]Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, Hamilton, Ontario, Canada. [6]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA. [7]Hatter Institute for Cardiovascular Diseases Research in Africa (HICRA), Department of Medicine, University of Cape Town, Cape Town, South Africa. [8]Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Ontario, Canada. [9]Cambridge Precision Medicine Limited, ideaSpace, University of Cambridge Biomedical Innovation Hub, Cambridge, United Kingdom. [10]Institute of Continuing Education, Madingley Hall, University of Cambridge, Cambridge, UK. [11]Facultad de Ciencias de la Salud, Universidade Internacional de La Rioja, Madrid, Spain. [12]Department of Diabetes and Endocrinology, University of KwaZulu-Natal, Durban, South Africa. [13]Omnigen Biodata Ltd, Cambridge, UK. [14]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom. [15]Clinical Pharmacology and Therapeutics Section, Institute of Medical and Biomedical Education and Institute for Infection and Immunity, St George's, University of London, London, UK. [16]MRC/UVRI and LSHTM, Entebbe, Uganda. [17]MRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [18]London School of Hygiene and Tropical Medicine, London, UK. [19]H3Africa Bioinformatics Network (H3ABioNet) Node, Centre for Genomics Research and Innovation, NABDA/FMST, Abuja, Nigeria. [20]These authors contributed equally: Tinashe Chikowore, Segun Fatumo. ✉e-mail: tinashe.chikowore1@wits.ac.za; segun.fatumo@lshtm.ac.uk

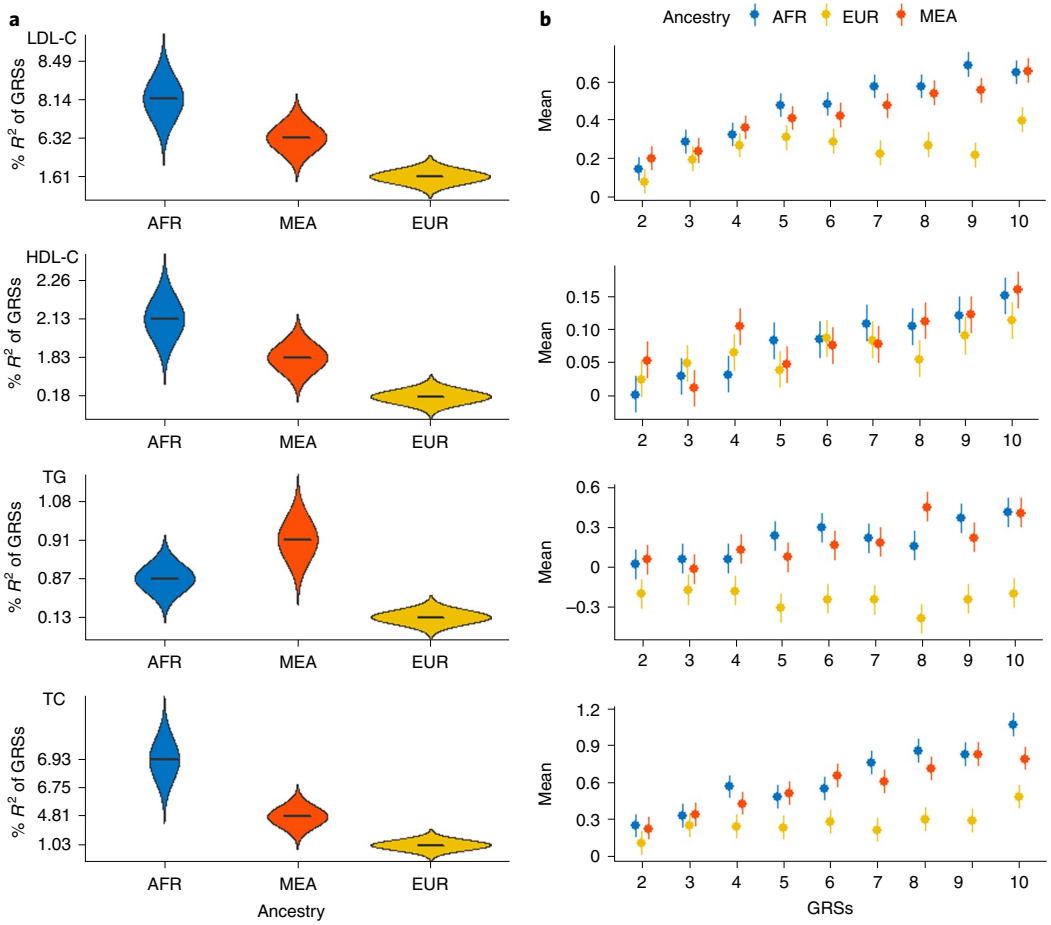


Fig. 1 | Performance of GRSs for lipid traits in the South African Zulu cohort using the MVP GWAS summary statistic results of various ancestry populations, including individuals of African American, European and multiethnic ancestry populations. a, Violin plots showing GRSs that explained the highest proportion of variance (R^2) for lipids derived from African American (AFR), European (EUR) and multiancestry (MEA) populations. **b**, GRSs in deciles compared to the first decile. The *y* axis shows the mean, and the *x* axis is the GRSs in deciles. The points show mean, and error bars represent standard errors of the mean. All South African Zulu cohorts ($n = 2{,}598$) were used in this analysis.

$P_T < 5 \times 10^{-08}$), and the one from individuals of European ancestry ($R^2 = 1.61\%$, $P_T < 5 \times 10^{-08}$, Fig. 1a and Extended Data Table 2). Although the African American-derived GRS predicted better in the South African Zulu cohort its prediction was lower in Ugandan cohort (Extended Data Table 3). Moreover, our African American-derived GRSs (coefficient ranging from 0.100 to 0.286) were better correlated with all serum lipid levels than the European GRSs (coefficients ranging from 0.091 to 0.123) in South African Zulu (Extended Data Fig. 2).

We proceeded to evaluate risk stratification based on the deciles of the GRSs for the lipid traits presented (Methods). We compared the effect sizes of serum lipid levels from the first GRS decile after correction for age, sex and ten principal components. In parallel, we observed that individuals in the top 10% of the GRSs had higher serum lipid levels than those in the first decile (Fig. 1b). Notably, multiancestry-derived GRS was the best-performing approach for HDL-C and TG (Fig. 1b). Individuals at the top 10% of the GRSs had a higher difference of 0.16 mmol liter^{-1} and 0.45 mmol liter^{-1} for HDL-C and TG levels, respectively, compared to individuals at the bottom 10% GRSs. For LDL-C and TC, the best-performing approach was the African American GRS, with a mean difference (first versus tenth decile) of 0.70 mmol liter^{-1} for LDL-C and 1.09 mmol liter^{-1} for TC (Fig. 1b) for those at the top 10% GRS decile.

We proceeded to evaluate the transferability of a GRS derived from an African American cohort in Ugandan and South African Zulu cohorts (Fig. 2a). Using TC as an example, we noted that the same African American GRS of 286 SNPs performed poorly in the Ugandan cohort ($R^2 = 0.045\%$) but much better in the South African Zulu cohort (6.345%) (Fig. 2b). The correlations of the GRS with lipid traits were lower among the Ugandan cohort compared to the South African Zulu cohort (Fig. 2c). Of all the lipid traits, predictability was lowest for TGs, possibly due to the nonfasting of participants before blood collection for lipid analysis. TGs, unlike TC and HDL, are sensitive to dietary intake, which might have affected their accurate estimation and consequently its prediction[15].

We then sought to evaluate the contribution of minor allele frequencies to the poor transferability of the GRSs between the Ugandan and South African Zulu cohorts. We compared allele frequencies of the SNPs in the African American-derived GRSs in the Ugandan and South African Zulu cohorts (Fig. 2d). We noted that there were marked differences in age, body mass index and allele frequencies between these cohorts, which might have contributed to the poor transferability of the African American GRS (Extended Data Fig. 3 and Fig. 2e). The South African Zulu cohort recruited participants from an urbanized setting compared to the Ugandan cohort. Therefore the urban and rural environmental differences might also be playing a part in the poor transferability of the African American-derived GRSs between the Ugandan and South African Zulu cohorts. This finding suggests that both genetic and environmental factors might be responsible for the differences in the performance of GRSs in the Ugandan cohort.

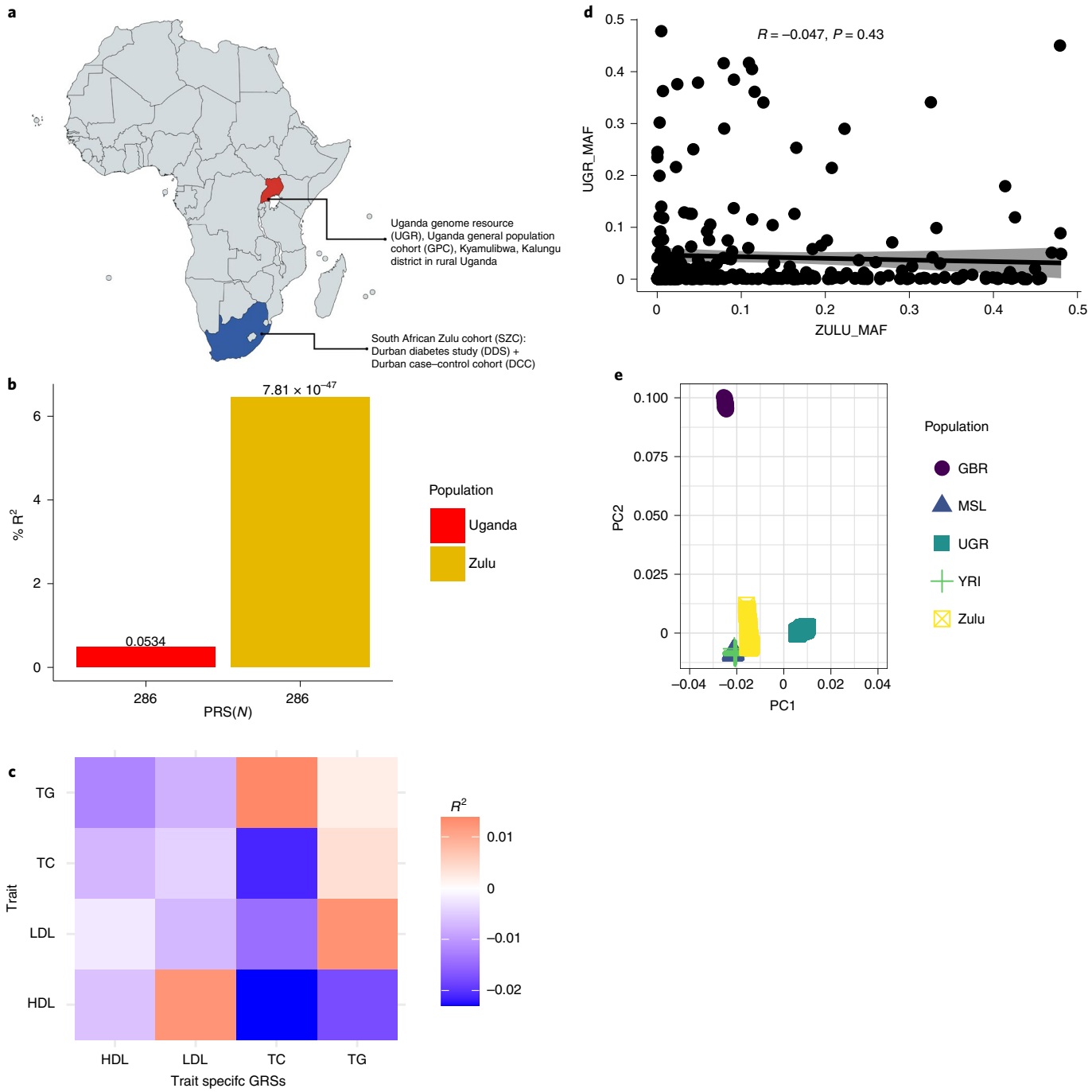


Fig. 2 | GRSs of individuals of African ancestry with dyslipidemia. a, Map of Africa showing sample collection points in Kyamulibwa in Kalungu district, Uganda and Durban, Kwazulu-natal province, South Africa**. b**, Bar plot showing comparative performance of polygenic prediction of TC using the same GRS comprising 286 SNPs, which was developed in Ugandan cohort ($n = 6{,}407$) and then replicated in the South African Zulu cohort ($n = 2{,}598$). The *y* axis is the prediction accuracy (R^2), and the *x* axis is the number of SNPs in the GRS for TC used. **c**, Correlation coefficients between African American-derived GRSs and serum lipid levels in the Ugandan cohort. **d**, Scatter plot for the correlation of the same minor allele frequencies (MAF) between the South African Zulu and Ugandan cohorts (R = Pearson correlation, one-sided test). PC, principal component. **e**, Scatter plot for the principal component analysis of the 1000 Genomes Project reference populations with the South African Zulu and Ugandan cohorts (GBR, British; MSL, Mende; UGR, Uganda genome resource; Zulu, South African Zulu; YRI, Yoruba).

Next, we then assessed the ability of the GRS to identify people with high lipid levels compared to conventional risk factors. We computed residuals of the linear model of TC adjusted for age and sex in the South African Zulu cohort. We then selected individuals at the top 10% of the residual density plot as 'cases' and the remaining 90% deciles as 'controls' (Extended Data Fig. 4a). For example, the average TC level in cases was 6.51 mmol liter^{-1} compared to 4.30 mmol liter^{-1} in controls, representing a difference of 2.21 mmol liter^{-1}. Using logistic regression models, we evaluated the prediction of the African American GRSs trained from the Ugandan cohort in the South African Zulu cohort. The areas under the curve were 55.5% (95% confidence interval [CI], 53.4–57.6%) for clinical

factors, including type 2 diabetes, body mass index, age, sex and five principal components, and 63.8% (95% CI, 61.8–65.9%) for GRSs only (Extended Data Fig. 4b). Moreover, the net reclassification index for the model of the clinical factors increased by 42% after adding the GRSs to this model, further supporting our results that the GRS was better at identifying individuals with high TC compared to conventional clinical factors. However, lipid profiles rather than conventional risk factors are used to assess for dyslipidemia in the clinical setting. Lipid profiles are easier to collect and interpret than GRSs, thereby limiting the clinical application of the GRS. Nonetheless, GRSs might find use in the risk stratification of children and young adults long before they start to exhibit elevated lipid levels[15].

Consistent with previous reports, GRSs derived from individuals of African ancestry performed significantly better in sub-Saharan Africans than GRSs derived from individuals of European ancestry[10,16–18]. The performance of GRS derived from data of African American individuals for LDL-C (R^2 = 8.14%) was much higher than the performances reported by Johnson et al. (ranging from 1.99% to 4.48% in African American, Asian American, white and Hispanic individuals for LDL-C)[18]. This difference suggests that GRSs computed using African Ancestry discovery GWASs may lead to better polygenic predictions of lipids in individuals of African descent. However, continental Africans are characterized by high genetic diversity, which may affect the performance and transferability of GRSs within Africa[12].

Moreover, our results suggest poor transferability of GRS between South African Zulu and Ugandan populations. This might be due to differences in environmental (Extended Data Fig. 3) and genetic factors (Fig. 2D) between the South African Zulu and Ugandan cohorts[19,20]. The poor performance of GRS within the same ancestry population hinders the implementation of GRS in preventative healthcare. It may lead to inaccurate results when applied to different ethnic groups within sub-Saharan Africa. This further suggests the need for more efforts to optimize polygenic prediction in Africa. A limitation of this study is the none inclusion of diet and regular physical activity for the prediction of dyslipidemia. Nevertheless, we included crucial clinical factors, including body mass index, which is strongly associated with diet and regular physical activity; hence, the overall performance of our GRSs were robust.

In conclusion, using GRSs derived from data of individuals of African ancestry performed better in predicting lipid traits in sub-Saharan African populations than GRSs derived from data of individuals of European ancestry. However, the GRS are likely to have variable performances across sub-Saharan African populations, as shown by the differences seen between South African Zulu and Ugandan populations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-022-01835-x.

Received: 18 August 2021; Accepted: 20 April 2022;
Published online: 02 June 2022

References

1. Sanna, S. et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLOS Genet.* **7**, e1002198 (2011).
2. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
3. Asselbergs, F. W. et al. Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am. J. Hum. Genet.* **91**, 823–838 (2012).
4. Willer, C. J. et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* **40**, 161–169 (2008).
5. Chasman, D. I. et al. Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLOS Genet.* **5**, e1000730 (2009).
6. Kathiresan, S. et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.* **40**, 189–197 (2008).
7. Lu, X. et al. Exome chip meta-analysis identifies novel loci and East Asian-specific coding variants that contribute to lipid levels and coronary artery disease. *Nat. Genet.* **49**, 1722–1730 (2017).
8. Liu, D. J. et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
9. Hoffmann, T. J. et al. A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).
10. Duncan, L. et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 1–9 (2019).
11. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 1–11 (2020).
12. Gomez, F., Hirbo, J. & Tishkoff, S. A. Genetic variation and adaptation in Africa: implications for human evolution and disease. *Cold Spring Harb. Perspect. Biol.* **6**, a008524 (2014).
13. Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).
14. Bentley, A. R., Callier, S. L. & Rotimi, C. N. Evaluating the promise of inclusion of African ancestry populations in genomics. *NPJ Genom. Med.* **5**, 1–9 (2020).
15. Bloetzer, C. et al. Screening for cardiovascular disease risk factors beginning in childhood. *Public Health Rev.* **36**, 1–9 (2015).
16. Cavazos, T. B. & Witte, J. S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Hum. Genet. Genomics Adv.* **2**, 100017 (2021).
17. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
18. Johnson, L., Zhu, J., Scott, E. R. & Wineinger, N. E. An examination of the relationship between lipid levels and associated genetic markers across racial/ethnic populations in the multi-ethnic study of atherosclerosis. *PLOS One* **10**, e0126361 (2015).
19. Lucchese, B. Implications of African genetic diversity. *Nat. Rev. Nephrol.* **5**, 663–663 (2009).
20. Yu, N. et al. Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* **161**, 269–274 (2002).

Methods

Study population. The target data for GRS construction were taken from the South African Zulu cohort, a combination of the Durban Diabetes Study (DDS) and the Durban Case-Control Study (DCC) KwaZulu-Natal South Africa. DDS is a population-based cross-sectional study of individuals aged >18 years residing in the urban black communities in Durban, KwaZulu-Natal, South Africa. DCC is a case–control study of individuals aged >40 years with diabetes recruited from tertiary hospitals in Durban. Data collection was conducted from 2009 to 2013 for the DCC and from 2013 to 2014 for the DDS. The survey questionnaire included socioeconomic factors, health information, lifestyle factors, blood pressure, anthropometric measurements (including height, weight, and hip and waist circumferences), biomarkers for communicable and noncommunicable diseases and genetic data. Of the 2,804 individuals surveyed, 1,204 were from the DDS and 1,600 were from the DCC; more detailed information on the study design and quality controls has been published previously[21,22]. Informed consent was obtained from all DDS and DCC participants. The DDS was approved by the University of KwaZulu-Natal Biomedical Research Ethics Committee (BF030/12) and the UK National Research Ethics Service (14/WM/); the DCC was approved by the University of KwaZulu-Natal Biomedical Research Ethics Committee (BF078/08) and the UK National Research Ethics Service (11/H0305/6).

The comparative cohort was taken from the Uganda genome resource (UGR), which is the genomic and phenotypic resource generated from the Uganda General Population Cohort (GPC). The GPC is a population-based cohort study founded in the late 1980s, and it has over 22,000 participants from 25 neighboring villages in Kyamilibwa in rural Uganda. This open-cohort study was established to investigate the trends of HIV infection in Uganda. However, the cohort's focus now is to examine the role of host genetic variants associated with communicable and noncommunicable diseases in rural Ugandans[22]. Informed consent was obtained from all participants, and the Uganda GPC was approved by Uganda Virus Research Institute Research and Ethics Committee (UVRI-REC HS 1978) and the Uganda National Council for Science and Technology (UNCST SS 4283).

Measurement of lipid traits. Nonfasting serum lipid levels were measured using the Cobas Integra 400 Plus Chemistry analyzer (Roche Diagnostics), an automated analyzer that uses four different technologies: absorption photometry, fluorescence polarization immunoassay, immune turbidimetry and potentiometry for accurate analysis. HDL-C and LDL-C were measured using the homogeneous enzymatic colorimetric assays[23,24].

Polygenic risk score. GWAS meta-analysis summary statistics results from the MVP were used as the discovery data sets in GRS computation for the specific lipids. For instance, LDL-C summary statistics from the multiancestry, African American and European cohorts were used for the development of the LDL-C GRSs. The MVP summary statistics results comprised an average of 30 million SNPs from more than 800,000 individuals of diverse ancestry. Of these, 61,796 were African American, and 241,54 were European. The multiancestry summary statistics comprised 25,747 individuals from Hispanic American, European and African American populations. Methods used for genotyping and quality control of MVP data have been previously described[25].

For GRS construction, SNPs from MVP serum lipid summary statistics were clumped based on their LD. We clumped SNPs at different R^2 thresholds, and a 500-kb clumping window with R^2 of 0.5 proved to be the best-fitting and best-performing model for all lipid traits. We also tested the best P-value threshold for selecting which clumped SNPs we would include in the final GRS for the range of 1 to 5×10^{-8}. The *P*-value threshold, which accounted for the highest proportion of the variance of the trait R^2, was selected as the best GRS for TC. The GRS was calculated by multiplying the weight of the SNPs with the number of risk alleles (0/1/2) carried by each individual using the algorithm implemented in the PRSice-2 software[26]. The GRS generated was incorporated into the generalized linear regression model to explain the serum lipids' performance while adjusting for age, sex, type 2 diabetes and five principal components, which were calculated using unrelated individuals and on pruned genotyped data sets using PLINK. An incremental R^2 was computed from each model by the PRSice algorithm and plotted against the P_T. R^2 is the difference between the R^2 of the fully adjusted model (GRS, age, sex, five principal components and diabetes status) and the R^2 of the null model (age, sex, five principal components and diabetes status); the best GRS achieved the highest proportion of R^2 (Fig. 1a).

The best-performing GRS was then categorized into deciles. The bottom decile was used as a reference and compared to other deciles. The difference in the effect sizes of the lipid levels across different GRS deciles was tested using linear regression while adjusting for age, sex, five principal components and diabetes status. We then performed logistic regression with the top decile of the GRS as cases with the remaining 90% as controls. The output of the logistic regression was used to compute the receiver operating curves in R. Furthermore, we used a net reclassification index to assess the ability of the GRS to identify individuals with high TC in the South African Zulu cohort. This reclassification was done by comparing the improvement in reclassification of a null model that comprised the conventional risk factors with that of a null model plus the GRSs using the PredictAbel package in R. The performance of the GRS from each lipid trait was compared among individuals of African ancestry, European ancestry and multiethnic ancestry populations using the ggplot2 R statistical package[27,28].

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

Data availability

Requests for resources and information should be directed to and will be fulfilled by the lead contact, S.F. (segun.fatumo@mrcuganda.org; segun.fatumo@lshtm.ac.uk). All individual-level data and phenotype, genotype and sequence data are available under managed access to researchers. Requests for access will be granted for all research consistent with the consent provided by participants. This would include any research in the context of health and disease that does not involve identifying the participants in any way. The array data have been deposited at the European Genome-phenome Archive (https://www.ebi.ac.uk/ega/, accession number EGAD00010000965). Requests for access to data may be directed to segun.fatumo@mrcuganda.org. Applications are reviewed by a data access committee, and access is granted if the request is consistent with the consent provided by participants. The data producers may be consulted by the data access committee to evaluate potential ethical conflicts. Requestors also sign an agreement that governs the terms on which access to data is granted. The genome-wide association summary statistics data are currently at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001672.v3.p1#:~:text=MVP%20is%20an%20ongoing%20prospective,health%20and%20disease%20among%20veterans.dbGaP Study Accession: phs001672.v3.p1. The data used to construct the PRS are available on the PGS catalog: https://www.pgscatalog.org/publication/PGP000313/ (PGS ID accession: PGP000313).

Code availability

We used publicly available software PRSice to compute PRS and its code is publicly available at https://www.prsice.info/. Other software programs used are listed and described in Methods.

References

21. Hird, T. R. et al. Study profile: the Durban Diabetes Study (DDS): a platform for chronic disease research. *Glob. Health Epidemiol. Genom.* **1**, e2 (2016).
22. Gurdasani, D. et al. Uganda Genome Resource enables insights into population history and genomic discovery in Africa. *Cell* **179**, 984–1002 (2019).
23. Sugiuchi, H. et al. Direct measurement of high-density lipoprotein cholesterol in serum with polyethylene glycol-modified enzymes and sulfated alpha-cyclodextrin. *Clin. Chem.* **41**, 717–723 (1995).
24. Sugiuchi, H. et al. Homogeneous assay for measuring low-density lipoprotein cholesterol in serum with triblock copolymer and α-cyclodextrin sulfate. *Clin. Chem.* **44**, 522–531 (1998).
25. Klarin, D. et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
26. Choi, S. W. & O'Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data.*Gigascience* **8**, giz082 (2019).
27. R Core Team. R: a language and environment for statistical computing https://www.R-project.org/ (R Foundation for Statistical Computing, 2019).
28. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2009). https://doi.org/10.1007/978-0-387-98141-3

Acknowledgements

A.B.K. is supported by the National Institutes of Health/National Human Genome Research Institute (CARDINAL grant 1U01HG011717) at The African Computational Genomics Research Group, Medical Research Council/Uganda Virus Research Institute (MRC/UVRI) and LSHTM (London School of Hygiene & Tropical Medicine Uganda Research Unit), Uganda and Malawi Epidemiology and Intervention Research Unit, Lilongwe/Karonga, Malawi; South African Medical Research Council (with funds received from the South African National Department of Health) and the UK Medical Research Council (with funds received from the UK Government's Newton Fund) (MRC-RFA-SHIP01/2015) for the Evolving Risk Factors for Cancers in African populations study (ERICA-SA). S.F. is an Intermediate international fellow funded by the Wellcome Trust (grant 220740/Z/20/Z) at the MRC/UVRI and LSHTM. T.C. is an international training fellow supported by the Wellcome Trust (grant 214205/Z/18/Z). D.G. was supported by the British Heart Foundation Centre of Research Excellence (RE/18/4/34215) at Imperial College and a National Institute for Health Research Clinical Lectureship (CL-2020-1.0.6-001) at St. George's, University of London. The DCC was funded by Servier South Africa, the South African Sugar Association and the Victor Daitz Foundation. We thank the Fogarty International Center of the National Institutes of Health of the United States for supporting Sounkou Mahamane Toure under grant U2RTW010673 of the West African Center of Excellence for Global Health

Bioinformatics Research Training. We thank the MVP staff, researchers and volunteers who have contributed to MVP, especially participants who previously served their country in the military and generously agreed to enroll in the study (for more details, see https://www.research.va.gov/mvp/ or Gaziano, J. M. et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* 2016;70:214–223). This research is based on data from the MVP, Office of Research and Development, Veterans Health Administration, and was supported by the Veterans Administration Cooperative Studies Program (award G002)

Author contributions

S.F. and T.C. conceptualized and designed the study. S.F., A.B.K. and T.C. performed the data analyses. S.M.T., M.V., T.M., O.S.S., C.K., F.P., E.Y., M.S.S., P.K., M.N. and A.M. conducted the study and/or collected data. S.F., T.C., A.B.K., M.C. and D.G. interpreted the data. A.B.K. wrote the first draft. S.F., T.C., M.C., D.G. and A.M. critically revised the article. All authors provided final approval of the version to be published. S.F. and T.C. supervised the project. All the authors read and provided final approval of the version to be published.

Competing interests

D.G. is employed part-time by Novo Nordisk. At the time of writing, M.C. is associated with Cambridge Precision Medicine Limited, UK. All other authors have no competing interests.

Additional information

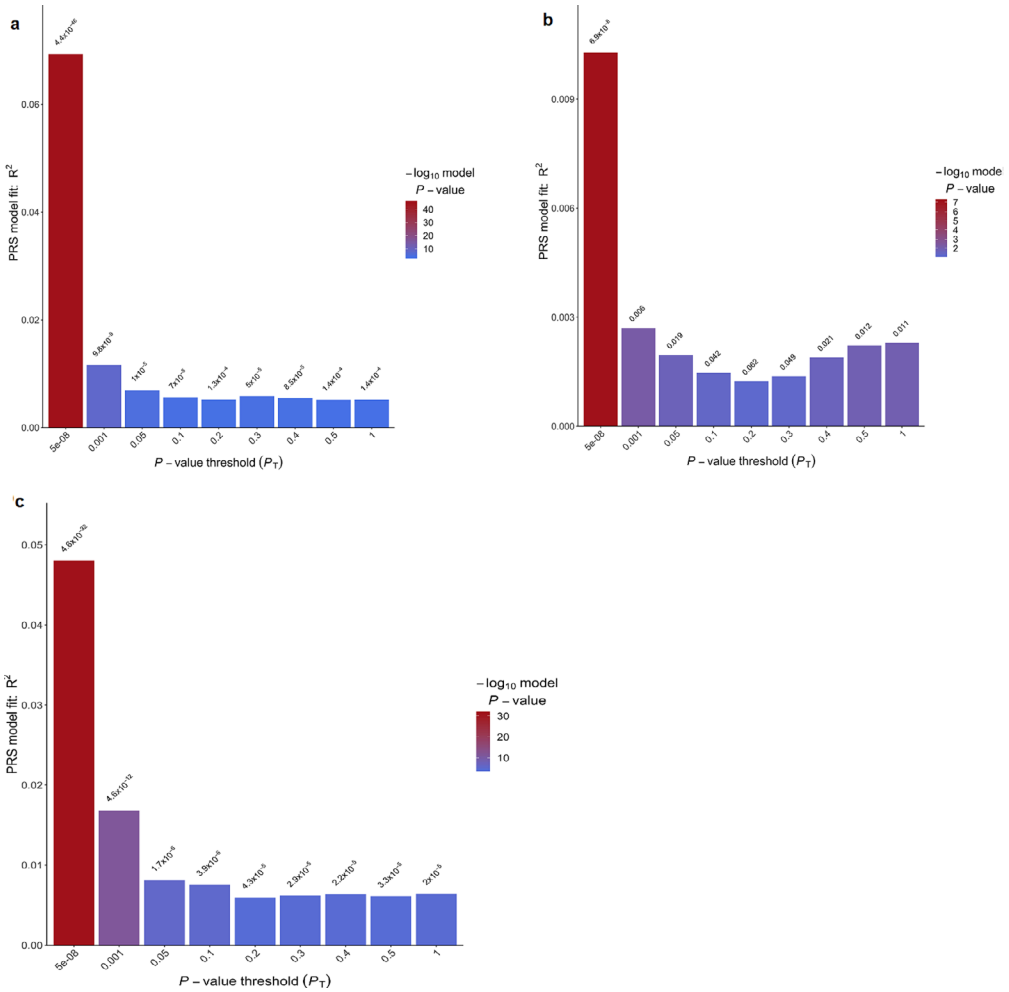
Extended data is available for this paper at https://doi.org/10.1038/s41591-022-01835-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-01835-x.

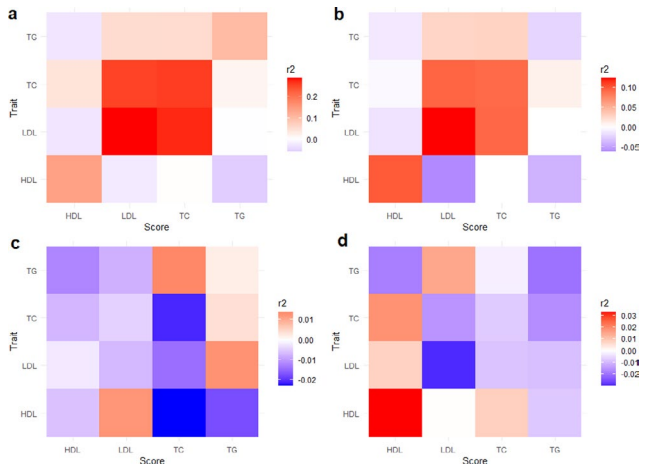
Correspondence and requests for materials should be addressed to Tinashe Chikowore or Segun Fatumo.

Peer review information *Nature Medicine* thanks Iftikhar Kullo, Mayowa Owolabi and Constantin Polychronakos for their contribution to the peer review of this work. Primary Handling editor: Anna Maria Ranzoni, in collaboration with the Nature Medicine team.

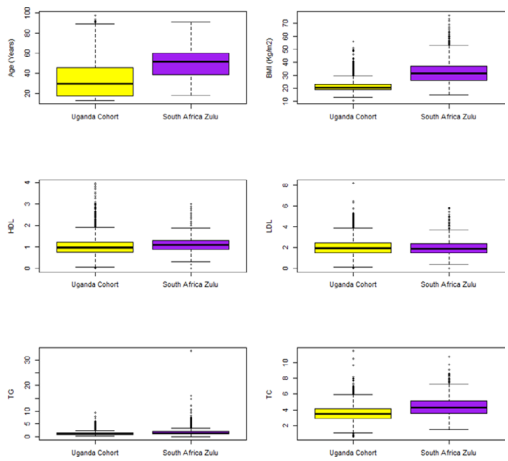
Reprints and permissions information is available at www.nature.com/reprints.



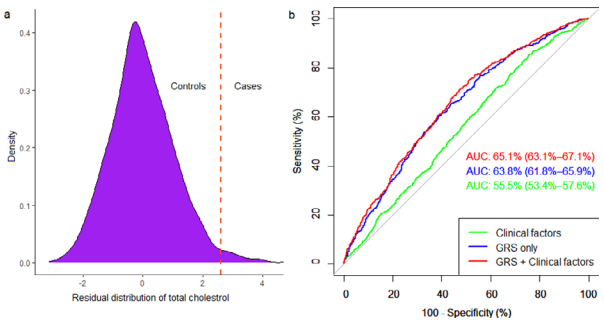
Extended Data Fig. 1 | Proportion of variance of TC explained by GRS in the South African Zulu samples using GRS derived from GWAS. (a) African American ancestry, (a) European ancestry and (a) multiethnic ancestry. The bars represent GRS calculated for subsets of markers at different p-value thresholds. The best GRS in red color was selected based on having the highest proportion of the variance (R2) for the trait in linear models adjusted for age, sex and principal components.



Extended Data Fig. 2 | Correlation coefficients between GRS and serum lipid levels. (a) African American derived GRS in the South African Zulu dataset. (b) European derived GRS in the South African Zulu. (c) African American derived GRS in the Ugandan cohort. (d) European derived GRS in the Ugandan cohort. The correlation coefficients r2 are given with colors corresponding to the direction and strength of r2. The r2 on the diagonal represents the strength of correlation of a GRS with its target lipid trait. The off-diagonal r2 represents the strength of correlation of a GRS with other lipid traits.



Extended Data Fig. 3 | Box plots showing the distribution of age, BMI and lipid traits among the Ugandan and South African Zulu cohorts. The horizontal line is the median values, error bars are 25th and 75th percentiles. Extreme values are maximum and minimum for respective traits. Data analysis were performed in all Ugandan (n = 6,407) and South African Zulu (n = 2,598) cohorts.



Extended Data Fig. 4 | The discriminative power of our polygenic risk score or GRS to successfully identify the individuals of African ancestry with dyslipidaemia. (a) Distribution of total cholesterol (TC) among South African Zulus. The top 10% deciles were named "cases," and the lower deciles were designated as "controls." (b) The area under the curve in South African Zulu.

Extended Data Table 1 | Best-fitting models of clumping using the TC serum lipid trait

Best fitting models of clumping using total cholesterol serum lipid trait

Clump-kb –r2	Best P_T	Coefficient	SE	SNPs	R^2 (%)	P-value
250kb-r2 0.1	0.0110	1071.03	145.54	2,047,936	0.83	2.08×10^{-13}
250kb-r2 0.2	0.0109	1157.86	152.56	2,557,000	0.89	3.67×10^{-14}
250kb-r2 0.5	0.0105	1421	185.75	3,572,766	0.91	2.30×10^{-14}
500kb-r2 0.1	0.0109	917.46	129.52	1,740,383	0.78	1.55×10^{-12}
500kb-r2 0.2	0.0109	1029.78	139.85	2,310,793	0.84	2.01×10^{-13}
500kb-r2 0.5	0.0105	1376.36	179.35	3,451,701	0.91	1.91×10^{-14}

P_T, *p*-value threshold; SE, standard error; SNPs, single nucleotide polymorphisms; R^2; GRS model fit,

Extended Data Table 2 | Best predictive polygenic risk scores of lipid traits in the South African Zulu cohort

Best predictive polygenic risk scores of lipid traits in South African Zulu cohort

Lipid traits	Populations	Adjusted R^2 (%)	Crude R^2 (%)	Predictive R^2	P-value†	Coefficient	SE	SNPs	P-value
HDL-C	AFR	12.123	9.996	2.13	5×10^{-08}	22.342	2.826	201	3.97×10^{-15}
	EUR	10.81	9.996	0.81	5×10^{-08}	0.0296	0.006	1,384	1.28×10^{-06}
	MEA	11.828	9.996	1.83	5×10^{-05}	0.0328	0.004	4,737	3.19×10^{-13}
LDL-C	AFR	14.303	6.166	8.14	5×10^{-08}	115.76	7.391	309	6.83×10^{-53}
	EUR	7.775	6.166	1.61	5×10^{-08}	0.1063	0.015	1,011	2.34×10^{-11}
	MEA	12.484	6.166	6.32	5×10^{-08}	0.1603	0.011	1,302	4.74×10^{-41}
TG	AFR	6.805	5.929	0.87	5×10^{-05}	182.90	37.13	879	8.97×10^{-07}
	EUR	6.061	5.929	0.13	0.2001	-0.0017	0.001	475,919	0.0573
	MEA	6.839	5.929	0.91	0.00045	0.0706	0.014	10,729	5.48×10^{-07}
TC	AFR	15.238	8.305	6.93	5×10^{-08}	196.13	13.49	325	4.43×10^{-46}
	EUR	9.332	8.305	1.03	5×10^{-08}	0.1541	0.028	1,087	6.85×10^{-08}
	MEA	13.106	8.305	4.81	5×10^{-08}	0.2384	0.019	1,447	4.60×10^{-32}

† P-value threshold, SE, standard error; SNPs, single nucleotide polymorphisms; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol; AFR, African ancestry; EUR, European ancestry; MEA, multi-ethnic ancestry

Extended Data Table 3 | Best predictive polygenic risk scores of lipid traits in the Ugandan cohort

Best predictive polygenic risk scores of lipid traits in Ugandan cohort

Lipid traits	Populations	Adjusted R^2 (%)	Crude R^2 (%)	Predictive R^2	P-value†	exit	SE	SNPs	P-value
HDL-C	AFR	9.556	9.553	0.003	5×10^{-08}	-0.8021	1.731	173	0.6432
	EUR	9.562	9.553	0.009	5×10^{-08}	14.171	17.57	1,053	0.4201
	MEA	9.614	9.553	0.061	0.0046	7.0337	3.379	2,427	0.0374
LDL-C	AFR	12.663	12.637	0.026	5×10^{-08}	-6.742	4.909	271	0.1696
	EUR	12.692	12.637	0.055	5×10^{-08}	-48.158	24.18	774	0.0464
	MEA	12.662	12.637	0.025	5×10^{-08}	-27.288	20.43	1,008	0.1817
TG	AFR	2.737	2.735	0.002	5×10^{-05}	2.2983	7.591	720	0.7620
	EUR	2.786	2.735	0.051	0.0529	-97.963	53.41	122,574	0.0666
	MEA	2.782	2.735	0.047	0.0027	4.3846	2.494	491	0.0788
TC	AFR	17.439	17.391	0.048	5×10^{-08}	-20.679	10.71	286	0.0534
	EUR	17.393	17.391	0.002	5×10^{-08}	20.809	51.18	864	0.6843
	MEA	17.431	17.391	0.04	0.0268	9.1977	5.165	1,095	0.0751

† P-value threshold, SE, standard error; SNPs, single nucleotide polymorphisms; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol; AFR, African ancestry; EUR, European ancestry; MEA, multi-ethnic ancestry; MAA, multivariate of the African American.

nature portfolio

Corresponding author(s): Segun Fatumo

Last updated by author(s): Mar 15, 2022

Reporting Summary

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

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	We used publicly available software PRSice to compute PRS and its code is publicly available at https://www.prsice.info/. Other software programs used are listed and described in the Methods section of the manuscript.

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All genome-wide association summary statistics data are currently publicly available either at https://www.ebi.ac.uk/gwas/downloads/summary-statistics and/or https://dbgap.ncbi.nlm.nih.gov/. The multi-ethnic and African American PRS's will be made publicly available on the Polygenic Score (PGS) catalogue https://www.pgscatalog.org/. All individual-level data, phenotype, genotype and sequence data are available under managed access to researchers via the European Genome-phenome Archive (EGA) EGAD00010000965. Requests for access to the phenotypic data will be granted for all research consistent with the consent provided by participants. This would include any research in the context of health and disease that does not involve identifying the participants in any way.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	Uganda genome resource (UGR) is the genomic, and phenotypic resource generated from the Uganda General Population Cohort (GPC). The GPC is a population-based cohort study founded in late 1980, and it has over 22,000 participants from 25 neighbouring villages in Kyamilibwa in rural Uganda. Of these individuals 6,407 consented for genetic study and were available and included in this study.
Data exclusions	No sample excluded from this study
Replication	We performed replication to assess the performance, portability and predictivity of African Americans derived GRS in 2,598 South African Zulu cohort
Randomization	Not applicable as this is not a therapeutic randomization study
Blinding	Not applicable as this is not a therapeutic randomization study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Overall, our study had 6,407 Ugandan and 2,598 South African Zulu cohorts . The majority of our study participants were females; rural Ugandan (n=3,660; 57.1%) from the Uganda Genome Resource (UGR) cohort and South African Zulu (n=1,919; 73.8%) cohort. Our study participants mean ages were and 34.1 (18.3) years in the Ugandan cohort and 33.1 (15.1) years in the South African Zulu cohort. Mean serum lipid was 1.01 mmol/L, 2.05 mmol/L, 1.17 mmol/L and 3.57 mmol/L, for HDL-C, LDL-C. TG and TC, respectively in the UGR cohort and the corresponding values were respectively 1.08 mmol/L, 1.99 mmol/L, 1.60 mmol/L and 4.37 mmol/L in the South African Zulu study cohort.
Recruitment	From 2010-2011, the research questions have included the epidemiology and the genetics of communicable and non-communicable diseases (NCDs) to address the limited data on the burden and risk factors of NCDs in sub-Saharan Africa. The cohort comprises all residents (52% aged > 13 years, men and women in equal proportions) within one-half of a rural sub-county, residing in scattered houses, and largely farmers of three major ethnic groups. Data collection was conducted from 2009 to 2013 for the DCC and from 2013 to 2014 for the DDS. The survey questinnaire included socioeconimic factors, health information, lifestyle factors, anthropometric measurements (including height, weight, systolic, diastolic blood pressure, and hip and waist circumference), biomarkers for communicable and non-communicable diseases, and genetic data.
Ethics oversight	The Uganda GPC was approved by Uganda Virus Research Institute Research and Ethics Committee (UVRI-REC #HS 1978) and the Uganda National Council for Science and Technology (UNCST #SS 4283).

The DDS was approved by the University of KwaZulu-Natal Biomedical Research Ethics Committee (UKZN BREC) (BF030/12) and the UK National Research Ethics Service (14/WM/); the DCC was approved by UKZN BREC (BF078/08) and the UK National Research Ethics Service (11/H0305/6).

Note that full information on the approval of the study protocol must also be provided in the manuscript.