Polygenic prediction of lipid traits in sub-Saharan Africans

Segun Fatumo (segunfatumo@gmail.com)
MRC/UVRI & LSHTM Uganda Research Unit

Abram Kamiza
University of Witwatersrand

Sounkou Toure
University of Science and Technologies of Bamako, Bamako, Mali

Marijana Vujkovic
University of Pennsylvania School of Medicine https://orcid.org/0000-0003-4924-5714

Tafadzwa Machipisa
University of Cape Town

Opeyemi Soremekun
MRC/UVRI and LSHTM

Christopher Kintu
MRC/UVRI and LSHTM

Manuel Corpas
Cambridge Precision

Fraser Pirie
University of KwaZulu-Natal

Elizabeth Young
Omnigen Biodata Ltd

Dipender Gill
Imperial College London https://orcid.org/0000-0001-7312-7078

Manjinder Sandhu
Imperial College

Pontiano Kaleebu
Uganda Virus Research Institute

Moffat Nyirenda
Malawi Epidemiology and Intervention Research Unit https://orcid.org/0000-0003-2120-4806

Ayesha Motala
UKZN

Tinashe Chikowore
Wits

Brief Communication

Keywords: Lipid traits, PRS, GWAS, African ancestry, Zulu, Uganda

Posted Date: August 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-824992/v1

Version of Record: A version of this preprint was published at Nature Medicine on June 2nd, 2022. See the published version at https://doi.org/10.1038/s41591-022-01835-x.

Abstract

Polygenic risk scores (PRS) can enhance risk stratification and are useful for precision medicine interventions. Here we show that African American Genome-wide Association Study (GWAS) derived PRS enhance prediction of lipid traits in Sub-Saharan Africans. Our PRS prediction varied greatly between South African Zulus (LDL-C, R2 = 8.49%) and Ugandans (LDL-C, R2= 0.043%), potentially attributable to environmental factors. Moreover, the PRS shown here had a higher discriminatory ability (AUC = 74.6%) than conventional risk factors (AUC= 67.8%) to identify extreme phenotypes. This work highlights the utility of PRS derived from relevant ethnic groups for identifying high-risk cases missed by conventional clinical factors.

Background

Genome-wide association studies (GWAS) have successfully identified and characterised 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 dyslipidaemia and aid prediction for precision medicine. Polygenic risk scores (PSR) can be generated 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]. PRS derived from European ancestry tend to perform poorly in genetically diverse populations, including Africans [10], partly due to the unique differences in linkage disequilibrium (LD) patterns, allele frequencies and environmental exposures [12] between populations. Lack of precise PRS in Africans hinders risk stratification and targeted treatment essential for precision medicine and may exacerbate health disparities.

Recent studies have indicated that using multivariate approaches and multi-ancestry summary statistics enhance PRS performance [13,14]. Moreover, previous studies suggested that using summary statistics from African Americans may improve PRS performance in sub-Saharan Africans [15]. We therefore aimed to determine the optimal approach for lipid traits in sub-Saharan Africans using publicly available GWAS summary statistics.

We computed PRS using PRSice-2[7]. Of the many PRS computed at various P-value thresholds that ranged from 1 to 5E-08, the PRS that explained the highest variance (R^2) of the trait was selected as the best performing for the African (univariate and multivariate), European and Multi-ancestry GWAS (Methods). In the South African Zulu target dataset (Table S1), the best performing PRS for low-density lipoprotein cholesterol (LDL-C) was from the multivariate approach derived using high-density lipoprotein cholesterol (HDL-C), LDL-C and triglyceride (TG) African American summary statistics (R^2 = 8.48%, P_T <5 x 10^{-08}). This was followed by the African American univariate (R^2 = 8.14%, P_T <5 x 10^{-08}) multi-ancestry approach, derived from individuals of African ancestry, then European and Hispanic American ancestry (R^2 = 6.32%, P_T <5 x 10^{-08}) and finally individuals of European ancestry (R^2 = 1.61%, P_T <5 x 10^{-08}, Fig.1A and Table.S1). Moreover, the African American derived PRS (coefficient range 0.100 to 0.286) were better

correlated with all serum lipid levels that the European PRS (coefficients ranged from 0.091 to 0.123) in South African Zulu (Fig.S2).

We proceeded to evaluate risk stratification based on the deciles of the PRS for the lipid traits presented (Methods). Analysis of variance (ANOVA) was used to compare the means of serum lipid levels. In parallel, we observed that individuals in the top 10% of the PRS had higher serum lipid levels than individuals in the bottom 10% of the PRS (Fig. 1). Notably, the multi-ancestry derived PRS was the best performing approach for HDL-C (Fig.1E) and TG (Fig.1G). Individuals at the top 10% of the PRS had a higher mean difference of 0.16 mmol/L and 0.45 mmol/L for HDL-C and TG levels, respectively, than individuals at the bottom 10% PRS. For LDL-C and total cholesterol (TC), the best performing approach was the African American multivariate with the mean difference of 0.70 mmol/L for LDL-C (Fig.1F) and 1.09 mmol/L for TC (Fig.1H) among those at the top 10% of the PRS deciles.

We then sought to compare polygenic predictions in Uganda (East Africa) and South African Zulus (Southern Africa) using similar discovery data sets. The predictive performances of PRS were low in the Ugandan population (Table.S2) together with its correlations with lipid traits (Fig.S2C). The multivariate approach of the African American GWAS was the best performing PRS in the Uganda dataset (R^2 = 0.113%, P_T = 0.0012) (Table.S2). We proceeded to evaluate the transferability of a PRS derived from a multivariate African America discovery in Uganda to the South African Zulu. Using TC to explore this, we noted that the same African American PRS of 286 SNPs predicted poorly in Uganda (R^2 = 0.048%) but much better in the South African Zulu's (R^2 = 7.061%) (Fig.S3). Environmental factors might be responsible for these differences.

We then assessed the ability of the PRS to identify people with extreme lipid levels, compared to conventional risk factors. We computed residuals of the linear model of total cholesterol adjusted for age and sex in South African Zulu. We then selected individuals at the top 10% of the residual density plot as "cases" and those at the bottom deciles as "controls" (Fig.2A). For example, the average TC level in cases was 6.51 mmol/L compared to 2.76 mmol/L in controls, representing a difference of 3.75 mmol/L. Using logistic models, we evaluated the prediction of the African American PRS derived from Uganda in the South African Zulu's. The area under curves (AUC) were 67.6% for clinical factors including T2D, age, sex and five PCs and 74.7% for PRS only (Fig.2B). This indicates that the PRS was better at identifying individuals with hypercholesterolemia than the conventional risk factors.

Consistent with previous reports, in this study the PRS derived from individuals of African ancestry performed significantly better in sub-Saharan Africans than PRS derived from individuals of European ancestry [10,16–18]. The PRS performance for the African American multivariate approach for LDL-C (R^2 = 8.49%) was much higher than the performances reported by Johnson *et al.*, 2015 ranging from 1.99% to 4.48% in African American, Asian American, Caucasians and Hispanics for LDL-C [18]. This supports that PRS computed using African Ancestry discovery from a multivariate GWAS might lead to better polygenic predictions of lipids in Africa. However, the genetic diversity of African Americans and people residing in

Africa is different. Future studies are required to assess performance of PRS from African individuals within Africa.

Another limiting aspect is the poor transferability of PRS within Africa. This might be due to differences environmental exposure between the South African Zulus and Ugandans [19,20]. The poor performance of PRS hinders the implementation of PRS 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 optimise polygenic prediction in African individuals.

In conclusion, using PRS derived from the African American multivariate approach improved lipid PRS performance in sub-Saharan Africans, as compared to the other considered methods. This approach should be prioritised in studies evaluating PRS application in sub-Saharan Africans by ensuring an increase in the representation of African ancestry individuals in GWAS. Furthermore, the lipid PRS maybe clinically useful to identify individuals at high risk of dyslipidaemia in individuals of African ancestry that are missed by conventional risk factors.

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 Genetics* **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. *Nature Genetics* **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 Genetics* **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. *Nature Genetics* **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. *Nature Communications* **10**, 3328 (2019).

11. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine* **12**, 44 (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. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics* **50**, 229–237 (2018).

14. 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).

15. Bentley, A. R., Callier, S. L. & Rotimi, C. N. Evaluating the promise of inclusion of African ancestry populations in genomics. *npj Genomic Medicine* **5**, 1–9 (2020).

16. Cavazos, T. B. & Witte, J. S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Human Genetics and Genomics Advances* **2**, 100017 (2021).

17. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **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. *Nature Reviews Nephrology* **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 PRS construction were taken from the South African Zulu study, 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, anthropometric measurements (including height, weight, systolic blood pressure, diastolic blood pressure, and hip and waist circumferences), biomarkers for communicable and non-communicable 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 [1,2]. 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).

The comparative cohort was taken from the Uganda genome resource (UGR). The UGR is the genomic, phenotypic resource generated from the Uganda General Population Cohort (GPC). The GPC is a population-based cohort study founded in 1989, and it has over 22,000 participants from 25 neighbouring 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 non-communicable diseases in rural Ugandans. Details on the UGR cohort have been published previously [2].

Measurement of lipid traits

Non-fasting serum lipid levels were measured using the Cobas Integra 400 Plus Chemistry analyser (Roche Diagnostics), an automated analyser that employs 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 [3,4]. Dyslipidaemia was defined as LDL-C >3.3 mmol/L, TG >1.7 mmol/L, TC >5.2 mmol/L, and HDL-C <1.0 mmol/l [5].

Polygenic risk score

Extensive GWAS meta-analysis summary statistics results from the MVP were used as the discovery data in PRS generation. The MVP summary statistics results have over 30 million SNPs from more than 800,000 individuals of diverse ancestry. Of these, 61,796 are individuals of African origin, and 241,541 are individuals of European ancestry. The MVP summary statistics results have over 291,746 individuals with HDL-C, 297,218 individuals with LDL-C, 291,993 individuals with TG, and 297,626 with TC. The multi-ancestry summary statistics included 59,007 individuals of African ancestry, 227,817 individuals of

European ancestry and 25,747 Hispanic Americans. Detailed information on genotyping and quality control for the MVP data has been previously described [6].

For PRS construction, SNPs from MVP serum lipid summary statistics were clumped based on their linkage disequilibrium. We clumped SNPs at different r^2 thresholds, and a 500kb clumping window with r^2 of 0.5 proved to be the best fitting and best performing model for all lipid traits (Table.S3). We also tested the best P-value threshold for selecting which clumped SNPs we would include in the final PRS for the range of 1 to 5E-08. The P-value threshold, which accounted for the highest variance of the trait $R^{2,}$ was select as the best PRS for TC (Fig.S4). The PRS 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 [7]. The PRS generated was incorporated into the generalised linear regression model (GLM) to explain the serum lipid' performance while adjusting for age, sex, type 2 diabetes, and five principal components. An incremental R^2 was computed from each model by the PRSice algorithm and plotted against the P-value threshold (P_T). R^2 is the difference between the R^2 of the fully adjusted model and the R^2 of the null model; the best PRS achieved the highest R^2 (Fig 1).

Moreover, the best performing PRS was then categorised into deciles. The bottom decile was used as a reference and compared to other deciles. The difference in mean lipid levels across different PRS categories was tested using ANOVA. The performance of the PRS from each lipid trait was compared among individuals of African ancestry (AFR), European ancestry (EUR), multivariate of African American (MAA) and from the multiethnic ancestry (MEA) population using the ggplot2 R statistical package [8,9]. The multivariate approach of African Americans for HDL-C was derived from a combination of the summary statistics from HDL-C, LDL-C, TG and TC. For LDL, a combination of the summary statistics was derived from HDL-C, LDL-C and TG. We derived our summary statistics from TG, HDL-C, TC, and LDL-C for the TG multivariate approach. For TC, we combined summary statistics from HDL-C, TG and TC as described in a previous study [10].

References

1. Hird, T. R. *et al.* Study profile: the Durban Diabetes Study (DDS): a platform for chronic disease research. *Glob Health Epidemiol Genom* **1**, (2016).

2. Gurdasani, D. *et al.* Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* **179**, 984-1002.e36 (2019).

3. 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).

4. 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).

5. Noubiap, J. J. *et al.* Prevalence of dyslipidaemia among adults in Africa: a systematic review and meta-analysis. *The Lancet Global Health* **6**, e998–e1007 (2018).

6. 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).

7. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, (2019).

8. R Core Team R: A language and environment for statistical computing.https://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, 2019).

9. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, 2009). doi:10.1007/978-0-387-98141-3.

10. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics* **50**, 229–237 (2018).

Declarations

Contribution statement

SF and TC conceptualised the study. SF, ABK and TC performed the data analyses. ABK wrote the first draft. SF and TC supervised the project and reviewed the first draft. All the authors read and provided critical feedback on the paper.

Acknowledgements

SF is an Intermediate international fellow funded by the Wellcome Trust grant (220740/Z/20/Z) at the MRC/UVRI and LSHTM. TC is an international training fellow supported by the Wellcome Trust grant (214205/Z/18/Z). DG 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-2020-16-001) 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.

Data availability

The data generated or analysed during this study are included in this published article (and its supplementary information files).

Conflicts of interest

DG is employed part-time by Novo Nordisk.

At the time of writing, MC is associated to Cambridge Precision Medicine Limited, UK.

Figures

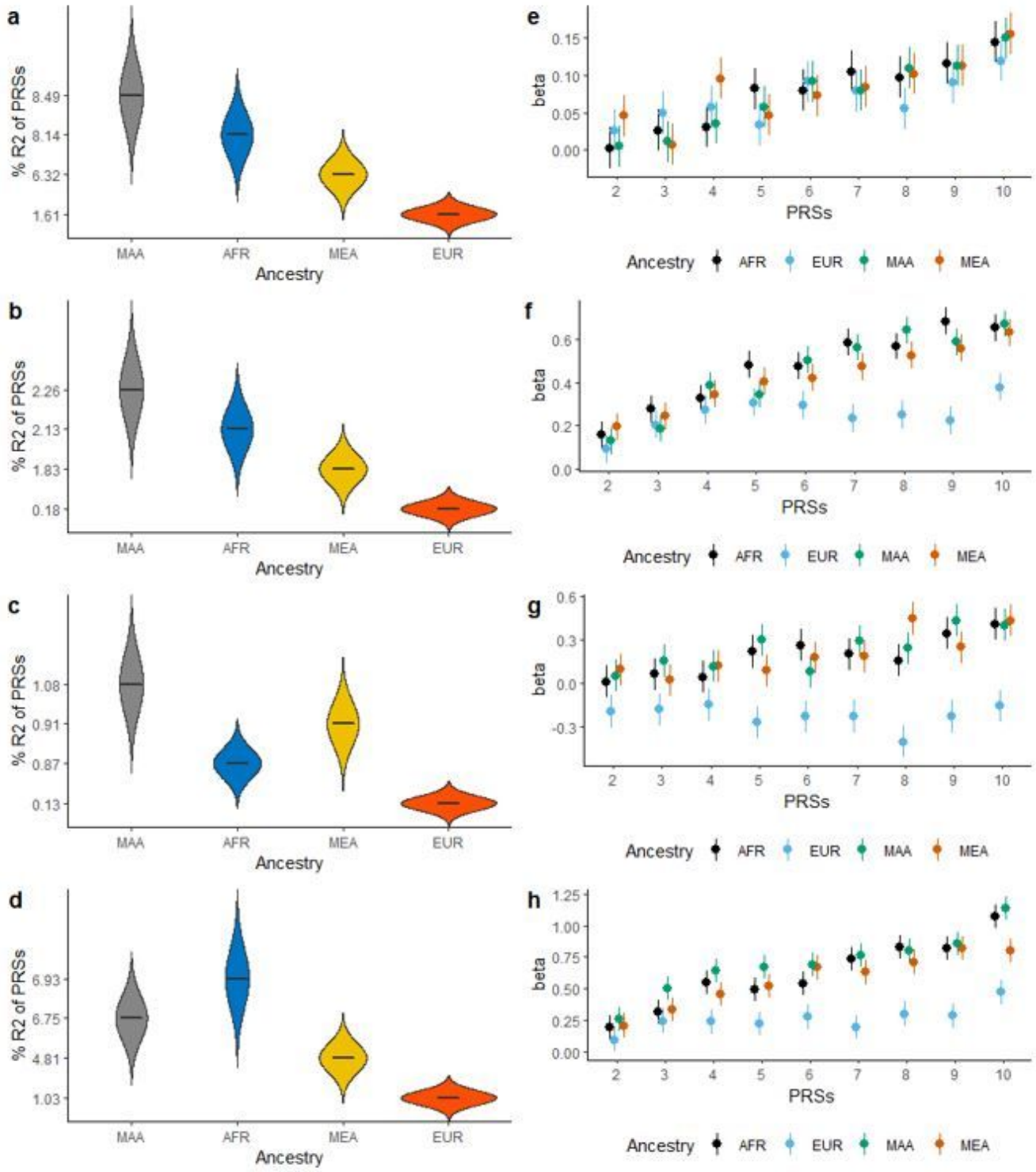

Figure 1

Performance of PRS for lipid traits in the South African Zulu dataset using the MVP GWAS summary statistic results of various ancestry populations including individuals of African American (AFR), multivariate African American (MAA), individuals of European ancestry (EUR) and multiethnic ancestry (MEA) populations. The R2 of PRS include (A) LDL-C, (B) HDL-C, (C) TG, (D) TC and the PRS include (E) HDL-C PRS decile plots, (F) LDL-C PRS decile plots, (G) TG decile PRS plots and (H) TC PRS decile plots.

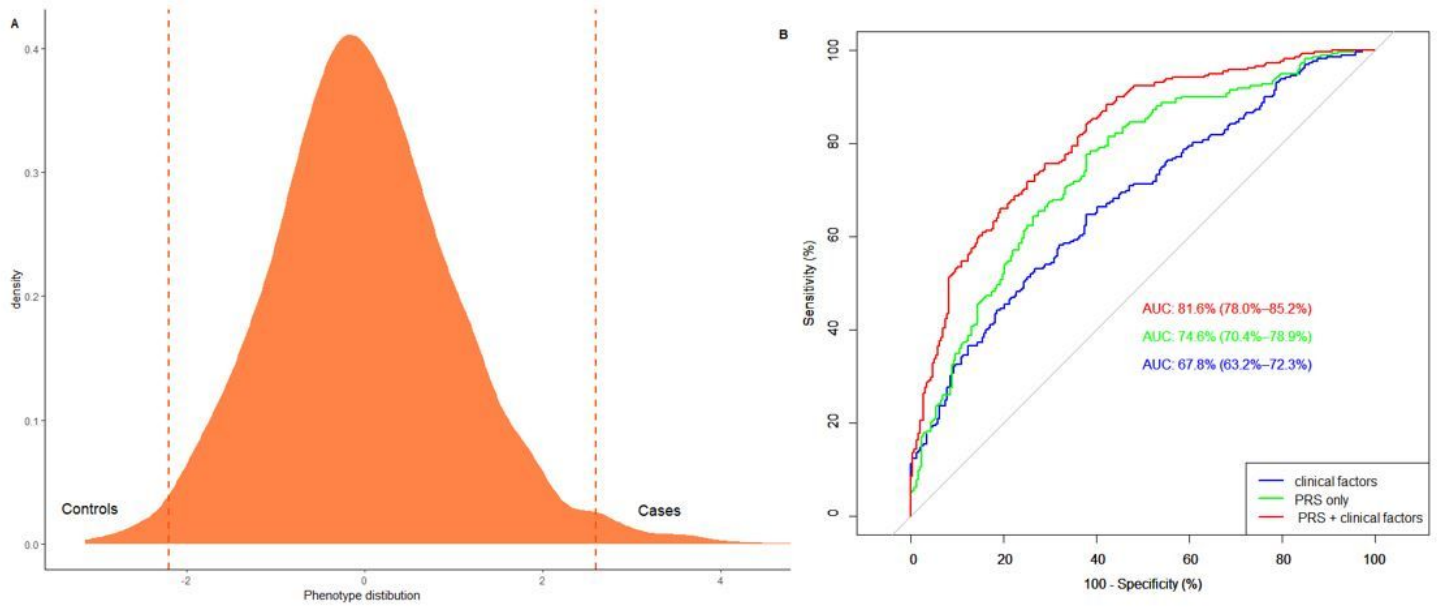

Figure 2

The discriminative power of our PRS to successfully identify individuals of African ancestry with dyslipidaemia. (A) Distribution of total cholesterol 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- flatFatumors.pdf
- flatFatumoepc.pdf
- supplementaryfiles.docx