

https://dspace.org/

In-Depth Survey of the Transferability of Genetic Associations of Lipid Traits to African Populations

Crosby, Michaela

2025
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/48236

Crosby, Michaela . (2025). In-Depth Survey of the Transferability of Genetic Associations of Lipid Traits to African Populations [Master`s dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace.

Downloaded from a DSpace repository. https://dspace.org/

IN DEPTH SURVEY OF THE TRANSFERABILITY OF GENETIC ASSOCIATIONS OF LIPID TRAITS TO AFRICAN POPULATIONS

Michaela Crosby

Student number: 2308127

Supervisor: Dr Dhriti Sengupta

Co-Supervisor: Prof Ananyo Choudhury



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

A research report (in the format of a "submissible" paper) submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements of the degree of Master of Science in Medicine (Genomic Medicine) by Research and Coursework

Johannesburg, 6th March 2025

Letter of Approval

NATIONAL HEALTH LABORATORY SERVICE
School of Pathology, University of the Witwatersrand
DIVISION OF HUMAN GENETICS
Hospital Street, Johannesburg, 2001 I PO Box 1038, Johannesburg, 2000
[T] +27 11 489 9223 I [M] +27 78 080 8841 I [F] +27 11 489 9226 I [E] human.genetics@nhls.ac.za

30 November 2024

Re: Research report submitted to the Faculty of Health Sciences, University of the Witwatersrand, by Ms Michaela Crosby in partial fulfilment of the requirements for the degree of Master of Science in Medicine (Genomic Medicine)

Dear Examiner

Thank you for agreeing to examine the research report submitted by Ms Michaela Crosby (student number 2308127) titled "In depth survey of the transferability of genetic associations of lipid traits, to African populations". This research report is being submitted by the candidate in the format of a "submissible" paper as per the Style Guide for Thesis, Dissertation and Research Reports (updated March 2016) by the University of the Witwatersrand Faculty of Health Sciences. This research report contributes 33% towards the final mark of the MSc (Med) Genomic Medicine degree for which the candidate is enrolled.

The candidate and her supervisors have chosen to submit the paper to the Journal of Human Molecular Genetics. This is an international peer-reviewed scientific journal published by Oxford University Press. The submissible paper attached is therefore written in accordance with the author guidelines of the stipulated journal. These guidelines have been attached for your reference in the appendices section of this document. The only deviation from these guidelines is that the manuscript has been presented as one complete document (including all tables and figures) for ease of marking, rather than being submitted separately.

The research protocol has also been attached as an appendix for the purpose of providing an extended literature review and further contextualise the study. The protocol has already been assessed and approved by the Faculty of Health Sciences Post-Graduate Assessors committee and is therefore not for examination.

Please refer to the Table of Contents for a complete list of the contents provided in this research report.

Yours sincerely,

Michaela Crosby (Candidate)

Dr Dhriti Sengupta (Supervisor)

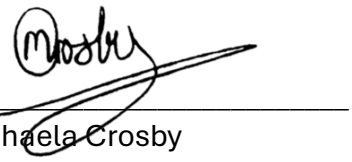
Prof Ananyo Choudhury (Co-Supervisor)

This report is intended solely to record the observations and/or opinion of the writer. It does not constitute a medico-legal report

Chairperson: Prof Jeffrey Mphahlele CEO: Prof Koleka Mlisana
Physical Address: 1 Modderfontein Road, Sandringham, Johannesburg, South Africa Postal Address: Private Bag X8, Sandringham, 2131, South Africa
Tel: +27 (0) 11 386 6000/ 0860 00 NHLS(6457) www.nhls.ac.za
Practice number: 5200296

Declaration

I, Michaela Crosby, declare that this research report (in the format of a "submissible" paper) is my own, unaided work. It is being submitted for the Degree of Master of Science in Medicine (Genomic Medicine) by Research and Coursework at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.



Michaela Crosby

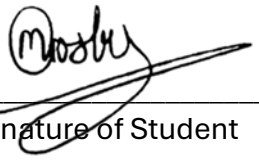
Student Number: 2308127

6th day of March 2025 in Johannesburg

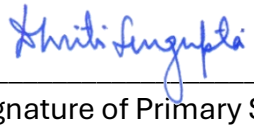
Contribution of Candidate

Student's contribution to the article(s) and agreement of co-author(s)

I, Michaela Crosby, student number 2308127, declare that this Research Report, titled "In depth survey of the transferability of genetic associations of lipid traits to African populations", is my own work and I contributed significantly towards the research findings presented in the paper intended for publication below.



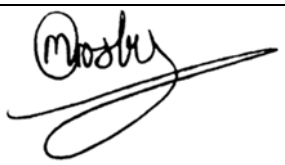
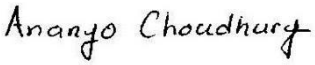
Signature of Student — 06/03/2025 Date



Signature of Primary Supervisor — 06/03/2025 Date

Agreement by co-authors:

By signing this declaration, the co-authors listed below agree to the use of the research findings intended for a published article by the student as part of their Research Report.

Authors	Name	Signature	Date
1st	Ms Michaela Crosby		06/03/2025
2nd	Prof Ananyo Choudhury		06/03/2025
3rd	Dr Dhriti Sengupta		06/03/2025

Presentations arising from this research

1. University of the Witwatersrand, Young and Early Career Geneticists Group Symposium (YEGG), 26th October 2024 (Poster presentation)

2. Southern African Society for Human Genetics Conference, 27th - 29th October 2024 (Poster presentation)

Abstract

The transferability of genetic associations detected in European and US based populations has been shown to be lowest in African populations with high variations among traits as well as different African regions. We report an assessment of the transferability of genetic associations from the Global Lipids Genetics Consortium (GLGC) study, the largest meta-analysis for lipid traits (LDL, HDL, and Triglyceride), to seven African ancestry cohorts comprising participants from eastern, western, and southern Africa. Independent genome-wide significant signals (p-value < $5x10^{-8}$) from the GLGC were identified using the LD clumping function in PLINK 2.0. The transferability of these signals were assessed in the seven cohorts, at the standard genome-wide significance threshold (p-value <$5x10^{-8}$) as well as a less stringent replication threshold (p-value <$5x10^{-6}$). For LDL-C and HDL-C, the GLGC-African-specific ancestry GWAS signals were considerably more transferable compared to the GLGC-multi-ancestry GWAS. Although, transferability estimates generally improved with the size of the replication cohort, there were several exceptions. Moreover, we noted major differences in replication among the African regions with LDL-C signals showing higher transferability in Southern Africa, HDL-C signals showing higher transferability in East and TG signals showing higher transferability to West African cohorts. Among the three lipid traits, LDL-C signals showed highest concordance in beta values with the GLGC beta-values. Our results suggest that transferability of signals depends on several factors including the genetic architecture of the trait being investigated. Therefore, the application of polygenic risk scores and other estimates for precision medicine requires independent in-depth evaluations for each trait in each African region.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr Dhriti Sengupta, for her invaluable guidance, encouragement, and insightful feedback throughout this project. Her expertise and mentorship have been instrumental in shaping this work. Equally, I would like to thank my co-supervisor, Prof Ananyo Choudhury, for all his support and constructive discussions, which enriched our research.

Special thanks to everyone at the Division of Human Genetics for fostering a supportive and collaborative environment that allowed me to learn and grow throughout the year, both in my personal and academic capacity.

I also wish to thank all my classmates for their unwavering support throughout this journey. I am grateful for all the shared experiences, teamwork and encouragement that have been a continuous source of motivation and inspiration throughout the year.

Lastly, I would like to extend heartfelt thanks to my family, whose constant love and support have inspired and motivated me every step of the way, and without whom, this journey would not have been possible.

Table of Contents

Letter of Approval i

Declaration ii

Contribution of Candidate iii

Presentations arising from this research iv

Abstract v

Acknowledgments vi

Table of Contents vii

List of Figures ix

List of Tables x

List of Abbreviations xi

Research report in the format of a “submissible” paper 1

Abstract (For Journal Submission) 2

Introduction 3

Results 5

Discussion 12

Methods 15

- Ethics 15
- GWAS summary statistics download 15
- Identification of independent lead signals 16
- Transferability assessment 16
- Assessing heterogeneity of effect size and risk allele frequencies across African cohorts 18

Data availability statement 18

Acknowledgments 18

Conflicts of Interest Statement 18

Funding 18

References 19

Appendices 21

Appendix A: Supplementary Tables and Figures 22

Supplementary Table 1: Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for LDL-C 22

Supplementary Table 2: Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for HDL-C 22

Supplementary Table 3: Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for TG 22

Appendix B: Ethics Clearance Certificate - M240770 23

Appendix C: Approved Research Protocol 25

Appendix D: Plagiarism Declaration and Turnitin Report 42

Appendix E: Author Guidelines for Human Molecular Genetics Journal 45

List of Figures

Figure 1. Number of LDL-C signals from the Multi-ancestry (blue) and African-specific ancestry (green) GWASs replicated at genome wide significance ($p<5\times10^{-8}$)(dark) and replication threshold ($5\times10^{-8}<p<5\times10^{-6}$) (light) in the African cohorts. 7

Figure 2. Number of HDL-C signals from the Multi-ancestry (blue) and African-specific ancestry (green) GWASs replicated at genome wide significance ($p<5\times10^{-8}$)(dark) and replication threshold ($5\times10^{-8}<p<5\times10^{-6}$) (light) in the African cohorts. 8

Figure 3. Number of TG signals from the Multi-ancestry (blue) and African-specific ancestry (green) GWASs replicated at genome wide significance ($p<5\times10^{-8}$) (dark) and replication threshold ($5\times10^{-8}<p<5\times10^{-6}$) (light) in the African cohorts.....9

Figure 4. Scatter Plots showing comparison of allele frequencies. Allele frequencies for signals for (a) LDL-C (b), HDL- C (c) TG detected in the GLGC African ancestry cohorts compared to allele frequencies observed the combined AWI-Gen GWAS. Allele frequencies for signals for (d) LDL-C (e), HDL- C and (f) TG detected in the GLGC Multi-ancestry cohorts compared to allele frequencies observed the combined AWI-Gen GWAS. 10

Figure 5. Scatter Plots showing comparison of beta-values. Beta-values of signals for (a) LDL-C (b), HDL- C (c) TG detected in the GLGC African ancestry cohorts compared to beta-values observed in the combined AWI-Gen GWAS. Beta-values for signals for (d) LDL-C (e), HDL- C and (f) TG detected in the GLGC Multi-ancestry cohorts compared to beta-values observed in the combined AWI-Gen GWAS. 11

Figure 6. The list of summary statistics files downloaded from GWAS Catalog for each lipid trait.....16

Figure 7. Flow diagram showing the process of extraction and analysis of GWAS summary statistics data 17

Figure 8. Illustration of the analysis approach, which evaluates the transferability of genetic signals from multi- and African-specific ancestry GLGC cohorts to African populations (AWI-Gen and Gurdasani), and compares HDL, LDL, and TG associations across East, West, and South African regions.....17

List of Tables

Table 1. Number of independent lead SNPs in the GLGC datasets that were included in both the continental African studies 6

Supplementary Table 1: Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for LDL-C 22

Supplementary Table 2: Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for HDL-C 22

Supplementary Table 3: Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for TG 22

List of Abbreviations

Abbreviation	Description
AADM	Africa America Diabetes Mellitus
AWI-Gen	Africa Wits-INDEPTH Partnership for Genomic Research
CMD	Cardiometabolic Disease
CVD	Cardiovascular Disease
DCC	Durban Case Control Study
DDS	Durban Diabetes Study
GLGC	Global Lipids Genetics Consortium
GWAS / GWASs	Genome-Wide Association Study Singular) / Genome-Wide Association Studies (Plural)
HDL / HDL-C	High-density Lipoprotein / High-density Lipoprotein Cholesterol
LD	Linkage Disequilibrium
LDL / LDL-C	Low-density Lipoprotein / Low-density Lipoprotein Cholesterol
PRS	Polygenic Risk Scores
SNP	Single Nucleotide Polymorphism
TC	Total Cholesterol
TG	Triglyceride
UGR	Uganda Genome Resource

Research report in the format of a "submissible" paper

To be submitted to the journal of **Human Molecular Genetics**

In depth survey of the transferability of genetic associations of lipid traits to African populations

Michaela E Crosby[1], Ananyo Choudhury[2], Dhriti Sengupta[2]

[1] Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand and the National Health Laboratory Service, 2050, Johannesburg, South Africa

[2] Sydney Brenner Institute of Molecular Biosciences, Faculty of Health Sciences, University of the Witwatersrand, 2050, Johannesburg, South Africa

Corresponding Author:

Michaela Crosby, Tel No: 079 857 6514, Email: michaela.e.crosby@gmail.com

Abstract (For Journal Submission)

The transferability of genetic associations detected in European and US based populations has been shown to be lowest in African populations with high variations among traits as well as different African regions. We report an assessment of the transferability of genetic associations from the Global Lipids Genetics Consortium (GLGC) study, the largest meta-analysis for lipid traits (LDL, HDL, and Triglyceride), to seven African ancestry cohorts comprising participants from eastern, western, and southern Africa. Independent genome-wide significant signals (p-value < $5x10^{-8}$) from the GLGC were identified using the LD clumping function in PLINK 2.0. The transferability of these signals were assessed in the seven cohorts, at the standard genome-wide significance threshold (p-value <$5x10^{-8}$) as well as a less stringent replication threshold (p-value <$5x10^{-6}$). For LDL-C and HDL-C, the GLGC-African-specific ancestry GWAS signals were considerably more transferable compared to the GLGC-multi-ancestry GWAS. Although, transferability estimates generally improved with the size of the replication cohort, there were several exceptions. Moreover, we noted major differences in replication among the African regions with LDL-C signals showing higher transferability in Southern Africa, HDL-C signals showing higher transferability in East and TG signals showing higher transferability to West African cohorts. Among the three lipid traits, LDL-C signals showed highest concordance in beta values with the GLGC beta-values. Our results suggest that transferability of signals depends on several factors including the genetic architecture of the trait being investigated. Therefore, the application of polygenic risk scores and other estimates for precision medicine requires independent in-depth evaluations for each trait in each African region.

Introduction

Despite the progress made in prevention and treatment, cardiovascular disease (CVD) remains the foremost cause of mortality across the world (1). Blood lipid levels are heritable genetic risk factors for CVD and vary globally due to environmental and genetic influences. For complex traits such as lipid biomarker levels, genome wide association studies (GWASs) that study genetic variation across the genome are employed to identify genetic risk factors (2). This is achieved by identifying single-nucleotide polymorphisms (SNPs) with allele frequencies that vary systematically based on phenotypic traits (3,4).

GWASs have successfully identified hundreds of genetic associations for lipid traits such as Low-Density Lipoprotein-Cholesterol (LDL-C), High-density Lipoprotein-Cholesterol (HDL-C) and Triglyceride (TG) levels. However, most GWASs to date have been performed in European ancestry populations, primarily residing in Europe and North America, lacking representation of diverse populations (5).

Approximately 75% of deaths associated with CVD occur in low- and middle-income countries where the prevalence is increasing (6). A sufficient portion of mortality due to CVD can be prevented or at least delayed by simple lifestyle modification and medical intervention. An in-depth characterisation of genetic variants that are associated to a trait in a particular population, has the immense potential for enabling the stratification of individuals by risk and prioritized targeting of individuals at high risk. Therefore, it is important to establish whether SNPs associated with cardiometabolic traits such as lipid traits, that are key markers of CVDs, are transferable to other populations.

The transferability of GWAS signals refers to whether the genetic basis of a trait or phenotype identified in one population, can be replicated or shared across different populations of other ancestries to infer the risk of certain variants causing disease (5). Thus far, the issue lies in the fact that the transferability of genetic associations detected in European and US-based populations have shown to be lowest in African populations, with recent studies showing notable differences in transferability even between various regions within Africa (7–9). This affects the generalisability and applicability of genetic findings, such as the identification of risk allele variants for a particular trait, to different population groups. The lack of transferability of

GWAS findings between populations can be explained by a combination of factors including differences in allele frequencies, genetic architecture, effect sizes and statistical power.

The current understanding of genetics of lipid-traits at the global level as well as that in continental African populations has been shaped on three major GWAS studies. The following paragraphs present a brief introduction to these studies.

The Global Lipids Genetics Consortium (GLGC) is a large-scale GWAS study that conducted a meta-analysis involving over 1,650,000 individuals across five genetic ancestry groups. This was done to identify variants associated with five key lipid traits across different populations, including African, East Asian, European, Hispanic and South Asian. These lipid traits included LDL-C, HDL-C, TGs, total cholesterol (TC) and non-high-density lipoprotein cholesterol (non-HDL-C) (1). The study revealed that utilising diverse ancestries improves signal transferability to African populations and genetic risk factor predictions (1). Of particular interest is the African-specific ancestry sub-cohort that consisted of 100,000 individuals including African Americans, Afro-Caribbeans, and African individuals with unspecified ethnolinguistic metadata.

The landmark study by Gurdasani et al. (2019) (7) was one of the major milestones in African GWASs (7). Based primarily on a population cohort from Uganda, known as the Uganda Genome Resource (UGR), the authors performed GWASs on more than 20 traits that included body composition traits and biochemical traits including three lipid traits. The study also included a meta-analysis of these summary statistics including three other summary statistics from continental African cohorts – the Durban Case Control (DCC), Durban Diabetes study (DDS) and African American Diabetes Mellitus (AADM) study (based predominantly in Nigeria). Although there were sample size differences between the cohorts and some were based on diabetes patients, which was not optimal for the GWAS, the results provided baseline data for future comparisons.

Africa Wits-INDEPTH Partnership for Genomic Research (AWI-Gen) is a pan-African population cross-sectional cohort with a focus on understanding the genetic and environmental risk factors that predispose older adults to cardiometabolic diseases (CMDs) in Africa (10). The cohort consists of ~12,000 participants, aged 40 to 60 years, sampled across four countries - Burkina Faso, Ghana, Kenya and South Africa. The AWI-Gen genetic data was used to perform

a GWAS for fasting serum TC, LDL-C, HDL-C, and TG on 10,603 individuals (8). The study also explored different approaches that could be used to address the challenge of managing complex population substructure between African regions, in Pan African GWASs.

The AWI-Gen lipid GWAS publication (10) also included a comprehensive assessment of transferability of GWAS signals from some of the largest global GWAS studies available at that time (GLGC (11) and PAGE (12)) to the cohorts included in the AWI-Gen and Gurdasani studies. The results demonstrated that increasing the sample size of both the discovery cohort and the replication cohort could lead to more robust transferability estimates. The study also noted considerable regional variation in transferability of signals to different African regions. However, the GLGC study dataset that was assessed was from a study conducted in 2013 (11) and had a much smaller sample size (~100,000 individuals total) compared to the more recent GLGC 2021 dataset, which has an almost 16-fold higher sample size compared to the previous study, potentially leading to major revision in these estimates. Moreover, the availability of summary statistics from a sizable African-ancestry sub-cohort enables a more reliable assessment of the extent of transferability of signals emerging from diaspora studies to populations living in Africa.

Our study leads on from this by making use of the larger GLGC 2021 dataset to assess the transferability of signals to continental African cohorts and in turn, inform whether large African ancestry based, or much larger global meta-analyses are better replicated in African populations. In addition, the research generates data to answer pivotal questions about whether the existing associations and polygenic risk scores (PRSs) are expected to perform differently in populations from different African regions. The transferability estimates also inform whether meta-analyses with African-American cohorts that is widely practised, is indeed the optimal approach for extending the sample size of GWASs from all African regions, or whether it is more suitable for specific African populations.

Results

To assess the transferability of genetic associations of lipid traits to African populations, we first checked the proportion of independent lead SNPs reported in the GLGC Multi-ancestry and the GLGC African-specific ancestry datasets that were included in the continental African studies (Table 1). For LDL-C, 13,533 independent lead SNPs were identified using linkage

disequilibrium (LD) clumping in the Multi-ancestry GLGC dataset. Of which, 5,683 SNPs were included in both the African GWAS datasets (AWI-Gen and Gurdasani meta-analysis). Similarly, of the 1,487 lead SNPs in the GLGC African-specific ancestry GWAS, 873 SNPs were found in both the continental African datasets. Similarly, about 40-60% of the independent lead SNPs for HDL-C and TG (both Multi-ancestry and African-specific ancestry datasets) were included in the African datasets.

Table 1. Number of independent lead SNPs in the GLGC datasets that were included in both the continental African studies

	Multi-ancestry		**African-specific Ancestry**	
Trait	**Number of independent lead SNPs**	**Lead SNPs found in both the African GWASs**	**Number of independent lead SNPs**	**Lead SNPs found in both the African GWASs**
LDL-C	13 533	5 683	1 487	873
HDL-C	16 502	7 708	1 051	599
TG	15 202	6 763	645	381

LDL-C=Low-Density Lipoprotein Cholesterol; HDL-C=High-Density Lipoprotein Cholesterol; TG=Triglyceride

The analysis of LDL-C loci in various African cohorts reveals distinct patterns in the replication of genetic associations. As seen in Figure 1 of Multi-ancestry LDL-C signals, a total of 77 SNPs were replicated at genome wide significance threshold ($p<5\times10^{-8}$), with the majority belonging to the AWI-Gen cohort, specifically AWI-Gen South but little to no replication from AWI-Gen West and East (1 and 0 SNPs, respectively). Replication estimated at a more relaxed replication threshold ($p<5x10^{-6}$) (See Supplementary Figures 1; 2 and 3 in Appendix A) followed a similar trend (Figure 1), with the AWI-Gen south and DDS cohorts having the most number of SNPs replicated at genome-wide significance threshold (11 and 12 respectively).

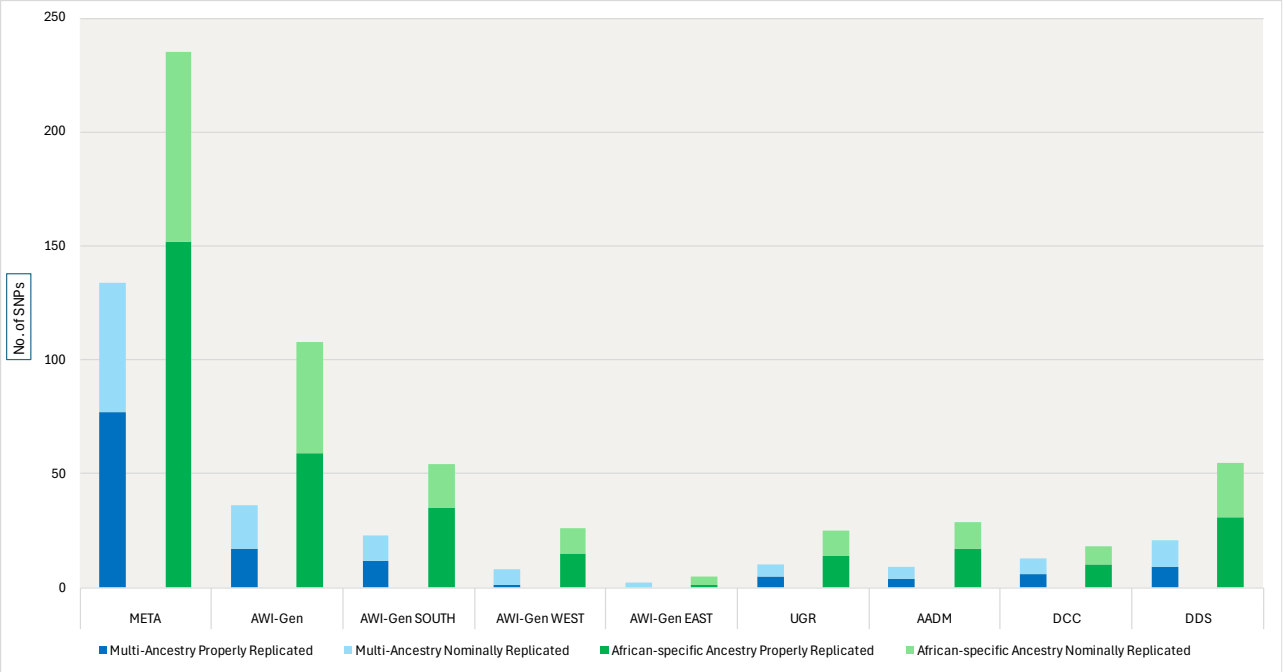


Figure 1. Number of LDL-C signals from the Multi-ancestry (blue) and African-specific ancestry (green) GWASs replicated at genome wide significance ($p<5\times10^{-8}$)(dark) and replication threshold $(5\times10^{-8}<p<5\times10^{-6})$ (light) in the African cohorts. META=AWI-Gen and Gurdasani meta-analysis; AWI-Gen=Africa Wits-INDEPTH Partnership for Genomic Research; AWI-Gen SOUTH=Africa Wits-INDEPTH Southern Africa; AWI-Gen WEST=Africa Wits-INDEPTH West Africa; AWI-Gen EAST=Africa Wits-INDEPTH East Africa; UGR= Uganda Genome Resource; AADM=Africa America Diabetes Mellitus; DCC=Durban Case Control Study; DDS=Durban Diabetes Study

African-specific ancestry GWAS signals demonstrated significantly higher replication rates across all continental African cohorts compared to Multi-ancestry GWAS based signals. Replication at both genome wide significant (dark green) and replication thresholds (light green) were highest in the combined African GWAS (META) cohort, which replicated the largest number of loci overall, followed by AWI-Gen combined and UGR. Multi-ancestry loci showed overall lower replication rates, with META leading in both genome wide significant (dark blue) and replication threshold signals (light blue). As expected, smaller cohorts, such as AADM, DCC, AWI-Gen SOUTH, AWI-Gen WEST, and AWI-Gen EAST, exhibited comparatively lower replication rates for both Multi-ancestry and African-specific analyses, demonstrating the importance of sample size in the ability to replicate signals.

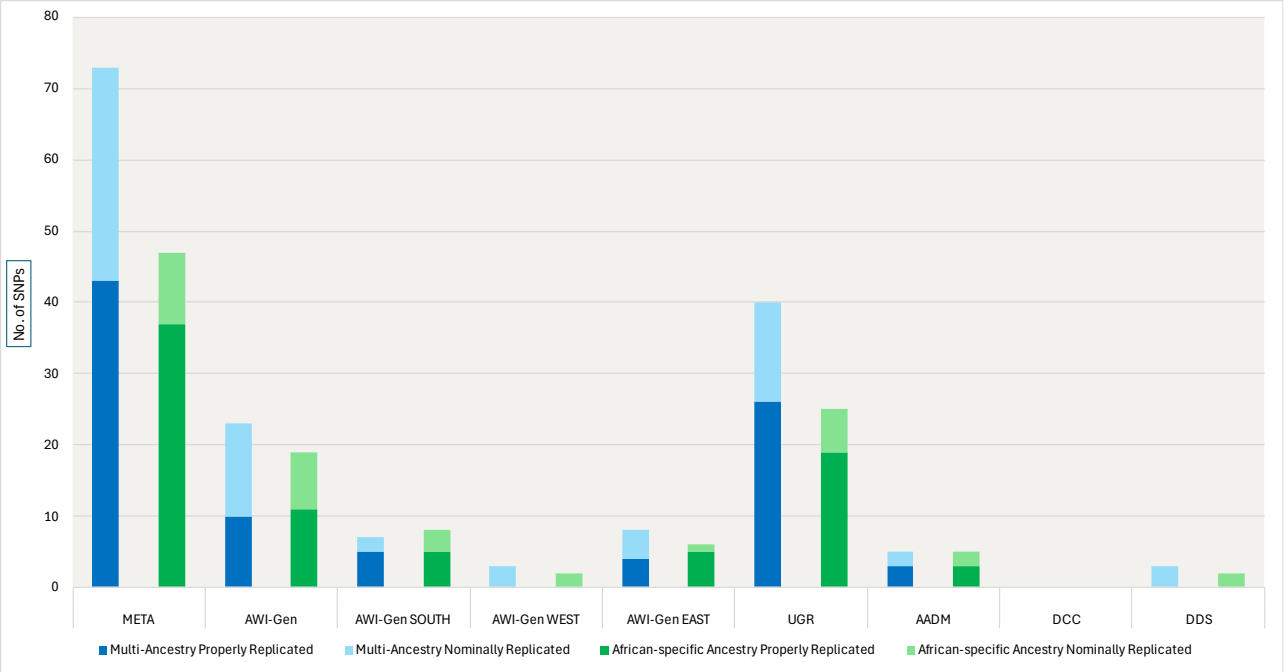


Figure 2. Number of HDL-C signals from the Multi-ancestry (blue) and African-specific ancestry (green) GWASs replicated at genome wide significance ($p<5\times10^{-8}$)(dark) and replication threshold ($5\times10^{-8}<p<5\times10^{-6}$) (light) in the African cohorts. META=AWI-Gen and Gurdasani meta-analysis; AWI-Gen=Africa Wits-INDEPTH Partnership for Genomic Research; AWI-Gen SOUTH=Africa Wits-INDEPTH Southern Africa; AWI-Gen WEST=Africa Wits-INDEPTH West Africa; AWI-Gen EAST=Africa Wits-INDEPTH East Africa; UGR= Uganda Genome Resource; AADM=Africa America Diabetes Mellitus; DCC=Durban Case Control Study; DDS=Durban Diabetes Study

Figure 2 illustrates the replication estimates of HDL-C loci across various continental African cohort GWAS. As observed for LDL-C the combined dataset (META) showed the highest number of replicated SNPs both at genome-wide significant (dark blue) and at the replication thresholds (light blue). Also, the signals from the African-specific ancestry GWASs showed higher replication rates compared to the Multi-ancestry-based cohort. However, in contrast to higher replication in LDL-C in the Southern African cohort, for HDL-C, the UGR showed significantly higher replication compared to other cohorts, even the full AWI-Gen dataset.

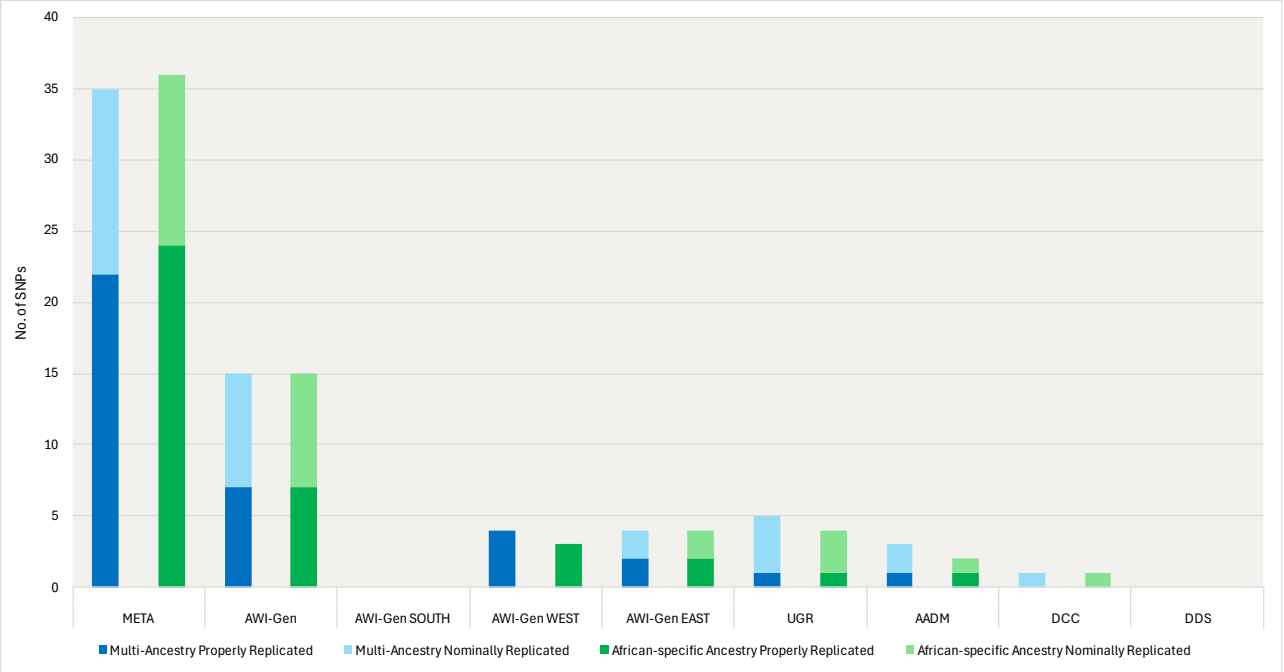


Figure 3. Number of TG signals from the Multi-ancestry (blue) and African-specific ancestry (green) GWASs replicated at genome wide significance ($p<5\times10^{-8}$) (dark) and replication threshold ($5\times10^{-8}<p<5\times10^{-6}$) (light) in the African cohorts. META=AWI-Gen and Gurdasani meta-analysis; AWI-Gen=Africa Wits-INDEPTH Partnership for Genomic Research; AWI-Gen SOUTH=Africa Wits-INDEPTH Southern Africa; AWI-Gen WEST=Africa Wits-INDEPTH West Africa; AWI-Gen EAST=Africa Wits-INDEPTH East Africa; UGR= Uganda Genome Resource; AADM=Africa America Diabetes Mellitus; DCC=Durban Case Control Study; DDS=Durban Diabetes Study

The stacked plot in Figure 3 illustrates the replication of TG genetic associations across various continental African cohorts. As observed for other lipid-traits, the combined African GWAS showed the highest number of replicated signals from the Multi-ancestry study, with 22 loci replicated at the genome wide significant threshold ($P < 5x10^{-8}$) and an additional 13 replicated at the replication threshold ($5x10^{-8} < P < 5x10^{-6}$). However, unlike other traits, here we did not observe any major difference between replication of Multi-ancestry and African-specific ancestry signals.

Figure 4 summarises a comparison of allele frequency signals identified in the GLGC African-specific ancestry GWAS (Figure 4 a, b, c) and the Multi-ancestry GWAS (Figure 4 d, e, f) to the frequencies observed in the AWI-Gen dataset (consisting of three AWI-Gen regional cohorts). As expected, the African-specific ancestry dataset shows higher concordance in comparison to the Multi-ancestry cohort across all three traits.

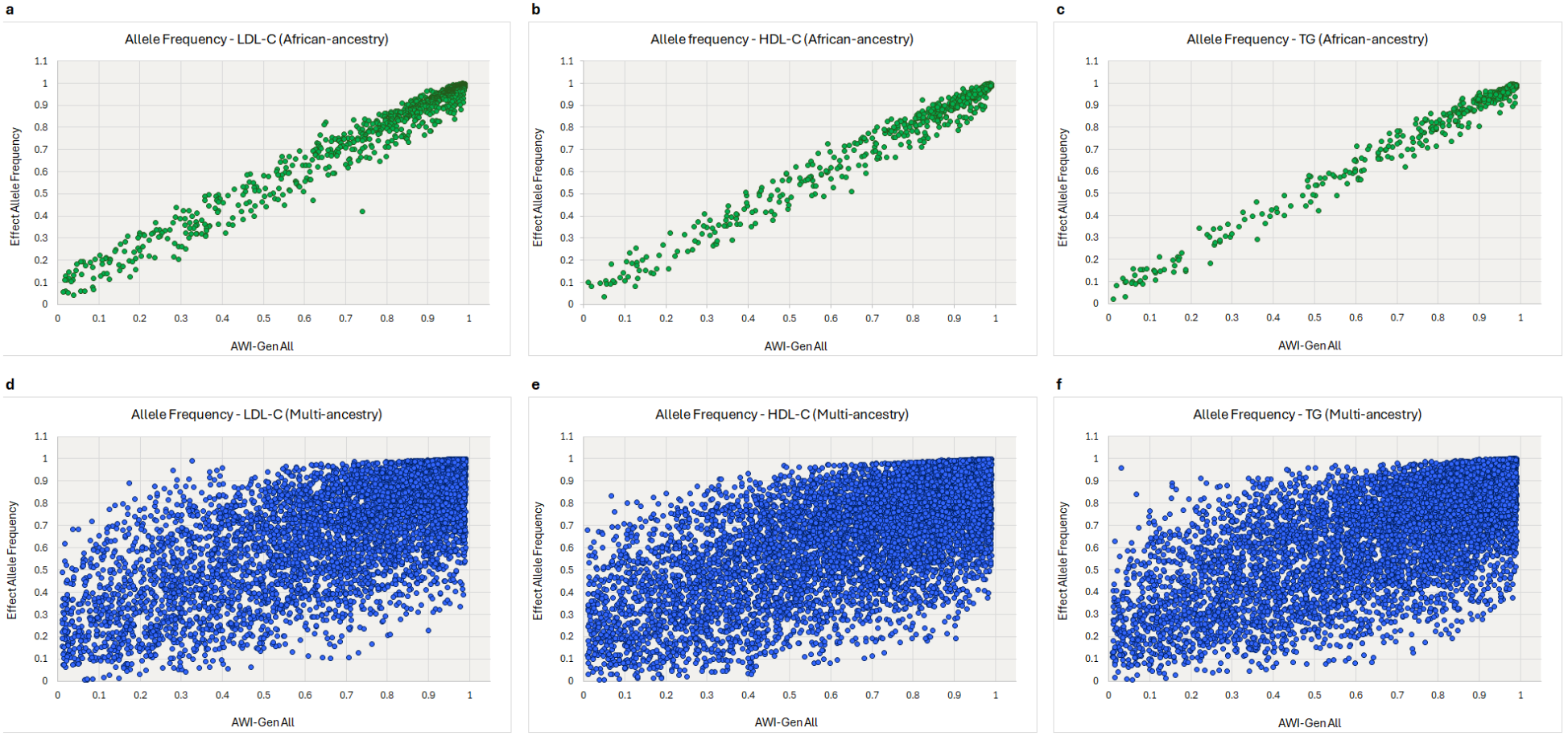


Figure 4. Scatter Plots showing comparison of allele frequencies. Allele frequencies for signals for (a) LDL-C (b), HDL- C (c) TG detected in the GLGC African ancestry cohorts compared to allele frequencies observed the AWI-Gen GWAS. Allele frequencies for signals for (d) LDL-C (e), HDL- C and (f) TG detected in the GLGC Multi-ancestry cohorts compared to allele frequencies observed the combined AWI-Gen GWAS. AWI-Gen=Africa Wits-INDEPTH Partnership for Genomic Research.



Figure 5. Scatter Plots showing comparison of beta-values. Beta-values of signals for (a) LDL-C (b), HDL- C (c) TG detected in the GLGC African ancestry cohorts compared to beta-values observed in the AWI-Gen GWAS. Beta-values for signals for (d) LDL-C (e), HDL- C and (f) TG detected in the GLGC Multi-ancestry cohorts compared to beta-values observed in the combined AWI-Gen GWAS. AWI-Gen=Africa Wits-INDEPTH Partnership for Genomic Research.

A similar comparison of beta-values (Figure 5) shows a much higher concordance between the beta values from the combined AWI-Gen dataset and GLGC African-specific ancestry GWAS (Figure 5 a, b, c) compared to the GLGC Multi-ancestry GWAS (Figure 5 d, e, f). These differences were more pronounced in the beta-values for HDL-C (correlation coefficient 0.56 in African-specific ancestry GWAS and 0.11 in Multi-ancestry GWAS) and TG (correlation coefficient 0.59 in African-specific ancestry GWAS and 0.13 in Multi-ancestry GWAS) compared to LDL-C (correlation coefficient 0.66 in African-specific ancestry GWAS and 0.33 in Multi-ancestry GWAS) suggesting that the impact of a substantial proportion of known signals on these traits could be restricted to particular ancestries only.

Discussion

The level of transferability of associations found in a GWAS based on one ancestry to other ancestries is key to generalizability of these associations and the applicability of polygenic risk score models. Previous investigation of transferability of lipid-traits to African populations has shown conspicuous variations, both among traits as well as among different African regions (8). Similarly, differences in transferability of PRS models for lipid traits to different African regions has also been noted (9). However, these estimates were based on GWAS summary statistics available at the time, which has been increased by more than 10 fold by the current phase of GLGC (1). Accordingly, the number of loci identified in the GLGC GWASs are several fold higher than previous studies. Therefore, we have estimated transferability based on this larger dataset and investigated whether any novel trends are revealed by this much larger discovery GWAS.

In the current study we have employed two thresholds for assessing transferability. The first is the standard genome-wide significant threshold of $5x10^{-8}$ for GWAS studies. The second is a replication threshold which in essence is an approximation of the Bonferroni correction of p-value. However, as two discovery GWASs identified different number of independent signals for the three traits, instead of using six different thresholds, we selected a threshold of $5x10^{-6}$ that is more conservative than the Bonferroni corrected p-value threshold estimates for all six GWASs (Multi-ancestry and African-specific ancestry GWAS for the three lipid traits) and provides scope for adjusting potential errors in determination of independence of signals. This replication p-value threshold therefore has generated a slightly conservative estimate of transferability and fine-tuning it could lead to higher estimates of transferability.

In the previous evaluation, the transferability estimates showed a correlation with sample size of the replication cohort with larger replication cohorts showing better transferability. In alignment, for all three traits and both GWASs (African-specific ancestry and Multi-ancestry) the combined continental African dataset (based on meta-analysis AWI-Gen and Gurdasani (N=~24000)) showed highest replication. This was followed by the next larger dataset, the full AWI-Gen dataset (N=~10600) for TG and LDL-C. However, in case of HDL-C the smaller regional cohort UGR (~6000) showed notably higher transferability instead. For all the three traits, a higher replication rate (number of SNPs replicated/number of independent lead SNPs present in both GLGC and replication cohorts) was observed for the GLGC African-specific ancestry GWASs compared to the much larger Multi-ancestry GWASs. We also observed a trend towards a regional bias with two of the Southern African cohorts replicating more signals for LDL-C, and UGR replicating more signals for HDL-C, compared to other regional GWASs. This was surprising given that based on genetic similarity, we expected to observe higher replication in West African cohorts. Interestingly for TG, among the regional cohorts, the Southern African cohort showed the lowest replication, highlighting the immense inter trait variability in the replication of associations in different African regions. While the involvement of factors like impact of unique ancestries, and region-specific gene-environment interactions can be speculated, further work will be needed to validate these results and understand the factors driving these differences.

The study also highlights that despite insufficient sample sizes, which ultimately reduce statistical power and the ability to detect genetic associations, for some of the traits the smaller cohorts such as DDS and AWI-Gen East showed considerably higher replication rates than other larger regional cohorts. Most of these differences were undetectable in the previous comparison (8) underlining the importance of conducting GWASs from different parts of the continent, even if with minimal sample size, as the unique ancestral composition in a region could allow for replication of signals that could be missed by much larger datasets from other parts for the continent. In addition to the size of the discovery cohort, the ancestral composition could be a key determinant for transferability of associations especially to non-European populations. However, the evaluations of lipid-trait signal transferability performed so far have been based on European and Multi-ancestry discovery cohorts (5,8). This was due to the lack of large ancestry-specific GWASs from non-European populations. The ancestry-specific summary statistics generated by the new GLGC study, enabled a more nuanced

estimation of the relationship between ancestral composition and sample size of discovery cohort and the transferability of signals. One of the key takeaways of this study is the notably higher transferability of African-specific ancestry LDL-C and HDL-C associations compared to those from Multi-ancestry GWAS. This is indicative of the distinct genetic architecture of African populations, which may not be adequately captured by global datasets and demonstrate that indeed smaller but more ancestrally similar GWASs could have a much higher transferability compared to much larger Multi-ancestry GWASs. The differences observed in the transferability of genetic associations across datasets, especially between Multi-ancestry and African-specific loci, further emphasise the importance of inclusive, diverse genomic research to advance health equity.

One of the major limitations of this study is that our approach for exact replication and including only those SNPs that were present in both Choudhury et al. 2021 and Gurdasani et al. 2019 summary stats, allowed us to assess the transferability about 40-60% of the signals from the GLGC study. While it might be impossible to assess the transferability of a sizable proportion of these signals, as these are either absent or extremely low frequency in continental African populations, use of alternative reference panel for imputation, amending minor allele frequency cut-offs and choice of alternative SNPs to represent LD blocks (i.e. credible-set based replication approach) could substantially increase the number of SNPs being investigated. Due to intrinsic complications involved in applying such approaches, we have not included these in the current analyses. However, consideration of these in future studies could not only increase the number of SNPs being assessed for transferability but also the confidence and robustness of the results. Also, the allele frequency and beta comparisons reported here are based on combined African meta-analysis that includes populations from different African regions. Given the level of variation in replication, regional based comparison of these estimates could be more informative and reveal additional trends and will be a major focus of our future work. Our study demonstrates an overall higher transferability and effect size concordance of African-specific ancestry signals from GLGC to diverse African cohorts compared to signals from a 15-fold larger Multi-ancestry GWAS and highlights major differences in transferability of different lipid-traits to the three African regions. These results point to the importance of including more diverse African cohorts in genetic research, with adequate sample sizes and broader regional coverage, to enhance the generalizability of findings. As genetic research progresses, particularly in the context of precision medicine, it is

crucial to adopt a tailored approach that reflects the unique genetic and environmental factors present in African populations. This approach could lead to more effective healthcare interventions and treatments tailored to the specific genetic makeup of African individuals.

Methods

Ethics

This study was approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (Wits) (Ethics Clearance Certificate - M240770)

GWAS summary statistics download

The first step involved downloading the summary statistics files for the lipid GWAS studies from the GWAS Catalog (https://www.ebi.ac.uk/gwas/home) (14). For the GLGC study (11), for each lipid trait (LDL, HDL and TG), the Global multi ancestry meta-analysis and the African ancestry meta-analysis summary statistics file were downloaded. For each of three lipid traits, summary statistics from seven African cohorts (7,8) – UGR (~6,000 participants from Uganda), AADM (~4,000 participants, predominantly from Nigeria), DDS (~1,000 participants from Durban), DCC (~1,300 participants from Durban), AWI-Gen East (~2,000 participants from Kenya), AWI-Gen West (~4,000 participants from Ghana and Burkina Faso) and AWI-Gen South Africa (~5,000 participants from South Africa) were also downloaded (Figure 6).

We restricted our analysis to measured HDL, measured TG and calculated LDL, and not measured total cholesterol (TC), since TC is a major component of the calculated LDL. Moreover, a previous study (8) showed that both calculated LDL and measured TC show very similar GWAS results, therefore TC was not needed in our analysis.

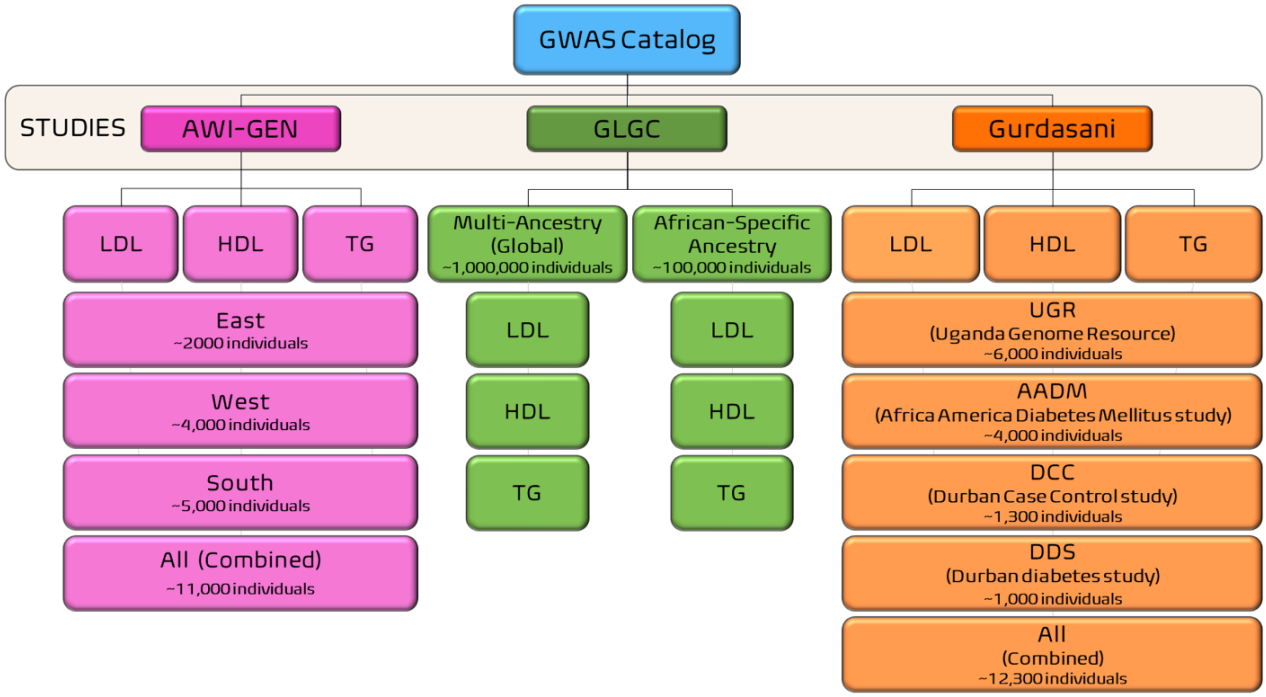


Figure 6. The list of summary statistics files downloaded from GWAS Catalog for each lipid trait.

Identification of independent lead signals

Independent lead signals from each of the Multi-ancestry and African-specific ancestry GLGC summary statistics files were identified using the Linkage Disequilibrium (LD) clumping function in PLINK 2.0 (15). The following settings were used for LD clumping: *--clump-p1* $5e^{-8}$ (clumps are formed around central "lead variants", starting with those with the lowest p-value, the maximum p-value allowed for those SNPs here is $5e^{-8}$); *--clump-kb 500* (the maximum distance in KB from the lead variant for SNPs to be considered to be clumped with it) and *--clump-r2 0.5* (variants with LD r^2>0.5 with the lead variant will be included in the clump if they are within a distance of 500kb).

Transferability assessment

For each trait, the independent lead SNPs (identified from Multi-ancestry and African-specific ancestry GLGC summary statistics files) were extracted from each of the African cohort summary statistics files using the R merge function. Figure 7 demonstrates the steps involved in identifying the final list of SNPs that are used for subsequent analysis. To ensure comparability of datasets, only the signals that are present in both the Gurdasani and AWI-Gen studies were considered for the transferability analysis.

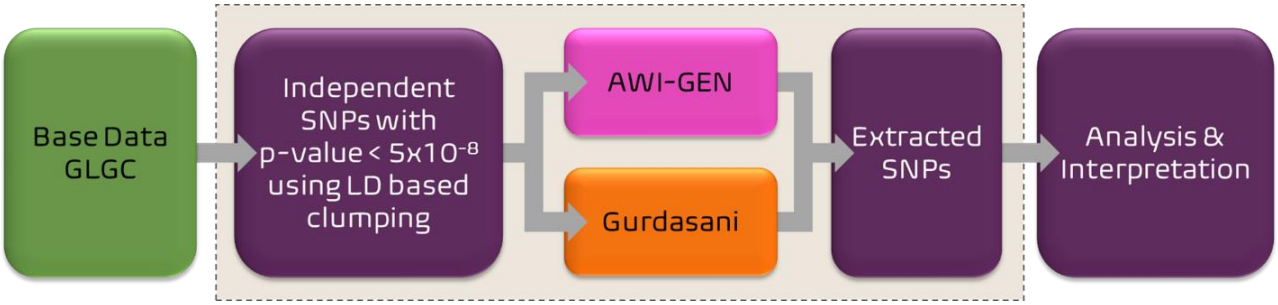


Figure 7. Flow diagram showing the process of extraction and analysis of GWAS summary statistics data

We have employed two thresholds for assessing transferability. The first is the standard genome-wide significant threshold of $p< 5x10^{-8}$ for GWAS studies. The second is a replication threshold, which is an approximation of the Bonferroni correction of p-value i.e. if an independent signal showed a p-value less than (*0.05/Number of independent lead signals*) in an African cohort, it was considered to be transferable to that cohort. The replication threshold was calculated using in-house shell scripts for all the independent target and base datasets as shown in Figure 6. The level of transferability from the Multi-ancestry GLGC and African-specific ancestry GLGC to the full AWI-Gen and Gurdasani meta-analyses was also performed.

Assessment for transferability was done for each of the three lipid traits: HDL, LDL and TG. This was followed by comparing the trends in transferability from the two GLGCs to the three continental African regions the cohorts fall under (East, West and South), also for each of the three lipid traits to investigate intra-continental variation in transferability. (Figure 8)

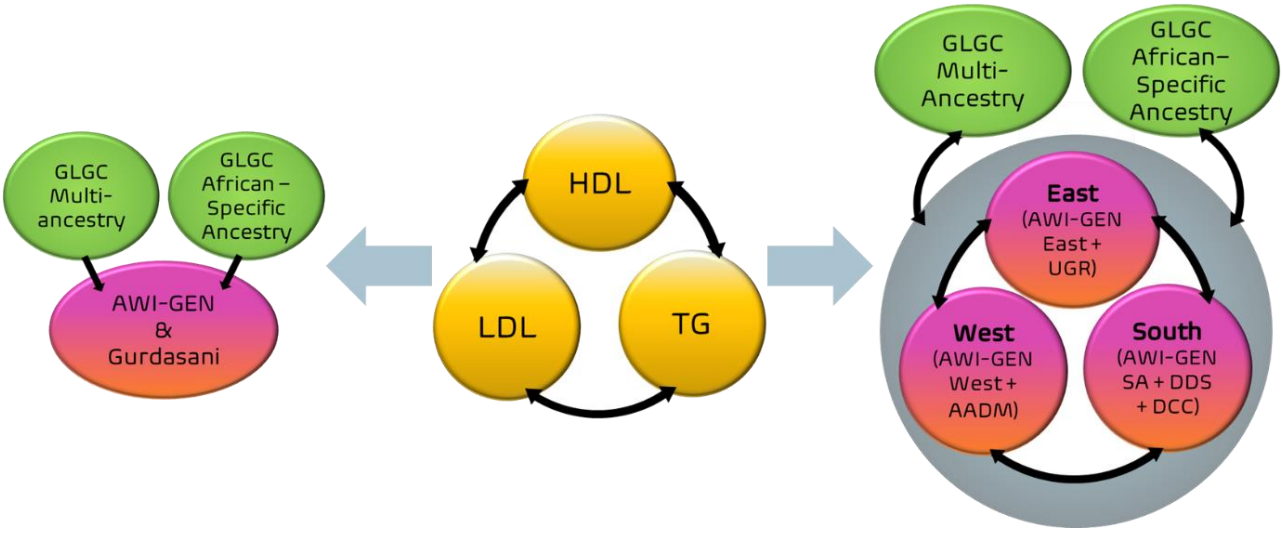


Figure 8. Illustration of the analysis approach, which evaluates the transferability of genetic signals from multi- and African-specific ancestry GLGC cohorts to African populations (AWI-Gen and Gurdasani), and compares HDL, LDL, and TG associations across East, West, and South African regions

Assessing heterogeneity of effect size and risk allele frequencies across African cohorts

Lastly, for the final list of independent lead SNPs in the GLGC Multi-ancestry and African-ancestry dataset, the risk alleles, risk allele frequencies and risk allele effect sizes were compared against the same for AWI-Gen and Gurdasani data. For some SNPs, the risk allele and non-risk alleles were swapped between the GLGC and continental African studies. We changed the direction of beta in such cases, as well as plotted the frequency of the other allele to maintain consistency.

Data availability statement

The GWAS summary statistics used in this study are downloaded from GWAS Catalog (https://www.ebi.ac.uk/gwas/summary-statistics). (14)

Acknowledgments

We would like to thank Prof Scott Hazelhurst for providing access to the Wits computational cluster.

Conflicts of Interest Statement

The authors have no conflict of interest to declare.

Funding

There is no funding required for this project since it involves analysis of publicly available data downloaded from GWAS Catalog as well as data that is already on the Wits Core computing cluster. All the in-silico tools being used are publicly available and free of charge.

References

1. Graham, S.E., Clarke, S.L., Wu, K.-H.H., Kanoni, S., Zajac, G.J., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., *et al.* (2021) The power of genetic diversity in genome-wide association studies of lipids. *Nature*, **600**, 675–679.

2. Bush, W.S. and Moore, J.H. (2012) Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*, **8**, e1002822.

3. Marees, A.T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C. and Derks, E.M. (2018) A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, **27**, e1608.

4. Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T. and Posthuma, D. (2021) Genome-wide association studies. *Nat Rev Methods Primers*, **1**, 1–21.

5. Kuchenbaecker, K., Telkar, N., Reiker, T., Walters, R.G., Lin, K., Eriksson, A., Gurdasani, D., Gilly, A., Southam, L., Tsafantakis, E., *et al.* (2019) The transferability of lipid loci across African, Asian and European cohorts. *Nat Commun*, **10**, 4330.

6. Gaziano, T.A., Bitton, A., Anand, S., Abrahams-Gessel, S. and Murphy, A. (2010) Growing Epidemic of Coronary Heart Disease in Low- and Middle-Income Countries. *Curr Probl Cardiol*, **35**, 72–115.

7. Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C.S., Prado-Martinez, J., Bouman, H., Abascal, F., Haber, M., Tachmazidou, I., *et al.* (2019) Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell*, **179**, 984-1002.e36.

8. Choudhury, A., Brandenburg, J.-T., Chikowore, T., Sengupta, D., Boua, P.R., Crowther, N.J., Agongo, G., Asiki, G., Gómez-Olivé, F.X., Kisiangani, I., *et al.* (2022) Meta-analysis of sub-Saharan African studies provides insights into genetic architecture of lipid traits. *Nat Commun*, **13**, 2578.

9. Kamiza, A.B., Toure, S.M., Vujkovic, M., Machipisa, T., Soremekun, O.S., Kintu, C., Corpas, M., Pirie, F., Young, E., Gill, D., *et al.* (2022) Transferability of genetic risk scores in African populations. *Nat Med*, **28**, 1163–1166.

10. Ramsay, M., Crowther, N., Tambo, E., Agongo, G., Baloyi, V., Dikotope, S., Gómez-Olivé, X., Jaff, N., Sorgho, H., Wagner, R., *et al.* (2016) H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Glob. Health Epidemiol.*, **1**, e20.

11. Global Lipids Genetics Consortium (2013) Discovery and refinement of loci associated with lipid levels. *Nat Genet*, **45**, 1274–1283.

12. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., *et al.* (2019) Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, **570**, 514–518.

13. Choudhury, A., Aron, S., Botigué, L.R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y.J., *et al.* (2020) High-depth African genomes inform human migration and health. *Nature*, **586**, 741–748.

14. GWAS Catalog https://www.ebi.ac.uk/gwas/home (accessed Apr 12, 2024).

15. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*, **81**, 559–575.

Appendices

Appendix A: Supplementary Tables and Figures

Supplementary Table 1: Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for LDL-C

Supplementary Table 2: Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for HDL-C

Supplementary Table 3: Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for TG

Appendix B: Ethics Clearance Certificate - M240770

Appendix C: Approved Research Protocol

Appendix D: Plagiarism Declaration and Turnitin Report

Appendix E: Author Guidelines for Human Molecular Genetics Journal

Appendix A: Supplementary Tables and Figures

Supplementary Table 1: *Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for LDL-C*

COHORTS	META	AWI-Gen	AWI-Gen SOUTH	AWI-Gen WEST	AWI-Gen EAST	UGR	AADM	DCC	DDS
Multi-ancestry LDL-C									
Properly Replicated (P<5x10-8)	77	17	12	1	0	5	4	6	9
Nominally Replicated (5x10-8<P<5x10-6)	57	19	11	7	2	5	5	7	12
African-specific ancestry LDL-C									
Properly Replicated (P<5x10-8)	152	59	35	15	1	14	17	10	31
Nominally Replicated (5x10-8<P<5x10-6)	83	49	19	11	4	11	12	8	24

META=AWI-Gen and Gurdasani meta-analysis; AWI-Gen=Africa Wits-INDEPTH Partnership for Genomic Research; AWI-Gen SOUTH=Africa Wits-INDEPTH Southern Africa; AWI-Gen WEST=Africa Wits-INDEPTH West Africa; AWI-Gen EAST=Africa Wits-INDEPTH East Africa; UGR= Uganda Genome Resource; AADM=Africa America Diabetes Mellitus; DCC=Durban Case Control Study; DDS=Durban Diabetes Study; LDL-C=Low-Density Lipoprotein Cholesterol

Supplementary Table 2: *Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for HDL-C*

COHORTS	META	AWI-Gen	AWI-Gen SOUTH	AWI-Gen WEST	AWI-Gen EAST	UGR	AADM	DCC	DDS
Multi-ancestry HDL-C									
Properly Replicated (P<5x10-8)	43	10	5	0	4	26	3	0	0
Nominally Replicated (5x10-8<P<5x10-6)	30	13	2	3	4	14	2	0	3
African-specific ancestry HDL-C									
Properly Replicated (P<5x10-8)	37	11	5	0	5	19	3	0	0
Nominally Replicated (5x10-8<P<5x10-6)	10	8	3	2	1	6	2	0	2

META=AWI-Gen and Gurdasani meta-analysis; AWI-Gen=Africa Wits-INDEPTH Partnership for Genomic Research; AWI-Gen SOUTH=Africa Wits-INDEPTH Southern Africa; AWI-Gen WEST=Africa Wits-INDEPTH West Africa; AWI-Gen EAST=Africa Wits-INDEPTH East Africa; UGR= Uganda Genome Resource; AADM=Africa America Diabetes Mellitus; DCC=Durban Case Control Study; DDS=Durban Diabetes Study; HDL-C=High-Density Lipoprotein Cholesterol

Supplementary Table 3: *Number of SNPs replicated at different p-value thresholds in the African cohorts, from the multi- and African-Ancestry GLGC for TG*

COHORTS	META	AWI-Gen	AWI-Gen SOUTH	AWI-Gen WEST	AWI-Gen EAST	UGR	AADM	DCC	DDS
Multi-ancestry TG									
Properly Replicated (P<5x10-8)	22	7	0	4	2	1	1	0	0
Nominally Replicated (5x10-8<P<5x10-6)	13	8	0	0	2	4	2	1	0
African-specific ancestry TG									
Properly Replicated (P<5x10-8)	24	7	0	3	2	1	1	0	0
Nominally Replicated (5x10-8<P<5x10-6)	12	8	0	0	2	3	1	1	0

META=AWI-Gen and Gurdasani meta-analysis; AWI-Gen=Africa Wits-INDEPTH Partnership for Genomic Research; AWI-Gen SOUTH=Africa Wits-INDEPTH Southern Africa; AWI-Gen WEST=Africa Wits-INDEPTH West Africa; AWI-Gen EAST=Africa Wits-INDEPTH East Africa; UGR= Uganda Genome Resource; AADM=Africa America Diabetes Mellitus; DCC=Durban Case Control Study; DDS=Durban Diabetes Study; TG=Triglyceride

Appendix B: Ethics Clearance Certificate - M240770

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research and Innovation)

TO: Ms M Crosby
School of Pathology
Division of Human Genetics
Faculty of Health Sciences

E-mail: Michaela.e.Crosby @gmail.com

CC: Supervisor: Dr D Sengupta and Professor A Choudhury
Dhriti.Sengupta@wits.ac.za
and <HREC-Medical Research Office@wits.ac.za>

FROM: Mr Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 29 August 2024

REF: R14/49

PROTOCOL NO: **M240770** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *In depth survey of the transferability of genetic associations of lipid traits to African populations*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R49 Ms M Crosby

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M240770

NAME: (Principal Investigator) Ms M Crosby **DEGREE:** MSc

DEPARTMENT: School of Pathology
Division of Human Genetics
Faculty of Health Sciences

TITLE: *In depth survey of the transferability of genetic associations of lipid traits to African populations*

DATE CONSIDERED: Ad hoc

DECISION: Approved unconditionally

CONDITIONS: Sub-study under M22/10/108

NOTE: If contact information regarding student study participants is required, please contact the Registrar's office <Nicoleen.Potgieter@wits.ac.za>

SUPERVISORS: Dr D Sengupta and Professor A Choudhury

APPROVED BY:
Professor P Ruff, Co-Chairperson, HREC (Medical)

DATE OF APPROVAL: 29 August 2024 **EXPIRY DATE:** 28 August 2029

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report** in the format available at https://wits.ac.za/research/researcher-support/research-ethics/ethics-committees/. This report is due annually, for the duration of the Clearance Certificate, beginning on the first anniversary of Date of Approval above. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator

31 August 2024
Date

Appendix C: Approved Research Protocol

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Private Bag 3 Wits, 2050
Fax: 027117172119
Tel: 02711 7172076

Reference: Mrs Sandra Benn
E-mail: sandra.benn@wits.ac.za

30 June 2024
Person No: 2308127
PAG

Miss ME Crosby
11 Lantana Road
Morningside Manor Ext. 2
2196
South Africa

Dear Miss Michaela Crosby

Master of Science in Medicine: Approval of Title

We have pleasure in advising that your proposal entitled *In depth survey of the transferability of genetic associations of lipid traits, to African populations* has been approved. Please note that any amendments to this title have to be endorsed by the Faculty's higher degrees committee and formally approved.

Yours sincerely

Mrs Sandra Benn
Faculty Registrar
Faculty of Health Sciences

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

CANDIDATE'S SURNAME: [Please print] **Crosby**	FIRST NAME/S: **Michaela Emma**	STUDENT NUMBER: **2308127**
CURRENT QUALIFICATIONS: **BHSc (Hons) Bioethics & Health Law; BSc Human Physiology, Genetics & Psychology**		
TEL:	CELL: **0798576514**	E-MAIL: **2308127@students.wits.ac.za** FAX:
DEGREE FOR WHICH PROTOCOL IS BEING SUBMITTED: **MSc (Med) Genomic Medicine**		
PART-TIME OR FULL-TIME: **Full-time**		
FIRST REGISTERED FOR THIS DEGREE:	TERM : **Term 1**	YEAR: **2024**
DEPARTMENT: **Human Genetics, School of Pathology**		
TITLE OF PROPOSED RESEARCH: **In depth survey of the transferability of genetic associations of lipid traits, to African populations**		
CANDIDATE'S SIGNATURE: 		DATE: **28/03/2024**
SUPERVISOR 1 (NAME & SURNAME): **Dhriti Sengupta**		% Supervision **60%**
SUPERVISOR'S QUALIFICATIONS **PhD**		
SUPERVISOR'S DEPARTMENT **Sydney Brenner Institute of Molecular Bioscience (SBIMB)**		
SUPERVISOR'S ADDRESS / TEL / E-MAIL: **Dhriti.sengupta@wits.ac.za**		
SUPERVISOR 2 (NAME & SURNAME): **Ananyo Choudhury**		% Supervision **40%**
SUPERVISOR'S QUALIFICATIONS **PhD**		
SUPERVISOR'S ADDRESS / TEL / E-MAIL: **Ananyo.choudhury@wits.ac.za**		
SUPERVISOR 3 (NAME & SURNAME):		% Supervision
SUPERVISOR'S QUALIFICATIONS		
SUPERVISOR'S ADDRESS / TEL / E-MAIL:		

SYNOPSIS OF RESEARCH:(Brief summary of proposed research project; between 200-300 words only; with sub-headings: an introduction and justification for study, aim/s, proposed methodology and expected outcome/s)
[Use reverse side of this page if more space is required]

Introduction and Justification

One of the key questions in genetics of complex traits is whether the same variant always has the same or similar effect on a phenotype across global populations. So far, the transferability of genetic associations detected in European and US based population has been shown to be lowest in African populations. Moreover, some studies have reported the possibility of considerable differences in the

level of transferability of European based associations even to different African regions. The level of transferability has also been found to vary widely between traits.

Aims

Our study aims to investigate the transferability of genetic associations from the largest meta-analysis for lipid traits (LDL, HDL, and Triglyceride), to seven African ancestry cohorts originating from three African regions. We will also compare effect size magnitudes, effect size directions and risk allele frequencies of key lipid signals between the African cohorts.

Methodology

The GWAS summary statistics for LDL-C, HDL-C and Triglycerides from the GLGC study, the AWI-Gen study (East, West and South) and Gurdasani study (four cohorts UGR, AADM, DDS, DCC) will be downloaded from GWAS Catalog. Independent genome-wide significant signals (*p-value* < $5x10^{-8}$) will be filtered from the GLGC summary stat file using the LD clumping function in PLINK 2.0. A custom script will record the number of such signals(N). The transferability of these signals will then be assessed in the seven cohorts, if a signal is observed to have *p-value* < 0.05/N in any of the cohorts, it will be considered as being transferable to the cohort. The level of variation in transferability to the seven cohorts as well as the transferability to three African regions will be assessed. Finally, the trends in transferability will also be compared between the three lipid traits.

Expected outcome

The proposed study is expected to provide insights into the suitability of currently existing lipid trait risk models in African populations.

WITS ETHICS NOT REQUIRED: WITS ETHICS PENDING: WITS ETHICS APPROVED: (circle appropriate symbol)* ***Please note the final human ethics clearance certificate or animal ethics certificate must be available prior to starting research**	Yes (No) (Yes) No Yes No	IF Y SUPPLY ETHICS CLEARANCE CERTIFICATE AS ATTACHMENT AND INCLUDE ETHICS NUMBER HERE:
As supervisor/s, I/we confirm that I have read the protocol which has been submitted for assessment.		
SIGNATURE OF SUPERVISOR/S:	Khuti Sengupta	Ananyo Choudhury
SIGNATURE PG OFFICE STAFF ……	REGISTERED YES….. NO…..	STAMP



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

In depth survey of the transferability of genetic associations of lipid traits, to African populations

Research Protocol

Michaela Crosby (2308127)

Master of Science (Med) Genomic Medicine

University of Witwatersrand

Supervisors:

Dr Dhriti Sengupta, PhD

Sydney Brenner Institute for Molecular Bioscience (SBIMB)

Prof Ananyo Choudhury, PhD

Sydney Brenner Institute for Molecular Bioscience (SBIMB)

Table of Contents

1 Introduction 1

1.1 Transferability of Signals 2

1.1.1 What Is Transferability? 2

1.1.2 Why Is Transferability an Issue? 2

1.2 GWAS for Lipid Traits 3

1.2.1 Global Lipids Genetics Consortium 3

1.2.2 Lipid GWAS in Gurdasani et al. 2019 study 3

1.2.3 Lipid GWAS in the AWI-Gen Study 4

2 Study Aims and Objectives 5

2.1 Aim 5

2.2 Objectives 5

2.2.1 Primary Objective 5

2.2.2 Secondary Objective 6

3 Methods 6

4 Data Analysis 8

5 Ethics 8

6 Timing 9

7 Funding 10

8 Limitations 10

9 References 11

1 Introduction

Despite advancements in prevention and treatment, cardiovascular disease (CVD) remains the top cause of death worldwide (Graham et al., 2021). Blood lipid levels are heritable genetic risk factors for CVD and vary globally due to environmental and genetic influences. For complex traits such as lipid biomarker levels, genome wide association studies (GWASs) that study genetic variation across the genome are employed to identify genetic risk factors (Bush and Moore, 2012). This is done through identifying single-nucleotide polymorphisms (SNPs) whose allele frequencies vary systematically as a function of phenotypic trait (Marees et al., 2018; Uffelmann et al., 2021).

GWASs have successfully identified hundreds of genetic associations for lipid traits such as Low-Density Lipoprotein-Cholesterol (LDL-C), High-density Lipoprotein-Cholesterol (HDL-C) and Triglyceride (TG) levels. However, most GWASs to date have been performed in European ancestry populations, primarily residing in Europe and North America, lacking representation of diverse populations (Kuchenbaecker et al., 2019).

About 75% of CVD-associated deaths occur in low- and middle-income countries where incidences are rising (Gaziano et al., 2010). A sufficient portion of mortality due to CVD can be prevented or at least delayed by simple lifestyle modification and medical intervention. An in-depth characterisation of genetic variants that are associated to a trait in a particular population, has the immense potential for enabling the stratification of individuals by risk and prioritized targeting of individuals at high risk. Consequently, it is important to determine whether SNPs associated with cardiometabolic traits such as lipid traits, that are key markers of CVDs, are transferable to other populations.

1.1 Transferability of Signals

1.1.1 What Is Transferability?

The transferability of GWASs signals refers to the ability of the genetic basis of a trait or phenotype to be shared across different populations to infer the risk of certain variants causing disease (Kuchenbaecker et al., 2019). Transferability refers to whether the results from a GWAS performed on one population could be transferred/replicated to other populations of other ancestries.

1.1.2 Why Is Transferability an Issue?

Thus far, the issue lies in the fact that the transferability of genetic associations detected in European and US-based populations has been shown to be lowest in African populations, with recent studies showing notable differences in transferability even between various regions within Africa (Gurdasani et al., 2019; Choudhury et al., 2022; Kamiza et al., 2022). This affects the generalisability and applicability of genetic findings, such as the identification of risk allele variants for a particular trait, to different population groups.

The lack of transferability of GWAS findings between populations can be explained by a combination of factors including differences in allele frequencies (e.g. when one SNP is commonly present in one population but extremely rare or absent in other population), genetic architecture (e.g. caused by differences in linkage disequilibrium between a variant and causal variants), effect sizes (e.g. caused by gene-by-gene or gene-by-environment interactions) and statistical power (when a GWAS has a smaller sample size and is not powered enough to detect associations observed in larger GWAS).

The results from GWASs are made freely accessible through the GWAS Catalog database (GWAS Catalog, no date) as summary statistics files. These files include information on SNP (chromosome, position, SNP ID and alleles), effect size (defined as the contribution of a SNP to the genetic variance of the trait), allele frequency and

p-value. Therefore, the assessment of transferability based on summary statistics data is simple, straightforward and computationally efficient.

1.2 GWAS for Lipid Traits

In this section we describe the largest global GWASs as well as two major GWASs based on lipid traits that were performed in continental African populations.

1.2.1 Global Lipids Genetics Consortium

The Global Lipids Genetics Consortium (GLGC) is a large-scale GWAS study that conducted a meta-analysis involving over 1,650,000 individuals across five genetic ancestry groups to identify variants associated with five key lipid traits across different populations, including African, East Asian, European, Hispanic and South Asian. These lipid traits included LDL-C, HDL-C, TGs, total cholesterol (TC) and non-high-density lipoprotein cholesterol (non-HDL-C) (Graham et al., 2021). The study revealed that utilising diverse ancestries improves signal transferability to African populations and genetic risk factor predictions (Graham et al., 2021). Of particular interest is the African-specific ancestry sub-cohort that consist of 100,000 individuals including African Americans, Afro-Caribbeans, and African unspecified individuals.

1.2.2 Lipid GWAS in Gurdasani et al. 2019 study

The landmark study by Gurdasani and colleagues was one of the major milestones in African GWASs (Gurdasani et al., 2019). Based primarily on a population cohort from Uganda, known as the Uganda Genome Resource (UGR), the authors performed GWASs on more than 20 traits that included body composition traits and biochemical traits including three lipid traits. The study also included a meta-analysis of these summary statistics with summary statistics from three other continental African cohorts – the Durban Case Control (DCC), Durban Diabetes study (DDS) and African American Diabetes Mellitus (AADM) study (based predominantly in Nigeria). Though sample size difference between cohorts and some of them being based on diabetes

patients was not optimal for the GWAS, the results provided baseline data for future comparisons.

1.2.3 Lipid GWAS in the AWI-Gen Study

Africa Wits-INDEPTH Partnership for Genomic Research (AWI-Gen) is a pan-African population cross-sectional cohort with a focus on understanding the genetic and environmental risk factors that predispose older adults to cardiometabolic diseases (CMDs) in Africa (Ramsay et al., 2016). The cohort consists of ~12,000 participants between the ages of 40 to 60 years sampled across from four countries - Burkina Faso, Ghana, Kenya and South Africa. The AWI-Gen genetic data was used to perform a GWAS for fasting serum TC, LDL-C, HDL-C, and TG on 10,603 individuals (Choudhury et al., 2022). The study also explored different approaches that could be used to address the challenge of managing complex population substructure between African regions, in Pan African GWASs.

The AWI-Gen lipid GWAS publication also included a comprehensive assessment of transferability of GWAS signals from some of the largest global GWAS studies available at that time (GLGC (Global Lipids Genetics Consortium, 2013) and PAGE (Wojcik et al., 2019)) to the cohorts included the AWI-Gen and Gurdasani et al. 2019 studies. The results demonstrated that an increase in sample size of both the discovery cohort and the replication cohort could lead to more robust transferability estimates. The study also noted considerably regional variation in transferability of signals to different African regions. However, the GLGC study dataset that was assessed was from a study conducted in 2013 and had a smaller sample size (~100,000 individuals total).

The GLGC 2021 dataset, which has almost a 16-fold higher sample size compared to the previous study, has the potential to lead to major revision in these estimates. Moreover, the availability of summary stats from a sizable African-Ancestry sub-cohort

enables a more reliable assessment of the extent of transferability of signals emerging from diaspora studies to populations living in Africa.

The proposed study would inform researchers whether large African ancestry based or much larger global meta-analysis based PRSs could be more informative for predicting risk in African populations. Also, the research will generate data to answer pivotal question about whether the existing PRSs are expected to perform differentially in populations from different African regions. The transferability estimates would also inform whether meta-analysis with African-American cohorts that is widely practised is indeed the optimal approach for extending the sample size of GWASs from all African regions or is more suitable for particular African populations.

2 Study Aims and Objectives

2.1 Aim

To investigate the transferability of genetic risk associations from the largest GWAS meta-analysis for three lipid traits (LDL-C, HDL-C and Triglycerides from the GLGC 2021 study) to four African ancestry cohorts included in the Gurdasani study (UGR, DDS, DCC, AADM) study as well as three regional cohorts (East, West and South African) from the AWI-Gen study. We also aim to compare the effect size magnitudes, effect size directions and risk allele frequencies of key lipid signals between the African cohorts.

2.2 Objectives

2.2.1 Primary Objective

Objective 1: Obtain the summary statistic files for each lipid trait from each study (**Figure 1**).

Objective 2: Use LD clumping to identify the most significant independent lead signals from each genomic region (with a *p-value* threshold of $5x10^{-8}$) and extract the details for these signals from each African cohort summary stat file (**Figure 2**).

Objective 3: Based on Bonferroni adjusted *p-value* threshold (and effect size direction consistency), assess the level of transferability of signals, for each of the three lipid traits, from GLGC to these African cohorts.

2.2.2 Secondary Objective

Compare the effect size magnitude and direction, as well as risk allele frequencies of the three lipid traits' associated SNPs between the seven African cohorts.

3 Methods

Primary Objective 1: Data download

The first step will involve downloading the summary statistics files from GWAS Catalog from the lipid GWAS studies. For the GLGC study, for each lipid trait (LDL, HDL and TG) the Global multi ancestry meta-analysis and the African Ancestry meta-analysis summary statistics file (e.g. https://www.ebi.ac.uk/gwas/studies/ GCST90239656 for LDL) will be downloaded. For each of three lipid traits, summary statistics from seven African cohorts – UGR (~6,000 participants from Uganda), AADM (~4,000 participants, predominantly from Nigeria), DDS (~1,000 participants from Durban), DCC (~1,300 participants from Durban), AWI-Gen East (~2,000 participants from Kenya), AWI-Gen West (~4,000 participants from Ghana and Burkina Faso) and AWI-Gen South Africa (~5,000 participants from South Africa) will be also downloaded (**Figure 1**).

We have restricted our analysis to measured HDL, measured TG and calculated LDL (and not measured TC) since TC is a major component of the calculated LDL. Moreover, our previous study (Choudhury *et al.*, 2022) has shown that both calculated LDL and measured TC show very similar GWAS results.

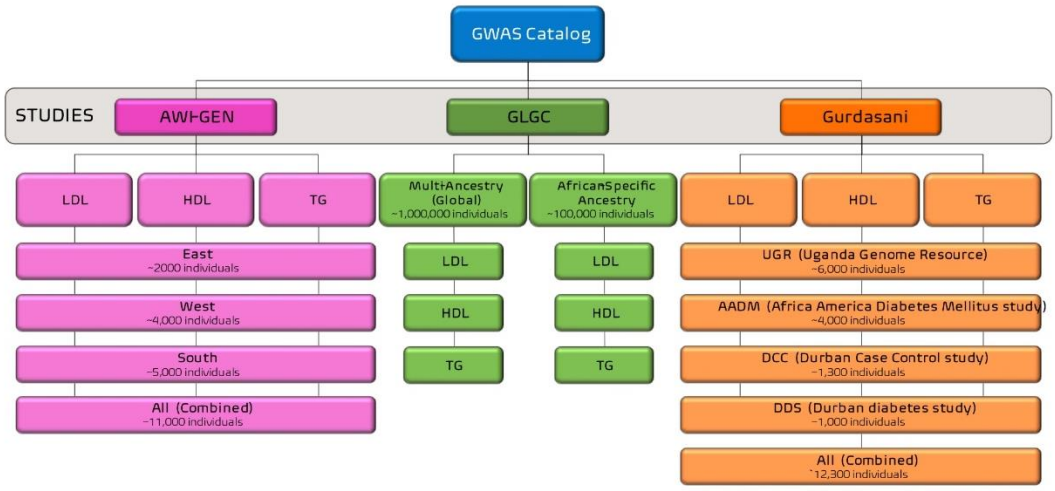


Figure 1: *The list of summary statistics files to be downloaded from GWAS Catalog for each lipid trait.*

Primary Objective 2: Identification of independent lead signals

The most significant signal for each genomic region (if *p-value* < $5x10^{-8}$) in the multi-ancestry GLGC file will be identified using the Linkage Disequilibrium (LD) clumping function in PLINK 2.0 (Purcell et al., 2007). This will provide us with a list of independent lead signals from each of the GLGC summary statistics file (**Figure 2**). This step will be repeated for the African-ancestry GLGC file, to identify the independent lead signals contained in this file.

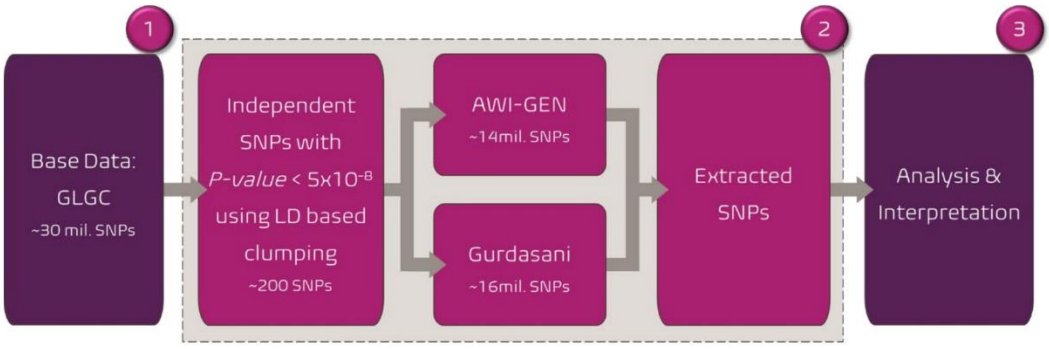


Figure 2: *Flow diagram showing the process for achieving the primary objectives, using rough SNP estimates*

Primary Objective 3: Transferability assessment

Next, we would only extract the SNPs included in the list of independent lead signals (generated in Objective 2) from each of the African cohort summary stat files for

analysis. To ensure comparability of datasets only the signals that are present in both Gurdasani et al. 2019 and AWI-Gen study will be included. If an independent signal shows a *p-value* that is less than (0.05/Number of independent lead signals) in an African cohort, it will be considered to be transferable to that cohort. This will be done for several independent target and base datasets (**Figure 1**). The level of transferability or significance from the multi-ancestry GLGC and African- ancestry specific GLGC to the combined AWI-GEN & Gurdasani meta-analyses, will be performed to assess which of the two GLGC datasets has higher transferability to the African populations. This will be done for each of the three lipid traits: HDL, LDL & TG. This will be followed by comparing the trends in transferability from the two GLGCs to the three different African regions the cohorts fall under (East, West & South), also for each of the three lipid traits to investigate intra-continental variation in transferability.

The Secondary Objective: Assessing heterogeneity of effect size and risk allele frequencies across African cohorts

Use the command line interface on the cluster, along with custom scripts/codes to extract the alleles, risk allele frequencies and effect size columns (including any other relevant data). Analyse and interpret the data in the form of various tables and graphs to compare the effect size magnitude, direction of effect and risk allele frequencies of the three lipid traits between the seven African cohorts.

4 Data Analysis

As this study will be comparing pre-calculated summary statistics, we have not performed any power estimation. Significance of allele frequency differences will be assessed using Chi-squared and Fischer tests implemented in PLINK 2.0. Data visualization will be performed using RStudio (v. 2023.12.1+402).

5 Ethics

This research project is a retrospective study using summary-level data so there are no major ethical considerations. The data was collected previously under earlier ethics

clearance (Clearance Certificate Number: M2210108) and no new data will be gathered. The study does not involve interacting with human or animal participants or the collection of new samples.

In accordance with the Protection of Personal Information Act (POPIA), the data used in this study contains no identifying information, such as names, dates of birth, addresses, identification, medical or hospital numbers. In addition, a gatekeeper permission letter has been obtained, granting permission for the use of the data in this research project.

A provisional sub-study ethics application form has been submitted to the Wits Medical Human Research Ethics Committee (HREC). Following the approval of the research protocol by the Post Graduate Assessors Committee, it will be submitted to the HREC who will then sit for ethics approval on the 9th May 2024. Once ethics approval has been obtained, the research project will begin.

6 Timing

Activity	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Literature Review		04/03									
Ethics Sub-study form submission		15/03									
Preparing Protocol											
Protocol submission			12/04								
Protocol presentation			06/04								
Ethics Approval Meeting				09/05							
Gathering summary data											
Data analysis & Result interpretation											
Report write-up											
Consolidation & conclusion											
Research Report Submission										30/11	

Figure 3: *GANTT reflecting the expected time frame for this research project.*

7 Funding

This is a retrospective study using pre-generated data, therefore does not require any funding. The data generation for the AWI-Gen study was funded by a five million USD National Institutes for Health (NIH) grant (U54HG006938).

8 Limitations

Potential limitations to the study are firstly the unequal sample size of the cohorts which might result in intrinsic differences in statistical power. Secondly, the transformations applied to lipid phenotypes in the GLGC study may differ from the transformations applied in the AWI-Gen and Gurdasani studies. However, we do not anticipate these limitations to substantially impact the transferability estimates.

Another limitation of this study is that only independent lead SNPs will be investigated for assessing transferability. Ideally, a more comprehensive investigation where instead of a single independent SNP, a credible SNP set per locus would have been more informative for transferability assessment, however, this was beyond the scope given the time period of the project.

9 References

Bush, W.S. and Moore, J.H. (2012) 'Chapter 11: Genome-Wide Association Studies', *PLoS Computational Biology*, 8(12), p. e1002822. Available at: https://doi.org/10.1371/journal.pcbi.1002822.

Choudhury, A. *et al.* (2022) 'Meta-analysis of sub-Saharan African studies provides insights into genetic architecture of lipid traits', *Nature Communications*, 13(1), p. 2578. Available at: https://doi.org/10.1038/s41467-022-30098-w.

Gaziano, T.A. *et al.* (2010) 'Growing Epidemic of Coronary Heart Disease in Low- and Middle-Income Countries', *Current problems in cardiology*, 35(2), pp. 72–115. Available at: https://doi.org/10.1016/j.cpcardiol.2009.10.002.

Global Lipids Genetics Consortium (2013) 'Discovery and refinement of loci associated with lipid levels', *Nature Genetics*, 45(11), pp. 1274–1283. Available at: https://doi.org/10.1038/ng.2797.

Graham, S.E. *et al.* (2021) 'The power of genetic diversity in genome-wide association studies of lipids', *Nature*, 600(7890), pp. 675–679. Available at: https://doi.org/10.1038/s41586-021-04064-3.

Gurdasani, D. *et al.* (2019) 'Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa', *Cell*, 179(4), pp. 984-1002.e36. Available at: https://doi.org/10.1016/j.cell.2019.10.004.

GWAS Catalog (no date). Available at: https://www.ebi.ac.uk/gwas/home (Accessed: 12 April 2024).

Kamiza, A.B. *et al.* (2022) 'Transferability of genetic risk scores in African populations', *Nature Medicine*, 28(6), pp. 1163–1166. Available at: https://doi.org/10.1038/s41591-022-01835-x.

Kuchenbaecker, K. *et al.* (2019) 'The transferability of lipid loci across African, Asian and European cohorts', *Nature Communications*, 10(1), p. 4330. Available at: https://doi.org/10.1038/s41467-019-12026-7.

Marees, A.T. *et al.* (2018) 'A tutorial on conducting genome-wide association studies: Quality control and statistical analysis', *International Journal of Methods in Psychiatric Research*, 27(2), p. e1608. Available at: https://doi.org/10.1002/mpr.1608.

Purcell, S. *et al.* (2007) 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *American Journal of Human Genetics*, 81(3), pp. 559–575.

Ramsay, M. *et al.* (2016) 'H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries', *Global Health, Epidemiology and Genomics*, 1, p. e20. Available at: https://doi.org/10.1017/gheg.2016.17.

Uffelmann, E. *et al.* (2021) 'Genome-wide association studies', *Nature Reviews Methods Primers*, 1(1), pp. 1–21. Available at: https://doi.org/10.1038/s43586-021-00056-9.

Wojcik, G.L. *et al.* (2019) 'Genetic analyses of diverse populations improves discovery for complex traits', *Nature*, 570(7762), pp. 514–518. Available at: https://doi.org/10.1038/s41586-019-1310-4.

Appendix D: Plagiarism Declaration and Turnitin Report

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Michaela Crosby (Student number: 23081287) am a student registered for the degree of MSc (Med) Genomic Medicine in the academic year 2024.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature:  Date: 30/11/2024

1

Final Turnitin Research Report Submission-1.pdf

ORIGINALITY REPORT

9%	6%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.nature.com Internet Source	2%
2	Ananyo Choudhury, Jean-Tristan Brandenburg, Tinashe Chikowore, Dhriti Sengupta et al. "Meta-analysis of sub-Saharan African studies provides insights into genetic architecture of lipid traits", Nature Communications, 2022 Publication	1%
3	www.biorxiv.org Internet Source	1%
4	Mats Nagel. "Changing perspectives: Towards detailed phenotyping in genetics", Thesis Commons, 2020 Publication	1%
5	repub.eur.nl Internet Source	<1%
6	www.cell.com Internet Source	<1%
7	worldwidescience.org Internet Source	<1%
8	www.frontiersin.org Internet Source	<1%
9	www.genomeweb.com Internet Source	<1%
10	Dhriti Sengupta, Ananyo Choudhury, Cesar Fortes-Lima, Shaun Aron et al. "Genetic-substructure and complex demographic history of South African Bantu speakers", Cold Spring Harbor Laboratory, 2020 Publication	<1%
11	Wu, Kuan-Han H.. "Integrating Electronic Health Records with Genetic Information to Advance Precision Medicine Approaches in Cardiovascular Diseases", University of Michigan, 2022 Publication	<1%
12	www.medrxiv.org Internet Source	<1%
13	wprod.sickkids.ca Internet Source	<1%
14	www.logomyway.com Internet Source	<1%
15	Carlberg, Daniel. "Genetic Association Studies in Allergic Rhinitis", Karolinska Institutet	<1%

(Sweden), 2021
Publication

16	Gibson, Greg. "A Primer of Human Genetics", Oxford University Press Publication	<1 %
17	Abram Kamiza, Sounkou Touré, Fend Zhou, Opeyemi Soremekun et al. "Discovery and fine-mapping of lipid traits loci in 125,000 individuals of African ancestry", Research Square Platform LLC, 2023 Publication	<1 %
18	N. R Wray. "Narrowing the Boundaries of the Genetic Architecture of Schizophrenia", Schizophrenia Bulletin, 12/08/2009 Publication	<1 %
19	Surina Singh, Ananyo Choudhury, Scott Hazelhurst, Nigel J. Crowther et al. "Genome-wide association study meta-analysis of blood pressure traits and hypertension in sub-Saharan African populations: an AWI-Gen study", Nature Communications, 2023 Publication	<1 %
20	Submitted to University of Witwatersrand Student Paper	<1 %
21	chs-nhlbi.org Internet Source	<1 %
22	genomemedicine.biomedcentral.com Internet Source	<1 %
23	repository.essex.ac.uk Internet Source	<1 %
24	Denis A. Baird, Jimmy Z. Liu, Jie Zheng, Solveig K. Sieberts et al. "Identifying drug targets for neurological and psychiatric disease via genetics and the brain transcriptome", PLOS Genetics, 2021 Publication	<1 %
25	Romain Icick, Alexey Shadrin, Børge Holen, Naz Karadag et al. "Identification of Novel Loci and Cross-Disorder Pleiotropy Through Multi-Ancestry Genome-Wide Analysis of Alcohol Use Disorder in Over One Million Individuals", Springer Science and Business Media LLC, 2023 Publication	<1 %

Exclude quotes Off
Exclude bibliography Off
Exclude matches Off

Appendix E: Author Guidelines for Human Molecular Genetics Journal

Instructions to Authors

Please read these instructions carefully and follow them strictly to ensure that the review and publication of your paper is as efficient and quick as possible. The Editors reserve the right to return manuscripts that are not in accordance with these instructions. Please note that all submissions should be accompanied by a manuscript submittal form, a copy of which can be found at the back of these instructions.

SCOPE

Human Molecular Genetics publishes papers of excellence, wide interest and broad significance in all aspects of human molecular genetics. Areas of interest include the molecular basis of human genetic disease, developmental genetics, chromosome structure and function, molecular aspects of cancer genetics, gene therapy, biochemical genetics and advances in gene mapping and understanding of genome organization. In general, mapping papers will be considered when they are of functional significance for the field of molecular human genetics. The molecular analysis of other species is also of interest where such studies are relevant to the understanding of the function of human genes.

REVIEW OF MANUSCRIPTS

Papers will normally be reviewed within three to four weeks of submission and will be published within six to ten weeks of acceptance. Due to space restrictions and to the volume of high quality papers submitted, the Editors reserve the right to return immediately those papers which are unlikely to be competitive for space in the journal and/or those which do not conform to the general editorial philosophy and standards of *Human Molecular Genetics*.

Type of Manuscript

The emphasis of the journal is on full research papers which may be of any length required for concise presentation and discussion of the data. Succinct and carefully prepared papers are favoured in terms of impact as well as readability. Several different categories of paper are published. While authors are invited to submit manuscripts in particular categories, the Editors reserve the right to determine in which category a paper is ultimately published, according to the criteria below.

Articles

Articles are definitive and comprehensive descriptions of major research findings of broad significance for readers of *Human Molecular Genetics*. The results described should be both novel and of wide interest. Articles may be of any length and may contain as many display items as appropriate for the subject matter.

Reports

Reports are descriptions of complete studies detailing results in areas of interest to readers of *Human Molecular Genetics*. Reports are typically less comprehensive than Articles and present data that are generally of narrower interest within the field of human molecular genetics. Reports may be of any length, although they are typically shorter than Articles.

Linkage Reports

Reports of novel localizations of genetic diseases can be presented as short Linkage Reports. Manuscripts should be presented in the same format as other Reports, but in as concise a manner as possible, including a description of the disease, family material, and the linkage analysis. Refinements of previously localized genetic disorders will generally not be considered and will be returned to the authors without review.

Mutation Reports

Analyses of mutations underlying human disease can be presented in a 1–2 page short report. The description of the disease, case and documentation of the mutation should be given. Evidence indicating that the mutation actually causes the disease should be included.

Commentaries and Mini-Reviews

Commentaries and Mini-Reviews are published by invitation only. However, proposals may be submitted in the form of a brief letter to the Editors at any time. The letter should state the topics and authors of the proposed review, and why the topic is of particular interest in the field.

PREPARATION OF MANUSCRIPTS

Three hard copies including all figures and tables should be submitted. Mini Review papers and full research papers will be automatically typeset from author-supplied word-processing disk. However, three hard copies should still be supplied.

Full research manuscripts (Articles and Reports) should be subdivided into the following sequence of sections: Title page, Abstract, Introduction, Results, Discussion, Materials and Methods, Acknowledgements, Abbreviations, References, Tables, Legends to Figures.

Mutation Reports do not have abstracts and should not be divided into headed sections; present the text in a single section. The Acknowledgements and References should be presented as in full research papers.

Mutation Reports are restricted in length to 1–2 printed pages. One printed page corresponds to approximately 1000 words without allowance for figures or tables.

General Format

The hard copies should be typed on A4 or American quarto (215mm×228mm). Dot matrix print or any print that is difficult to read is generally unacceptable. Please avoid the use of footnotes; use instead, and as sparingly as possible, parentheses within brackets. Underline only words or letters to appear in italics. Clearly identify unusual or handwritten symbols and Greek letters. Differentiate between the letter o and zero, and the letters l and 1 and the number 1. Mark the position of each figure and table in the margin.

Title page

The title should be short, specific and informative. The first name, initial(s), and surname of each author should be followed by his or her department, institution, city with postcode, and country. The fax and phone number of the corresponding author should also be provided. Any changes of address may be given in footnotes. The author(s) to whom reprints should be addressed should be indicated.

Abstract

The second page of every full research manuscript must contain the Abstract, which should not exceed 250 words. The Abstract should be comprehensible to readers before they have read the paper, and abbreviations and reference citations should be avoided. It is essential that the Abstract clearly states the biological importance of the work described in the paper.

Acknowledgements

These should be included at the end of the text and not in footnotes. Personal acknowledgements should precede those of institutions and agencies.

References

These should be numbered in order of appearance in the text and must be listed numerically. The citation of Journals, books and multi-author books and sequence data not published should conform with the following examples:

1. Huynh,T.V., Young,R.A. and Davis,R.W. (1988) In Glover,D.M. (ed.), *DNA Cloning—A Practical Approach*. IRL Press, Oxford, Vol.I, pp. 49–78.
2. Górecki,D.C., Monaco,A.P., Derry,J.M.J., Walker,A.P., Barnard,E.A. and Barnard,P.J. (1992) Expression of four alternative dystrophin transcripts in brain regions regulated by different promoters. *Hum. Mol. Genet.*, **1**, 505–510.
3. Smith,A. and White,B. (1991) *Nucleic Acids Res.* In Press.
4. Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor.
5. Muller,S.J. and Caradonna,S. (EMBL X52486).
6. Fairweather,N., Bell,C., Cochrane,S., Chelly,J., Wang,S., Mostacciuolo,M.L., Monaco,A.P. and Haites,N.E. (1994) Mutations in the connexin 32 gene in X-linked dominant Charcot–Marie–Tooth disease (CMTX1) *Hum. Mol. Genet.*, 3, 29–34.

References of the type Smith *et al.* (1989) should not be used.

Downloaded from https://academic.oup.com/hmg/issue/4/11 by guest on 19 November 2024

Personal communications (J.Jones, personal comm.) must be authorized in writing by those involved, and unpublished data should be cited in the text as (unpublished data). References to manuscripts submitted, but not yet accepted should be cited in the text as (B.Jones and L.Smith, in preparation) and should not be included in the list of references.

Illustrations

Wherever possible, figures should be submitted in their desired final size, to fit the width of a single column of text (88mm) or a double column of text (184mm). Any lettering should be approximately 2mm in height and should be in proportion to the overall dimensions of the figure. On the back of each figure, mark the number of the figure and the name of the first author, and also indicate the top margin.

Photographs should be of sufficient high quality with respect to detail, contrast and fineness of grain to withstand the inevitable loss of contrast and detail inherent in the printing process. When several prints of the same figure are provided, please indicate which print should be used for reproduction in the Journal.

There is a special charge for the inclusion of colour plates. Please contact the Publisher for further details.

Line drawings should have clear and sharp lines. No additional artwork, redrawing or typesetting will be done. Therefore all labelling should be on the original line drawing. Faint shading or stippling will be lost upon reproduction and should be avoided.

Figure legends should be included on the disk where appropriate and also on a separately numbered manuscript sheet. Define all symbols and abbreviations used in the figure. Figures and legends should be intelligible without reading the text of the manuscript.

Abbreviations

Try to restrict the use of abbreviations to SI symbols and those recommended by the IUPAC. Abbreviations should be defined and listed on a separate page with the footnotes. Standard units of measurement and chemical symbols of elements may be used without definition within the body of the paper. Acronyms formed from phrases are unacceptable.

Conventions

In general, the Journal follows the conventions of the CBE Style Manual (Council of Biology Editors, Bethesda, MD, 1983, 5th edn). Follow Chemical Abstracts and its indexes for chemical names. For guidance in the case of biochemical terminology follow the recommendations issued by the IUPAC-IUB Commission on Biochemical Nomenclature, as given Biochemical Nomenclature, published by the Biochemical Society, UK. Authors must follow the accepted nomenclature for human genes, as determined by the Nomenclature Committee of the Human Gene Mapping Workshops. See McAlpine,P. *et al.*, *Cytogenet. Cell Genet.*, 55, 5–76 (1990).

SUBMISSION TO THE EDITORS

Three hard copies of the manuscript including all figures and tables, disk if appropriate, and a completed manuscript submittal form should be submitted as follows:

From the Americas, the Pacific and Australia to:
The Editor Human Molecular Genetics, Department of Genetics, BRB 731, Case Western Reserve University, 2109 Adelbert Rd., Cleveland, Ohio 44106-4955, USA

Tel: (216) 368-0199 Fax: (216) 368-3030
E-mail: HMGJournal@po.CWRU.edu

From the Rest of the World to:
The Editor, Human Molecular Genetics, Molecular Genetics Group, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK

Tel: (+1865) 222397 Fax: (+1865) 65493

If a customs declaration is required, it is essential that the contents be listed as OF NO COMMERCIAL VALUE. Bills for Customs or VAT will be returned to the author.

If the manuscript is rejected, normally only the original figures are returned to the author. Submission of a paper implies that it reports unpublished work and that it is not under consideration for publication elsewhere. If previously published tables, illustrations or more than 200 words of text are to be included, then the copyright holder's permission must be obtained. Copies of any such permission letters should be included with the paper.

DISK SUBMISSION

To assist rapid publication, all papers should be submitted on disk. The computer format and word processor used must be clearly marked on the disk and on the manuscript submittal form. It is important to note that material submitted to the Editors in disk form must also be accompanied by three hard copies as above. See opposite for notes on disk preparation.

ACCESSION NUMBERS

Authors should be aware that all new sequence information, including that which extends a previously determined sequence already present in the database (and which already has an accession number) must be submitted to the Data Library for a new accession number. This number must be provided before acceptance of the manuscript.

REPRINTS

The Journal will provide 30 free reprints of all articles to authors. Reprint order forms are sent out with notification of acceptance and must be returned to Oxford University Press, Walton Street, Oxford OX2 6DP, UK if additional reprints are required. Late orders submitted after the Journal has been printed are subject to increased prices.

POLICY

It is understood that by publishing a paper in *Human Molecular Genetics*, the authors agree to make available to colleagues in academic (non-profit) research any of the cells, nucleic acids, antibodies etc. that are not available from commercial suppliers and are required to substantiate the scientific conclusions of the paper.

REVISION OF MANUSCRIPTS

When a manuscript is returned to the authors for revision prior to final acceptance, the revised version must be submitted within 1 month of the author's receipt of the referees' reports. Revised manuscripts returned after 1 month will be considered as new submissions subject to re-review.

PUBLICATION

Manuscripts will usually be published within two months of acceptance.

PAGE CHARGES

Page charges are NOT levied.

PROOFS

Disc edited proofs will be faxed to authors before typesetting translation. In the interest of speed, corrections should be returned within 24 hours. It is recommended that authors should spell-check their files before submission.

Downloaded from https://academic.oup.com/hmg/issue/4/11 by guest on 19 November 2024