Understanding the Genetic Basis of extreme levels of High and Low LDL-Cholesterol in African Populations

Natalie Smyth

Supervisors:
Prof. Michèle Ramsay
Prof. Frederick Raal
Dr. Dhriti Sengupta

A Thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

November 2025

Declaration

I, Natalie Smyth, declare that this Thesis is my own, unaided work, unless otherwise specified in the text. This thesis is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg, South Africa. It has not been submitted before for any degree or examination at this or any other university.

Natalie Smyth

November 2025

Dedication

To those who have supported me throughout this journey, especially my husband, Alex and my parents, Paul and Jenny. This would not have been possible without your encouragement, patience and guidance.

Presentations arising from this thesis

Smyth NA, Sengupta D, Raal F and Ramsay M. Genetic Contributions to high and low LDL-Cholesterol in African Populations. Poster presentation at the International Congress of Human Genetics conference in January 2023, Cape Town, South Africa.

Smyth NA, Sengupta D, Raal F and Ramsay M. Understanding the Genetic Contribution to high and low LDL-Cholesterol in African Populations. Oral presentation at the Southern African Society for Human Genetics Biennial Conference, 27-29 October 2024, Sun City, South Africa

Abstract

Cardiovascular diseases (CVDs) are a growing public health concern across Africa, with dyslipidaemia, particularly elevated low-density lipoprotein cholesterol (LDL-C), being a key modifiable risk factor. However, the genetics of lipid metabolism in African populations remain poorly understood due to limited representation in genomic studies. This thesis investigates both monogenic and polygenic contributions to extreme LDL-C levels in Sub-Saharan African populations, leveraging the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen) cohort.

To explore monogenic dyslipidaemia, we performed targeted next-generation sequencing (NGS) on 191 individuals with extreme high and low LDL-C levels from West, East, and South Africa, focusing on 16 genes implicated in LDL-C regulation, including *LDLR*, *APOB*, and *PCSK9*. This analysis identified 356 unique variants, including known pathogenic mutations and novel variants predicted to be functionally disruptive. While 28 variants were previously associated with phenotypic outcomes in ClinVar or HGMD, most were classified as variants of uncertain significance (VUS). Interpretation was complicated by the presence of multiple variants per individual, epistatic effects, and the absence of family-based cascade screening. Nevertheless, the findings highlight the complex genetic architecture associated with LDL-C levels in African populations and reinforce the need for functional studies and better population-specific variant classification.

In the polygenic arm of the study, we evaluated the predictive performance of four published polygenic scores (PGS) for LDL-C, developed from European, African-American, and multi-ancestry cohorts, across 10,603 participants from the AWI-Gen study. The variance explained (R^2) by each PGS ranged from 0.03 to 0.205, with the Graham and Zhang scores that were both developed in African–ancestry cohorts achieving the highest overall accuracy. Notably, regional differences in performance were observed, with PGS trained on African-American datasets performing better in South and East Africa than in West Africa, despite African-American having predominantly West African ancestral admixture. These results suggest that current PGS have limited and uneven predictive power in African populations, highlighting the need for ancestry-diverse discovery cohorts and region-specific calibration.

Together, these findings emphasize the importance of incorporating African genetic diversity into both monogenic and polygenic frameworks to improve CVD risk stratification and enable more equitable implementation of precision medicine on the continent.

Acknowledgements

To my supervisors, Prof. Michèle Ramsay, Dr. Dhriti Sengupta, and Prof. Frederick Raal – thank you for your unwavering guidance, mentorship, and patience. Your insight, encouragement, and belief in me have shaped this journey and brought me to this point. I am deeply grateful.

To my family and friends, – your support, especially over the past few months, has meant everything. Your belief in me, even when I struggled to believe in myself, carried me through.

To my colleagues at the SBIMB – thank you for the camaraderie and encouragement throughout. A special thank you to Jocelyne Gayenga and Busisiwe Mthembu for your invaluable help in the Biobank when I needed to focus on my PhD. To Dr. Carl Chen, I'm so grateful for your thoughtful advice and readiness to help. And to my fellow students – your kindness, shared frustrations, and solidarity have made this path less lonely.

Finally, to the research participants from the AWI-Gen study – thank you. Your willingness to share your time and data made this research possible, and I hold that privilege with deep respect and appreciation.

Table of Contents

List of Tables

List of Figures

List of Supplementary Tables

Chapter 2

List of Supplementary Figures

Chapter 3

List of Abbreviations

AADM: Africa America Diabetes Mellitus

ABCG5/8: Adenosine triphosphate Binding Cassette transporter

ACVD: Atherosclerotic Cardiovascular Disease

ANGPTL3: Angiopoietin-related protein 3

APOB: Apolipoprotein B

APOE: Apolipoprotein E

AWI-Gen: Africa Wits-INDEPTH Partnership for Genomic studies

BRDG1: BCR downstream signaling-1

CAD: Coronary Artery Disease

CHD: Coronary Heart Disease

CNV: Copy Number Variant

CVD: Cardiovascular Disease

DDD: Defined Daily Doses

FH: Familial Hypercholesterolaemia

FHBL: Familial Hypobetalipoproteinaemia

GATB: Glutamyl- TRNA Amidotransferase Subunit B

GLGC: Global Lipid Genetics Consortium

GWAS: Genome-Wide Association Study

H3Africa: Human, Heredity and Health in Africa

HeFH: Heterozygous Familial Hypercholesterolaemia

HDL-C: High-Density Lipoprotein Cholesterol

HMG-CoA: 3-hydroxy-3-methylglutaryl coenzyme A

HoFH: Homozygous Familial Hypercholesterolaemia

LDL-C: Low-Density Lipoprotein Cholesterol

LDL: Low-Density Lipoprotein

LDLR: Low Density Lipoprotein Receptor

LDLRAP1: LDLR adaptor protein 1

LIPA: Lysosomal acid lipase

LOVD: Leiden Open Variation Database

MAF: Minor Allele Frequency

MLPA: Multiplex Ligation-dependent Probe Amplification

MTTP: Microsomal triglyceride transfer protein

MYLIP: Myosin regulatory light chain interacting protein

NCD: Non-Communicable Disease

NGS: Next Generation Sequencing

NPC1L1: Niemann–Pick C1-like 1

PAGE: Population Architecture using Genomics and Epidemiology

PCSK9: Proprotein Convertase Subtilisin/Kexin Type 9

PGS: Polygenic Score

PGx: Pharmacogenomics

RMSE: Root Mean Square Error

SAMS: Statin Associated Muscle Symptoms

SAR1B: Secretion Associated ras related GTPase 1B

SI: Statin Intolerance

SNP: Single Nucleotide Polymorphism

SORT1: Sortilin

STAP1: signal transducing adaptor family member 1

VEP: Variant Effect Predictor

VLDL: Very Low-Density Lipoprotein

VUS: Variant of Uncertain Significance

WES: Whole Exome Sequencing

WGS: Whole Genome Sequencing

WHO: World Health Organization

Preface

This thesis is presented in the divided block format and includes drafts for two manuscripts that will be submitted for publication.

The thesis is divided into four chapters:

1. Introduction – this chapter describes the relevant literature, the study rationale, aims and objectives.
2. Manuscript 1: Targeted gene-panel sequencing reveals known and novel variants in Low-Density Lipoprotein Cholesterol associated genes in continental African populations – this chapter describes the results of sequencing individuals with high and low LDL-C levels.
3. Manuscript 2: Evaluation of Polygenic Scores for LDL-C in Sub-Saharan African Populations – this chapter describes the results of assessing the transferability of published polygenic scores for LDL-C prediction in the AWI-Gen cohort.
4. Discussion - this chapter summarizes the key findings of the study, research themes emerging from the study, limitations and potential future studies. The chapter ends with an overall conclusion.

Chapter 1: Introduction

Despite the global burden of cardiovascular disease (CVD), African populations remain understudied in terms of the genetic contributions to dyslipidaemia, a key CVD risk factor. There is a paucity of data as most genetic studies have focused on European-ancestry populations. This limits the generalizability of findings to diverse African populations. This dissertation contributes to addressing that gap by investigating both monogenic and polygenic contributors to extreme LDL-Cholesterol (LDL-C) levels in continental African populations and thus offering insights to genomic research for non-communicable diseases (NCDs) in Africa.

1.1. Cardiovascular Disease in Africa

The prevalence of NCDs is increasing worldwide, and they were responsible for more than 41 million deaths in 2021 (*WHO*, 2024). In Africa, it is expected that by the year 2030, NCDs will be the leading cause of death. CVD in particular, is rapidly increasing throughout the continent and is the most common NCD globally (Mayosi *et al.*, 2009). In Sub-Saharan Africa, cases of CVD have increased by 131.7% from 1990 to 2019 (Alhuneafat *et al.*, 2024). CVDs account for 38.3% of NCD deaths in Africa (Gouda *et al.*, 2019) and present a significant health and socioeconomic challenge. Healthcare systems across the continent, particularly in Sub-Saharan Africa, remain overwhelmed by the ongoing burden of infectious diseases and as a result, these systems are ill-equipped to manage the growing CVD epidemic (Peck *et al.*, 2014; 'Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.', 2020). To effectively address this issue, high-quality research is essential to develop evidence-based strategies aimed at reducing the impact of CVD in Africa.

CVD is defined as a collection of disorders that affect the blood vessels and heart that include coronary heart disease (CHD), rheumatic heart disease, coronary artery disease (CAD) and congenital heart disease. At least 75% of the deaths due to CVD arise in low- and middle-income regions such as Africa and is attributed to lack of early diagnosis, and an inadequate access to primary health care (WHO, 2017). Common risk factors include hypercholesterolaemia (elevated lipid levels), diabetes mellitus, obesity, hypertension and smoking (Mensah, 2013). A review article by Minja *et al*. (2022) discusses of the burden of metabolic risk factors to cardiovascular disease in four regions in Africa, including Sub-Saharan Africa. It illustrates that the highest disease burden is in Southern Africa and

hypercholesterolaemia, albeit not the largest contributor, is a significant risk factor (Minja *et al.*, 2022).

1.2. Dyslipidaemia

Dyslipidaemia is defined as abnormal (either elevated or decreased) lipid levels (Braun *et al.*, 2016). Lipids include cholesterol, triglycerides and phospholipids and are utilized in hormone production and cell membrane formation. Because lipids are insoluble in blood, they are transported as lipoprotein particles which are complexes composed of lipids and proteins. There are five major classes of lipoproteins, classified based on their density and function: chylomicrons, very low-density lipoproteins (VLDL), intermediate-density lipoproteins (IDL), low-density lipoproteins (LDL), and high-density lipoproteins (HDL) (Lin *et al.*, 2010). Elevated LDL-C levels have been shown to be a primary causal factor of atherosclerotic cardiovascular disease (ASCVD) ('Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report.', 2002),(Kalwick and Roth, 2025). Guidelines in terms of what defines an abnormal lipid level differ between countries. According to the WHO, standard healthy LDL-C levels are between 0.5 and 3.0 mmol/L.

LDL particles transport cholesterol from the liver to peripheral tissues where it further enables multiple cellular processes. The process which ultimately leads to the development of atherosclerosis starts when LDL penetrates the artery intima and undergoes oxidative alterations. These modified LDL particles cause an inflammatory response that attracts circulating monocytes, which then differentiate into macrophages at the site of damage. These macrophages engulf the modified LDL, transforming into foam cells and contributing to the accumulation of lipid-rich plaques, a key process in the development of atherosclerotic lesions (Bouhairie and Goldberg, 2015). There is strong and congruous evidence from genetic studies, prospective epidemiological cohorts, Mendelian randomization studies, and randomized intervention trials that have established that elevated LDL-C levels are not merely a biomarker of increased risk, but are, in fact, a causal factor in the pathogenesis of ASCVD (Ference *et al.*, 2017).

1.3. Genetic Architecture of Dyslipidaemia

The levels of LDL-C are controlled by a complex system of genetic and non-genetic factors, classifying dyslipidaemia as a multifactorial disorder. Elevated LDL-C levels are exacerbated by consuming processed foods with high fat content and low levels of physical activity, in the

context of predisposing genetic factors. The heritability of LDL-C is estimated to be between 40 and 60% (Weiss *et al.*, 2006). Rare genetic variants with large phenotypic effects cause monogenic forms of dyslipidaemia, but most cases are the result of many common genetic variants, each with small to moderate phenotypic effects, that cumulatively underlie polygenic forms of the disorder (Dron and Hegele, 2018).

1.3.1. Monogenic Dyslipidaemia

Monogenic disorders associated with elevated LDL-C levels make identification of a causative pathogenic variant less complicated since there is usually a single causal variant, compared to unravelling the complex genetic architecture of multifactorial disorders with the contribution of many genetic variants and other factors that each contribute a small amount to the phenotype (Antonarakis and Beckmann, 2006). Monogenic dyslipidaemias are characterized by the presence of a variant in a major gene involved in lipid metabolism in an individual, resulting in a large phenotypic effect, namely a change in lipid levels (Dron and Hegele, 2018). There is locus heterogeneity, with different individuals having mutations in different genes with different effects on their lipid profiles. These disorders can present as autosomal dominant, co-dominant or recessive traits. There are approximately 27 monogenic disorders of lipid metabolism caused by rare variants found in 25 different genes affecting lipid metabolism (Rahalkar and Hegele, 2008). Familial Hypercholesterolaemia (FH) in particular, is a genetic disorder marked by elevated LDL-C levels, which promote early-onset atherosclerotic plaque formation in the coronary arteries. This substantially elevates the risk of early-onset cardiovascular events, including angina and myocardial infarction. Majority of pathogenic variants linked to FH are identified in three major genes, namely Low-Density Lipoprotein Receptor (*LDLR*), Apolipoprotein B (*APOB*) and/or Proprotein subtilisin/kexin Type 9 (*PCSK9*) (Ison, Clarke and Knowles, 2025).

1.3.2. Multifactorial Dyslipidaemia

Comprehensive genetic studies of individuals with either high or low LDL-C levels have indicated a complex multifactorial aetiology for this trait (Dron and Hegele, 2019). Current research indicates that in clinically defined FH cohorts, a monogenic mutation can be identified in only about 40–60% of patients (Mariano *et al.*, 2020). The combined effect of many common small-effect variants namely single nucleotide polymorphisms (SNPs), also influence lipid traits. Recent discoveries over the past decade have shown that these small-effect SNPs with high population frequencies cumulatively increase or decrease lipid levels under a polygenic

framework (Dron and Hegele, 2018). Non-Mendelian cases of high LDL-C levels usually point to a polygenic inheritance pattern. The condition is influenced by both genetic factors and environmental factors such as a diet high in saturated fats, obesity, and low physical activity. Additionally, gene-environment interactions, epistasis and epigenetic factors may play a role in lipid homeostasis.

To understand the genetic architecture of the multifactorial basis of high LDL-C, genome-wide association studies (GWASs) and meta-analyses of large datasets have revealed many variants associated with higher LDL-C concentrations. The polygenic nature of varying lipid levels makes it possible to calculate polygenic scores (PGS) (i.e. the genetic contribution to a trait). A PGS is a calculated variable for an individual which summarizes the anticipated effect of the accumulation of trait-associated SNPs from GWASs. GWASs have permitted researchers to discover common variants throughout the genome, including in both coding and non-coding regions, that are significantly associated with phenotypic effects. The Global Lipid Genetics Consortium (GLGC) conducted the largest GWAS meta-analysis in the field of lipid research, identifying over 150 common SNP loci with moderate effects on lipid traits (Willer *et al.*, 2013).

The number and choice of SNPs in a PGS calculation and the assessment of resulting accuracy have evolved over the years. In 2015, a comparatively small number of SNPs was shown to have reasonable predictive power (Futema *et al.*, 2015). Conscious efforts were made to reduce the number of SNPs in the PGS. The initial PGS included 33 SNPs; however, sequential removal of variants with smaller effect sizes revealed that a weighted score based on just 6 SNPs performed comparably to a 12-SNP score in individuals of European ancestry. In mutation-negative individuals with elevated LDL-C levels, meta-analysis of the weighted 6-SNP score demonstrated consistently higher PGS values compared to healthy controls ($P < 2.2 \times 10^{-16}$) (Futema *et al.*, 2015). A PGS constructed with over 400 000 SNPs explained 13.8% of variation in LDL-C compared to the 12 SNP score which only explained 8.8% (Eyrich *et al.*, 2024). The scientists showed that there was a positive correlation between LDL-C levels and PGS and that individuals with a PGS in the top 1% has an almost equal risk of disease to those who carry a pathogenic monogenic variant (Eyrich *et al.*, 2024), (Khera *et al.*, 2018). This study, however, utilized a cohort from Denmark only, and thus transferability and reliability of the score in other populations can be questioned.

Majority of PGSs have been developed from studies including only European populations and these scores have been shown to perform poorly in other ethnic groups (Duncan *et al.*, 2019;

Martin *et al.*, 2019). Some of the contributing factors are dissimilar allele frequencies, and differences in linkage disequilibrium patterns and environment (Gomez, Hirbo and Tishkoff, 2014). In 2019, Martin *et al*. illustrated that prediction risk for 17 quantitative traits derived from European populations were not as accurate in 4 other populations within the UK Biobank, including South Asian individuals. Prediction accuracy was the lowest in East Asian and African individuals (Martin *et al.*, 2019).

Three key studies, by Choudhury *et al*., Graham *et al*., and Kamiza *et al*., have advanced understanding of PGSs for lipid traits in African populations. Choudhury *et al*. (2022) aimed to explore the genetic architecture of lipid levels across Sub-Saharan Africa through a large meta-analysis including ~24,000 continental participants. Their goal was to evaluate the performance of PGSs derived from African, European, and multi-ancestry GWAS. They assessed whether PGSs developed from the GLGC (European), Population Architecture using Genomics and Epidemiology (PAGE) (Multi-ancestry) and a Sub-Saharan African meta-analysis GWAS were able to accurately predict lipid levels, including LDL-C, in the AWI-Gen cohort. It was found that PGSs trained on African-ancestry data outperformed European-based scores in African cohorts, especially for LDL-C, reinforcing the importance of ancestry-matched discovery datasets (Choudhury *et al.*, 2022).

Graham *et al*. (2021) sought to assess how inclusion of diverse ancestries in GWAS discovery improves PGS performance across populations. They developed LDL-C PGSs from three different sources—European-only, admixed African (admAFR), and trans-ancestry GWAS—and evaluated them in multiple cohorts, including AWI-Gen and Africa America Diabetes Mellitus (AADM) cohorts. Their results showed that adding non-European cohorts to a lipid GWAS improved discovery and fine-mapping. PGSs derived from admixed African populations were shown to outperform European-based scores in African samples and that trans-ancestry scores achieved the best overall prediction accuracy. This underscores the value of multi-ancestry GWAS in developing broadly transferable risk models. The authors went further to assess the developed PGSs in various African regions.

Across the different geographic regions in Africa, the variance of LDL-C explained by the PGS model was higher in South Africa highlighting that even within Africa, regional differences in prediction accuracy of PGSs is present. This study however, utilized African American individuals in their discovery cohort and the score performed marginally better in African American populations than continental Africans. When assessing regional differences, as previously mentioned, the score performs best in the South African region of the AWI-Gen

cohort, however the adjusted R^2 value is similar to that of the West African in the AADM cohort (Graham *et al.*, 2021). This may indicate regional differences however more in-depth research is required to determine how large the differences are when using continental African populations from different regions.

Kamiza *et al*. (2022) focused specifically on evaluating the transferability of existing PGSs for lipid traits across different African populations. Using cohorts from Uganda and South Africa (Zulu population), they compared the performance of PGSs derived from African American, European, and multi-ethnic datasets. Their findings emphasised the heterogeneity within Africa, as the African American-derived PGS explained 8.14% of LDL-C variance in the Zulu population but only 0.026% in the Ugandan cohort. This demonstrated that even PGSs developed from African-ancestry populations abroad do not perform consistently across diverse African groups (Kamiza *et al.*, 2022).

Together, these studies highlight several important principles: ancestry-aligned PGS offer superior predictive power, multi-ancestry GWAS improve transferability, and within-continent genetic diversity must be accounted for. While all three studies aimed to improve LDL-C prediction in African individuals, their differing methodologies—ranging from discovery GWAS to cross-cohort PGS testing—offer converging evidence that local data and region-specific strategies are essential for equitable implementation of genomic risk prediction in Africa.

1.3.3. Treatment for Hypercholesterolaemia

Patients with high LDL-C have a greater risk of CVD, particularly those diagnosed with FH, as exposure to higher cholesterol levels from a younger age leads to earlier development of CVDs. Thus diagnosis, prevention and treatment are of utmost importance. It is widely accepted that statins are the first line treatment for individuals with high cholesterol (Clebak and Dambro, 2020). Statins works by inhibiting HMG-CoA, or 3-hydroxy-3-methylglutaryl coenzyme A, the rate-limiting enzyme in cholesterol metabolism, and as a consequence, increased LDLR activity (Rosenson, 2021). A review published in 2020 highlights that prolonged use of statins may lead to muscle pain, myopathy, muscle weakness, musculoskeletal injury, liver injury, and increased risk of diabetes (Clebak and Dambro, 2020). Response to statin treatment in patients is vastly variable when decreasing LDL-C and increasing HDL-C levels (Davidson and Toth, 2004). Genetic factors have been shown to contribute to this variability (Thompson *et al.*, 2009; Ruiz-Iruela *et al.*, 2020). Variants that

cause non-conservative amino acid substitutions in *PCSK9* are associated with a significant decrease in LDL-C levels in the plasma, and interestingly an increased response to statin therapy (Berge, Ose and Leren, 2006).

PCSK9 loss-of-function variants have shown to lower LDL-C levels by reducing the rate of LDLR breakdown and increasing LDLR recycling, thus making the gene an ideal target for drugs. Drugs such as evolocumab and alirocumab are monoclonal antibodies that inhibit PCSK9 enzyme activity. This enhances the clearance of LDL-C from the bloodstream into the liver, resulting in significant reductions in circulating LDL levels (Farhan *et al.*, 2025). There is also a study that shows that the use of PCSK9 inhibitors significantly reduces the risk of stroke and other major adverse cardiovascular events (Moustafa *et al.*, 2024; Farhan *et al.*, 2025).

1.3.4. Genes Implicated in LDL-C Metabolism and FH

LDL-C metabolism and FH involve a complex network of genes. While pathogenic variants in three primary genes, *LDLR*, *APOB*, and *PCSK9*, have been strongly associated with FH, a considerable proportion of clinically diagnosed cases lack identifiable mutations in these loci. This suggests the involvement of additional genes contributing to LDL-C regulation and the broader dyslipidaemia phenotype. Expanding the genetic scope beyond these well-established genes may help explain the missing heritability in FH and uncover novel contributors to lipid disorders.

LDLR is a gene with 18 exons and found on chromosome 19 (Francke, Brown and Goldstein, 1984). *LDLR* mutations account for the vast majority of FH in most populations (Leigh *et al.*, 2017). The function of the LDL receptor is to regulate LDL-C levels by binding and transporting LDL-C into the cells, thereby reducing levels in the blood stream. LDL-C is one of the two types of lipoproteins found in the blood stream that carry cholesterol, the other being HDL-cholesterol which is thought to be involved in reverse cholesterol transport. The impact of various *LDLR* variants on remaining LDLR activity differs significantly, resulting in a wide range of LDL cholesterol levels and varying risks for CAD among individuals with FH (Huijgen, Hutten, *et al.*, 2012; Huijgen, Kindt, *et al.*, 2012; Berberich and Hegele, 2019). Typically, this variation in LDLR function is categorized into two groups— "deficient" receptors (with residual LDLR activity <2%) and "defective" receptors (with residual LDLR activity ranging from 2–25%) (Watts *et al.*, 2014; Santos *et al.*, 2016). Pathogenic variants have been detected in all 5 functional domains of *LDLR* including the ligand or APOB-binding

domain, epidermal growth factor-like domain, *O*-linked sugar domain, transmembrane domain and the anchoring cytoplasmic tail domain (Goldstein and Brown, 2009) and can affect receptor-mediated endocytosis. Individuals who are heterozygous for a pathogenic variant have been shown to have a plasma LDL-C concentration on average double that of individuals who do not carry a pathogenic variant and thus have a significantly increased risk of CAD. Individuals who are homozygous tend to develop premature atherosclerosis which may cause death at an early age (Jeon and Blacklow, 2005). There are 5 classes of FH variants in *LDLR*, namely (1) null mutations whereby no protein is made, (2) transport-defective mutations which are either entirely or partially defective in reaching the surface of the cell, (3) binding defective mutations, (4) internalisation-defective mutations where the protein does not cluster in the pits at the surface of the cell, and (5) recycling-defective mutations (Jeon and Blacklow, 2005).

The lifecycle of LDLR (Figure 1) is key to understanding the interplay between the three main genes as well as a few auxiliary genes. When the intracellular cholesterol levels are low, *LDLR* is upregulated and the protein is synthesized within the endoplasmic reticulum (Wijers, Kuivenhoven and van de Sluis, 2015). Afterwards, it undergoes glycosylation within the Golgi apparatus and the mature LDL receptors travel to the hepatocyte cell surface. The circulating LDL particles within the blood stream bind to the apolipoprotein B-binding domain of the receptor. This complex is transported into the cell through vesicles which join with early endosomes (Wijers, Kuivenhoven and van de Sluis, 2015). When the LDLR-LDL complex is exposed to a lower pH in the endosome, the receptor releases the LDL particle, and it is recycled. The receptor can be recycled more than 100 times. The LDL particle is then degraded via the lysosomes and cholesterol is released into the cell (Wijers, Kuivenhoven and van de Sluis, 2015). PCSK9 or E3 ubiquitin-protein ligase MYLIP may degrade the LDLR in response to the increased intracellular cholesterol level. PCSK9 orchestrates this process both within and outside of the cell by shifting the pH. This causes the LDLR to remain in the endosome and it is thus eventually degraded along with the LDL particle (Burke *et al.*, 2017). Pathogenic variants found in any of the genes described can lead to either high or low LDL-C levels.

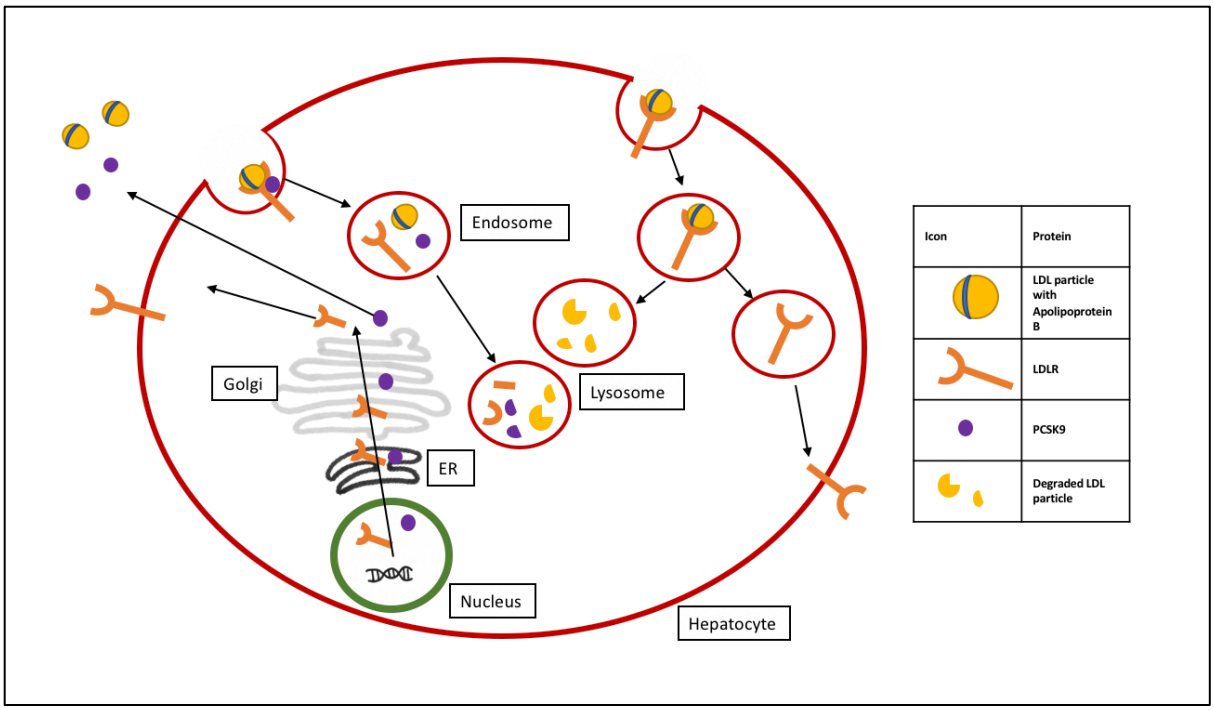


Figure 1: Lifecyle of LDLR. LDLR is responsible for transport of LDL particles into the cell for degradation. PCSK9 is responsible for the breakdown of LDLR.

The *APOB* gene is situated on chromosome 2, has 29 exons and generates 2 types of the protein, namely apoB-48 and apoB-100. The protein, apoB-100, is produced in the liver and is an important component of VLDL, ILDL and LDL. It is an amphipathic glycoprotein and is a key player in lipid metabolism as it is the ligand for the *LDLR* (Whitfield *et al.*, 2004). Variants located in the *APOB* receptor binding region (exons 26 and 29) lead to flawed binding and an accumulation of LDL-C in the bloodstream. This disorder, familial defective apoB, is a FH phenotype which accounts for approximately 5-10% of FH cases (Borén *et al.*, 2001). In the heterozygous form, this phenotype cannot be clinically differentiated from FH, however it has been shown to be less severe. Statin therapy has been shown to be effective to treat familial defective apoB as normal LDL receptors are present (Whitfield *et al.*, 2004). On the other hand, variants in *APOB* that lead to a truncated protein can cause Familial Hypobetalipoproteinaemia (FHBL) which is a rare autosomal codominant disorder. It is also characterized by low LDL-C levels. There are approximately 50 known variants that are associated with this disorder, most of which are found in exon 26 (Wakabayashi *et al.*, 2024). The prevalence of a truncated apoB in the general population is 1 in 3000, however the occurrence of homozygotes is estimated to be less than 1 in a million (Burnett, Hooper and Hegele, 1993). Heterozygous FHBL patients are typically asymptomatic and exhibit one third of the normal range of plasma levels of LDL-C and apoB. Some may experience mild liver dysfunction due to hepatic

steatosis, and 5–10% may go on to develop serious liver conditions such as cirrhosis or liver cancer (Tarugi and Lonardo, 1997; Lonardo *et al.*, 1998; Bonnefont-Rousselot *et al.*, 2009). Some cases have exhibited acanthocytosis, fat malabsorption, and mild vitamin deficiencies (Burnett, Hooper and Hegele, 1993; Wakabayashi *et al.*, 2024). Very few cases have presented with neurological issues which may be due to vitamin E and A deficiencies (Musialik *et al.*, 2020). The persistent reduction in serum lipids in heterozygous FHBL patients may offer protection against atherosclerotic cardiovascular diseases. In contrast, individuals with homozygous FHBL experience more severe symptoms, resembling abetalipoproteinemia due to mutations in the *MTTP* gene (Peloso *et al.*, 2019).

The *PCSK9* (chromosome 1) encodes a 692 amino acid protein and is synthesized in the kidney, liver and intestine (Horton, Cohen and Hobbs, 2009). The protein, PCSK9, starts with a signal sequence, followed by the pro-domain, the catalytic domain and lastly a c-terminal region which is rich in cysteine and histidine (Horton, Cohen and Hobbs, 2007). PCSK9 facilitates the breakdown of LDLR in the cells by targeting the receptor protein to the lysosome for destruction leading to less LDL-C uptake and thus an increased LDL-C plasma level (Horton, Cohen and Hobbs, 2007, 2009; Lakoski *et al.*, 2009). Variants in *PCSK9* can lead to either a gain-of-function or loss-of-function effect. The latter results in an accumulation of *LDLR* and thus a lowered LDL-C level in the bloodstream, whereas a gain-of-function variant will result in higher levels due to less *LDLR* activity (Horton, Cohen and Hobbs, 2009). The effects of loss-of-function *PCSK9* variants were determined in both in vitro and in vivo experiments and in 2005, Cohen et al. investigated possible variants in individuals with low plasma LDL-C levels (Cohen *et al.*, 2005). They discovered two inactivating variants, namely p.Y142X and p.C679X. Both variants were associated with an approximately 30-40% decrease in LDL-C plasma levels. A variant, p.Y142X, a nonsense mutation, is found in exon 3 and causes a truncation of the remainder of the protein. p.C679X, also a nonsense mutation, is found in exon 12, causes a truncation of the last 14 amino acids. Cohen and colleagues suggest that loss-of-function variants could increase the levels of LDLR in the liver cells and thus could potentially lower risk and protect against CAD (Cohen *et al.*, 2006). Interestingly, both variants were initially discovered in individuals of African descent (Cohen *et al.*, 2005). Less than 1% of FH cases are due to the gain-of-function variants in *PCSK9*. This group of variants acts by either increasing the activity of the PCSK9 protein or by presenting a new activity to the protein (Horton, Cohen and Hobbs, 2007). Studies in mice determined that mutant Pcsk9 led to high levels of the protein and a significant decrease in LDLR levels in the liver, however it was

shown that the levels of LDLR mRNA were not reduced leading to the conclusion that over expression of *PCSK9* lead to a reduction of LDLR through a post transcriptional mechanism (Maxwell and Breslow, 2004; Park, Moon and Horton, 2004; Horton, Cohen and Hobbs, 2009). Locations of variants do not appear to influence the functional consequence, for example, p.R46L and p.Y142X are both associated with lower levels of LDL-C and p.S127R is associated with higher levels of LDL-C, yet all 3 variants are found in the resulting prodomain of the protein (Horton, Cohen and Hobbs, 2007).

It is common to encounter individuals with autosomal dominant FH (based on a family history) who do not have mutations in the three previously discussed genes, suggesting that other genes may be involved.

LDLR adaptor protein 1 (encoded by the *LDLRAP1* gene), has been associated with LDL-C metabolism and an autosomal recessive disorder form of hypercholesterolaemia (ARH). Variants in this gene are rare with an estimated prevalence of 1 in 1 000 000 (Rogozik, Główczyńska and Grabowski, 2024; Yazıcı *et al.*, 2024). The LDLRAP1 protein is known to interact with the LDLR:LDL complex during transport into hepatocytes and thus a defective protein leads to a defect in the internalization process (Fellin *et al.*, 2015).

STAP1 (signal transducing adaptor family member 1) produces a docking protein which is also known as BRDG1 (BCR downstream signaling-1). STAP1 contains a domain with multiple tyrosine phosphorylation sites, which enable it to interact with membrane-associated proteins. This suggests that STAP1 plays a role in cholesterol metabolism through its interaction with these membrane proteins (Fouchier *et al.*, 2014; Lamiquiz-Moneo *et al.*, 2020). In 2014, it was shown by exome sequencing and linkage analysis in a single family in the Netherlands that *STAP1* variants are likely associated with autosomal dominant FH , however, a more recent study examining four rare pathogenic variants in *STAP1* concluded that this may not be the case (Fouchier *et al.*, 2014; Lamiquiz-Moneo *et al.*, 2020). The exact mechanism of action of the protein is still under debate however subsequent studies have detected disease-causing variants that may be associated with hypercholesterolaemia (Lamiquiz-Moneo *et al.*, 2021; Totoń-Żurańska *et al.*, 2023).

ApoE is an apolipoprotein that is known to be involved in lipid metabolism. It is expressed primarily in the liver and is involved in multiple pathways including but not limited to lipoprotein metabolism, signal transduction and angiogenesis (Marais, 2019; Khalil *et al.*, 2021). There are three major APOE isoforms (ε2, ε3, and ε4) that are determined by two SNPs,

rs429358 and rs7412, which alter the amino acid sequence of the APOE protein and affect its binding affinity for lipoproteins and receptors (Khalil *et al.*, 2021). APOE4 is known to be associated with higher APOB and LDL-C levels. ApoE2, on the other hand, is associated with lower APOB and LDL-C plasma levels (Rasmussen *et al.*, 2016). APOE4 is also associated with higher risk of Alzheimer's disease and APOE2 has been shown to be protective against Alzheimer's disease (Serrano-Pozo, Das and Hyman, 2021).

Angiopoietin-related protein 3, expressed by *ANGPTL3* and produced in the liver, falls within a family of proteins, namely angiopoietin-like proteins, which are vital regulators of the lipid metabolism (Kersten, 2017). Rare loss-of-function variants lead to lower concentrations of LDL-C (Ando *et al.*, 2003; Minicocci *et al.*, 2012). The disorder characterized by low lipid levels due to pathogenic variants in *ANGPTL3* is called familial hypobetalipoproteinaemia-2 (Kersten, 2017). It has been suggested that carriers of functional variants have a reduced risk of CAD (Kersten, 2017).

Two members of the human adenosine triphosphate (ATP) Binding Cassette (ABC) transporter family, ABCG5 and ABCG8, play an important role in cholesterol absorption. These ATP-dependent half-transporters join together to form a transport complex that regulates cholesterol absorption from the diet in the intestine and its subsequent excretion in the bile from the liver. Their role was uncovered through the study of sitosterolaemia, a rare genetic disorder marked by abnormal sterol levels in the blood, which results from variants in the genes that code for the ABCG5/G8 transport complex (Dean, Hamon and Chimini, 2001; Yu *et al.*, 2002; Berge, Ose and Leren, 2006; Viturro *et al.*, 2006). Individuals with sitosterolaemia usually present as homozygous or compound heterozygous variants in *ABCG5* and/or *8* and may present with xanthomas, hypercholesterolaemia and an increased risk of CHD which present similarly to FH (Yoo, 2016; Reeskamp *et al.*, 2020).

Niemann–Pick C1-like 1 (NPC1L1) has been shown to be important for cholesterol absorption in the intestines (Altmann *et al.*, 2004). Cohen et al. sequenced the *NPC1L1* gene and discovered that individuals with rare variants known to be associated with reduced cholesterol absorption had significantly lower mean LDL-C levels compared to individuals without the variants (Cohen *et al.*, 2006). *NPC1L1* is also the target of ezetimibe, which inhibits cholesterol absorption and genetic variations within the gene have shown to influence the response to ezetimibe in terms of reducing LDL-C (Garcia-Calvo *et al.*, 2005; Hegele *et al.*, 2005; Simon *et al.*, 2005; Wang, Williams and Hegele, 2005). A research team in the USA analysed multiple variants in *NPC1L1* and found that carriers of rare functional alleles presented with a 15%

greater LDL-C reduction following ezetimibe treatment compared to noncarriers (Simon *et al.*, 2005). Furthermore, Hegele and colleagues at the Robart's Institute in Canada showed that individuals with dyslipidaemia who carried an uncommon *NPC1L1* haplotype, including the minor alleles p.L272L, p.V1296V, and c.A25342C, had a more significant response to ezetimibe. These individuals showed a 36% LDL-C reduction, compared to a 18% decrease in those without the haplotype (Hegele *et al.*, 2005).

Table 1 shows additional genes that may be involved in the lipid metabolism and thus variants in these genes may affect LDL-C levels.

Table 1: Additional genes that may be involved in the LDL-C metabolism

Gene	Protein	Description and Pathology
LIPA	Lysosomal acid lipase	Hydrolyzes cholesteryl esters and triglycerides and thus produces free fatty acids and cholesterol in the cell. Variants are associated with rare lysosomal disorders, Wolman disease and cholesteryl ester storage disease (Li and Zhang, 2019).
SORT1	Sortilin	Transports different proteins to either the cell surface or lysosomes and endosomes within the cell (Conlon, 2019). Linked with LDL-C levels and cardiovascular disease (Coutinho *et al.*, 2013).
MYLIP	Myosin regulatory light chain interacting protein	Functions as a sterol-dependent regulator of cellular cholesterol uptake by promoting the ubiquitination and further degradation of the LDL receptor (LDLR) (Lindholm, Bornhauser and Korhonen, 2009). Associated with Hyperlipidaemia (Adi *et al.*, 2020).
MTTP	Microsomal triglyceride transfer protein	Traffics neutral lipids, triglycerides and cholesterol esters between vesicles (Wetterau and Zilversmit, 1984).

		Associated with abetalipoproteinemia and Familial hypobetalipoproteinaemia (FHBL) (Takahashi *et al.*, 2021; Wakabayashi *et al.*, 2024).
SAR1B	Secretion Associated ras related GTPase 1B	A component of the vesicular coat protein complex II, which forms a protective layer around vesicles that transport chylomicron (CM) contents through the secretory pathway. Mutations in Sar1b can disrupt the transport of chylomicrons between the endoplasmic reticulum and the Golgi apparatus, leading to impaired CM secretion and resulting in deficiencies of vitamins and essential fatty acids in affected individuals (Auclair *et al.*, 2023).
GATB	Glutamyl- TRNA Amidotransferase Subunit B	Novel association to LDL-C found in recent GWAS (Choudhury *et al.*, 2022).

1.3.5. Familial Hypercholesterolaemia (FH)

FH is one of the most common heritable diseases in the world (Dron and Hegele, 2018) and leads to the rapid progression of atherosclerosis, with severe outcomes such as myocardial infarction and death occurring at a young age particularly in those who are untreated or inadequately treated. Despite the availability of several effective cholesterol-lowering medications, a significant challenge in managing this condition is the lack of early detection and proper pharmacological intervention. While the most severe forms, like homozygous FH, typically present clear physical signs from childhood, milder forms of FH can remain undiagnosed until the first cardiovascular event. Therefore, early identification of these individuals is essential to minimize cholesterol exposure and reduce the risk of cardiovascular events (Raal, Hovingh and Catapano, 2018). Individuals with monogenic forms of FH are on average, three times more likely to develop ASCVD compared to those without the mutation, which is possibly due the consistent elevated cholesterol levels since birth (Do *et al.*, 2015; Ibrahim *et al.*, 2024).

The clinical diagnosis of FH in an individual is done by using three formal diagnostic criteria, namely the MEDPED Criteria, the Simon Broome Criteria, and the FH Dutch Lipid Clinic

Network Criteria. In the United Kingdom, the NICE guideline CG71 recommends diagnosing FH using the Simon Broome criteria, which classify cases as definite or possible FH based on LDL-C thresholds (≥4.9 mmol/L in adults or ≥4.0 mmol/L in children), family history of premature coronary heart disease, and/or the presence of physical signs of cholesterol deposition (NICE, 2023). In South Africa, the FIND FH program, which was launched in 2016, systematically identifies individuals with FH across the country using the Simon Broome criteria and genetically characterising FH-causing variants across different ethnic groups (Raal *et al*., 2020). In contrast, other countries apply Dutch Lipid Clinic Network (DLCN) or MEDPED criteria, which use point-based systems or population-specific cholesterol cut-offs, leading to variations in diagnostic thresholds and classification depending on regional reference ranges and healthcare practices. All three use principles to establish a diagnosis, namely, LDL-C level (>4.9mmol/L), history of CAD, family history, and physical symptoms such as xanthomas, xanthelasmas and corneal arcus (Cartier and Goldberg, 2016; Ison, Clarke and Knowles, 2025).

Molecular diagnosis is usually established when a pathogenic variant is identified in *LDLR*, *APOB* or *PCSK9*. Pathogenic variants detected in these three major genes lead to an autosomal dominant inheritance pattern. It is possible for an individual to have biallelic pathogenic variants, which can be either homozygous or compound heterozygous, and thus will have homozygous FH (HoFH). In these cases, the phenotype is more severe and occurs earlier in life (Ison, Clarke and Knowles, 2025).

1.3.5.1. Prevalence of FH

The worldwide prevalence of FH is estimated at 1 in 300 (Hu *et al.*, 2020) and has been extensively studied in high-income countries. An increasing number of studies continue to refine prevalence estimates and identify new variants, using diverse datasets and diagnostic criteria. To date, more than 2000 mutations have been reported in the LDLR gene, some of which are known to cause FH (Leigh *et al.*, 2017). Current variants and information on their potential causality of FH are listed on the Global Variome shared Leiden Open Variation Database (LOVD) which is frequently updated (https://databases.lovd.nl/shared/genes/LDLR).

Global estimates of FH prevalence often assume a uniform distribution across populations, overlooking continental and regional genetic diversity. In reality, most countries lack reliable FH registries, and diagnostic criteria vary widely, limiting the accuracy and comparability of prevalence estimates. Research-based estimates of prevalence include 1 in 217 in Denmark

(Benn *et al.*, 2016), 1 in 300 in the Netherlands (Besseling *et al.*, 2014) and 1 in 250 in the United States (de Ferranti *et al.*, 2016). Based on these values, the global prevalence of FH was estimated to range between 1 in 200 and 1 in 500 (Arca, 2017), with another systematic review suggesting 1 in 250 (Akioyamen *et al.*, 2017).

A meta-analysis in 2020 (Beheshti *et al.*, 2020) was performed to assess the worldwide prevalence and included 11 million subjects across multiple studies and found that 1 in 313 individuals in the general populations have heterozygous FH. Previously, the worldwide prevalence of homozygous FH was believed to be 1 in a million, however this study estimated a prevalence closer to 1 in 400 000. As previously mentioned, however, accurate prevalence across the world is difficult to determine as different regions present with different numbers. There are many possible explanations, for example whether individuals in a country are being correctly diagnosed either with genetic testing or clinically, individual population risk and the presence of founder mutations. The lack of data in multiple regions is also evident, highlighting the need for further investigations, both within countries and within different ethnicities.

In Africa specifically, there is a clear paucity of data with very few studies on prevalence published. In South Africa, the prevalence has been estimated to be between 1 in 70 and 1 in 100 in specific ethnic groups (Austin *et al.*, 2004). In 2018 we published a review to highlight the heterogeneity of FH within South Africa. There are multiple different ethnicities within the country, each with their own prevalence and most due to founder mutations (Figure 2) (Smyth, Ramsay and Raal, 2018). The prevalence in the Afrikaners (1 in 83), Ashkenazi Jewish population (1 in 67) and the Indians (1 in 100) is high, but there are no accurate estimates for Black South Africans. The review therefore emphasized the lack of data on black Africans in South Africans and

the rest of the continent. There was a study published in 1993 evaluating FH in Tunisia, where the prevalence was estimated to be 1 in 165 (Slimane *et al.*, 1993). Utilizing the data from two countries does not allow for an estimate of FH numbers across the African continent, as genetic variant vastly differs between regions.

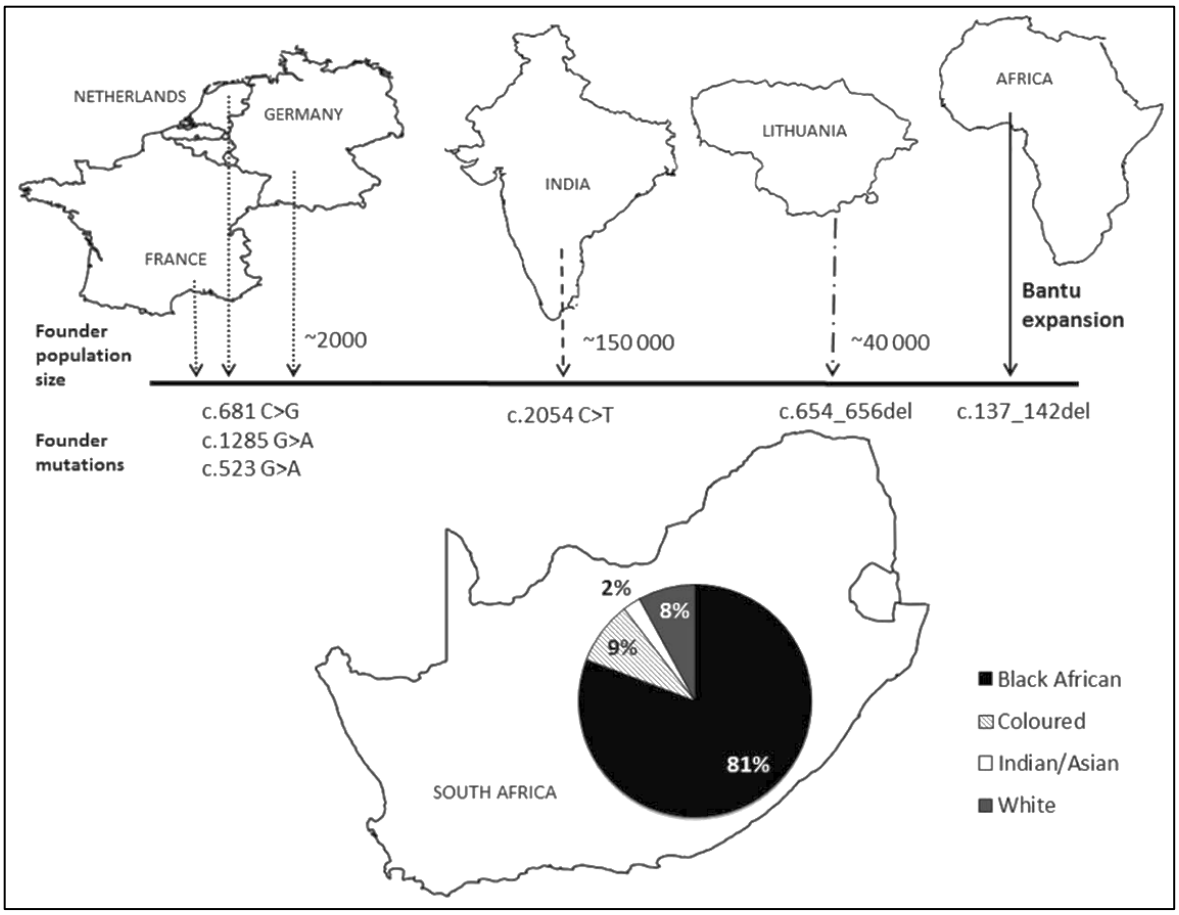


Figure 2: Illustration depicting the origins of founder mutations in LDLR in South African populations. Each founder mutation's origin is shown, along with the size of the founder population that migrated to South Africa. Pie chart depicts the percentage of each ethnic group in the entire South African population of ~65 million people. There are three founder mutations from Europe (present in White and Coloured South African families), one from India (present in Indian and Coloured populations), one from Lithuania (present in Jewish families) and one of African origin (found in black African families).

1.4. Identifying Disease-Causing and Disease-Associated Variants

Previously, DNA samples from individuals with dyslipidaemia were sequenced using Sanger sequencing of coding regions in one gene at a time. But with the development of massively parallel high-throughput DNA sequencing, also known as Next Generation Sequencing (NGS), the time and cost of variant detection has reduced. The flood of NGS data has led to the development of capable computational bioinformatic analyses. Whole genome sequencing (WGS) and whole exome sequencing (WES) have successfully been used to discover causative mutations in hundreds of different disorders and subsequently has identified new genes previously unknown to be involved in pathogenesis (Hegele *et al.*, 2015). For dyslipidaemias,

the identification of novel functional mutations in known genes has become the rule and not the exception and has led to the development of gene panels for targeted NGS. There are at least three gene panels for targeted NGS protocols used to detect variants causing FH and they all include the three major genes for autosomal dominant FH (*LDLR, APOB* and *PCSK9*). At the Robart's Institute in Canada, a targeted NGS panel termed LipidSeq, is used for diagnosis of dyslipidaemia causing variants. The panel uses the Illumina MiSeq platform (sequencing-by-synthesis) to screen coding regions and intron-exon boundaries of 69 candidate genes for dyslipidaemias. The cost of the panel is less than half the cost of Sanger sequencing of the coding regions of only *LDLR, APOB* and *PCSK9* (Hegele *et al.*, 2015). However, to validate findings from the LipidSeq panel, Sanger sequencing of the region surrounding the variants is performed. It has been shown that 5-20% of *LDLR* variants are copy number variations (CNVs) (Hegele *et al.*, 2015) and previously, CNVs were detected using a multiplex ligation-dependent probe amplification (MLPA). Studies involving *LDLR*, where NGS is used alongside MLPA, have shown a 90% concordance in CNV detection (Faiz *et al.*, 2013). A study published in 2019 showed a whole-gene duplication of *PCSK9* detected by NGS using a high depth of coverage analysis (300 X, on average) and was verified using MPLA (Iacocca *et al.*, 2018). NGS data and powerful bioinformatic tools have led to a more economic approach for CNV detection however there are limitations. For example, determining the breakpoints and location of the CNV, should it be a duplication, is challenging to detect using only short read NGS.

When studying African populations, there is a limitation to using a targeted NGS panel designed using only European genetic data. These panels will not detect variants in additional genes that, when mutated, contribute to causing the phenotype in other ethnic groups, such as Africans. Targeted NGS panels designed specifically for European populations might therefore not be appropriate for variant discovery in African populations, where other genes may be involved. The NGS technology remains a powerful approach for mutation discovery, but single molecule sequencing, using PacBio and Nanopore technologies, have the advantage of long sequence reads that enable the accurate detection of CNVs. In this study, however, I used Illumina sequencing-by-synthesis technology and a chosen gene panel enriched for genes in monogenic cases of increased or decreased LDL-C levels to detect variants, since this is an affordable and feasible approach in the context of the research project.

1.4.1. Assessing Variant Pathogenicity

Determining the pathogenicity of variants detected in NGS requires a multistep process. Bioinformatic techniques enable the classification of the potential effects of variants on the

encoded protein, for example, whether it is a synonymous or nonsynonymous amino acid substitution, splice variant, or indel. The next step is to search publicly available databases such as 1000 Genomes, dbSNP and gnomAD in order to determine whether variants are novel or have been previously reported and if so, to get information on the frequencies of variants in different populations.

There are many *in silico* tools that can be used for the interpretation of the function of sequence variants. Each tool utilizes a different algorithm, but their purpose is to predict the effect of a variant (Ohanian, Otway and Fatkin, 2012). The evolutionary conservation of an amino acid or nucleotide, the location within the protein sequence, and the biochemical change of the amino acid substitution are the criteria that can influence the effect of a missense mutation (Richards et al, 2017). In depth comparisons of the performance of widely used pathogenicity prediction methods have shown 65-80% accuracy when evaluating known diseases (Thusberg et al, 2011). An example of an analysis method to assess the impact of a variant would be the Variant Effect Predictor (VEP) which includes estimates of pathogenicity by PolyPhen and SIFT which are of the more commonly used tools for coding variants (McLaren *et al.*, 2016). A combination of different tools will be used as each program can have its own strengths and weaknesses depending on its particular algorithm.

1.4.2. Genetic Complexity of Dyslipidaemia

PGSs for LDL-C have emerged as promising tools for CVD risk stratification. Studies have shown that individuals in the highest PGS percentiles for LDL-C have substantially elevated risk of CAD, independent of monogenic mutations or measured LDL-C levels, suggesting PGSs capture cumulative lifelong exposure to high LDL-C (Khera *et al.*, 2018). It has been suggested that PGSs are a more reliable metric to assess CVD risk than LDL-C measurements and that it may assist in identifying groups of people or populations which have a lifelong exposure to elevated LDL-C and thus increased risk of CVD, and that may benefit from early treatment (Natarajan *et al.*, 2017).

Studies have revealed notable epistatic interactions between lipid-regulating genes that modulate LDL-C levels in ways not predictable from single-gene effects. In a cohort of 3,031 FH patients, heterozygous carriers of *ABCG5* or *ABCG8* variants exhibited lower LDL-C (~6.2 ± 1.7 mmol/L) than individuals with pathogenic *LDLR* mutations (~7.2 ± 1.7 mmol/L). Importantly, carrying both an *ABCG5* and an *ABCG8* variant did not further elevate LDL-C compared to *LDLR* mutation carriers alone (7.8 vs. 7.2 mmol/L, $p = 0.259$), illustrating a

protective, non-additive epistatic effect (Reeskamp *et al.*, 2020). Similarly, *PCSK9* variants can act epistatically with *LDLR*. Mouse models lacking both *Pcsk9* and *Ldlr* showed reduced apolipoprotein B secretion and less atherosclerosis compared to mice lacking *Ldlr* alone, indicating that *Pcsk9*'s effect depends on the presence of Ldlr and exemplifies gene–gene dependency (Sun *et al.*, 2018).

1.5. Dyslipidaemia in Africa

There remains a significant gap in genetic research involving African populations, underscoring the need to develop well-characterised African cohorts. Such efforts would create opportunities to uncover lipid-associated variants unique to African ancestries. It has been proposed that next-generation sequencing is the most effective approach for identifying these variants in both continental African and African diaspora populations (Wright, Housman and Taylor, 2016). Black South Africans with FH appear to be uncommon and are understudied; with only one publication since 2015. This study is somewhat flawed in that it analysed only selected regions of *LDLR*, *APOB*, and *PCSK9* for FH and missed important variants in regions that were not investigated. Nonetheless, in the 16 unrelated black South African patients of different language groups, they identified some FH-causing mutations. Eight *LDLR* mutations were identified, two of which were novel and predicted to be pathogenic by *in silico* tools (Ibe *et al.*, 2017). This study emphasises the genetic diversity and heterogeneity within African populations and highlights the importance of next-generation sequencing approaches, such as the use of extended gene panels, to comprehensively identify potentially pathogenic variants that may underlie FH in Black Africans, both in South Africa and across the continent.

Studies in African American populations have shown that, on average, individuals with a higher proportion of African ancestry tend to have lower LDL-C levels and higher triglyceride levels (Bentley and Rotimi, 2012). These findings highlight a significant genetic contribution to cholesterol levels and underscore the presence of trans-ethnic variation. Among the numerous GWASs conducted worldwide, only 0.16% of total participants are African and this relative proportion of African participants is declining (Martin *et al.*, 2018; Mills and Rahal, 2020; *GWAS Diversity Monitor*). Clinically evident dyslipidaemia arises from a complex interplay between genetic predisposition and secondary genetic, epigenetic, and environmental influences. Historically, hyperlipidaemia—particularly elevated LDL-C levels—has been considered uncommon in individuals of African descent (Agongo *et al.*, 2018), a perception largely attributed to dietary or environmental factors. However, this assumption persists despite evidence that 40–60% of the variance in plasma lipid levels is heritable. This raises an

important question: is the perceived lower prevalence a consequence of under diagnosis and limited data from large African cohorts, or could it reflect a higher frequency of lipid-lowering genetic variants in these populations? Emerging evidence supports the latter, including a study identifying *PCSK9* loss-of-function mutations in 3.7% of Black Zimbabweans, associated with a 27% reduction in LDL-C (Hooper *et al.*, 2007). Similarly, in African American populations, *PCSK9* loss-of-function variants were found in 2.6% of individuals and linked to a 28% decrease in LDL-C (Cohen *et al.*, 2006).

1.6. AWI-Gen

This study was performed using selected individuals from the AWI-Gen study cohort. The AWI-Gen research platform is a population cross-sectional study that includes over 12 000 participants (males and females; 40 to 60 years) from Burkina Faso (Nanoro), Ghana (Navrongo), Kenya (Nairobi) and South Africa (Agincourt, Dikgale and Soweto). The study was designed to explore genomic and environmental factors influencing body composition, distribution of body fat, and risk of developing CVD.

Established under the Human Heredity and Health in Africa (H3Africa) initiative, AWI-Gen generated a rich dataset that includes fasting lipid profiles (total cholesterol, HDL-C, LDL-C, and triglycerides), alongside genomic, behavioural, and environmental variables. Biospecimens were collected, including fasting venous blood, which enabled a range of biochemical assays to be performed. Standardised anthropometric and physiological measurements were also obtained, and participants completed comprehensive questionnaires covering demographics, health history, socioeconomic status, behavioural risk factors, and access to care (Ramsay *et al.*, 2016; Ali *et al.*, 2018).

The cohort is particularly valuable for investigating dyslipidaemia and cardiovascular risk in Africa due to its geographic diversity, phenotyping, and availability of genomic data. Importantly, several AWI-Gen sub-studies have already examined dyslipidaemia and its correlates across regions. For instance, in a sub-study conducted in rural northern Ghana, a relatively low burden of dyslipidaemia was reported, with only 4% of participants showing elevated LDL-C levels (Agongo *et al.*, 2018). In addition, the AWI-Gen study of multimorbidity in middle-aged African adults highlighted how dyslipidaemia frequently co-occurs with hypertension, obesity, and diabetes (Micklesfield *et al.*, 2023). The study showed significant regional variation in multimorbidity patterns, with lipid abnormalities contributing to cardiometabolic risk factors.

These studies underscore the need for region-specific genomic tools and clinical guidelines tailored to the unique metabolic profiles observed across African settings. In particular, the geographical and ethnic diversity within AWI-Gen offers a unique opportunity to explore how genetic variation, including both rare and common variants, contributes to lipid regulation and cardiovascular risk in underrepresented populations. Overall, the AWI-Gen cohort offers an ideal foundation for this PhD project, which aims to characterise the genetic determinants of LDL-C levels to investigate monogenic FH and low levels of LDL-C in African populations. Its comprehensive phenotyping, availability of genomic data, and representation of diverse African regions make it a powerful resource for understanding the molecular architecture of dyslipidaemia on the continent.

1.7. Study Rationale

There is a paucity of data and research regarding the genetic underpinning of high and low levels of LDL-C in African populations, presenting an opportunity for research. This study is nested in the AWI-Gen study, leveraging infrastructure, base-line data, an experienced team of investigators and field workers, and stored DNA samples (Ramsay et al. 2016; Ali et al. 2018). The study for this PhD aimed to develop a targeted NGS gene panel with genes involved in LDL-C levels to identify possible monogenic causal variants. The AWI-Gen genome-wide genotype data was used to assess the transferability of PGSs developed from European and other African-ancestry populations.

Due to dyslipidaemia being multifactorial, several confounding variables was noted. HIV infection and its treatment can affect lipid levels in the blood. In a cohort from Limpopo, South Africa, HIV infection was shown to be associated with low LDL-C levels (Vos *et al.*, 2017) and it was determined that if women were on HIV treatment, they had elevated levels of LDL-C (Anastos *et al.*, 2007). Obesity and diabetes mellitus are also potential confounders of LDL-C levels, together with lifestyle factors. To study the genetics of LDL-C levels, participants with these conditions could mask some of the genetic contributions, and were excluded, as appropriate and according to the study design and aim.

In the first part of the study the data generated from the NGS was used to identify potentially functional high-effect variants in the exonic regions, promoter regions and exon/intron junctions, and to assess the in silico predicted impact on high or low LDL-C levels in individuals. The concurrent presence of both pathogenic and protective variants for high LDL-

C levels, could contribute to understanding variant penetrance and may provide insights into dyslipidaemia.

The second part of the study focussed on assessing the predictivity of PGS models developed in different studies when applied to the AWI-Gen cohort to assess of a broader spectrum of genetic contributions to variation in LDL-C profiles.

1.8. Summary of Relevant Phenotypes for this study

The phenotypes referred to in this thesis are: (1) dyslipidaemia which is defined as an abnormal blood lipid profile that is based on measured values of LDL-C, HDL-C, TC and TG and specific cut off values that suggest increased risk for cardiovascular diseases, (2) LDL-C levels which refers to measured level of cholesterol carried within low-density lipoprotein particles and is a primary atherogenic lipid; extensive genetic, epidemiological, and interventional evidence shows that cumulative life-time exposure to higher LDL-C causally drives the development of atherosclerotic plaques, (3) atherosclerotic cardiovascular disease (ASCVD) is a form of CVD that refers to the build-up of plaque (referred to as atherosclerosis) in the arteries and has specific clinical consequences such as coronary artery disease, myocardial infarction, ischaemic stroke, and peripheral artery disease and is the major clinical endpoint of concern when considering lipid abnormalities, (4) familial hypercholesterolaemia (FH) is the heritable form of severe hyper-LDL-cholesterolaemia usually caused by pathogenic variants in specific genes (with and autosomal dominant monogenic inheritance pattern). Individuals with FH experience lifelong, markedly elevated LDL-C and consequently have a substantially increased and earlier onset risk of ASCVD unless lipid-lowering therapy is initiated. Finally, (5) the umbrella term cardiovascular disease (CVD) is a collection of disorders that effect the blood vessels and the heart, and includes ASCVD.

1.9. Aim and Objectives

Aim

To identify potential monogenic causes of extreme high and low LDL-cholesterol levels and to assess the transferability of PGSs developed in several studies to different African populations.

Objectives

1. To determine the genetic basis of high and low LDL-C levels in African populations by identifying potential functional variants in genes known to be associated with high and low LDL-C using a targeted next-generation sequencing approach.
2. To assess the performance of several polygenic risk score models for high LDL-C developed in European and non-continental Africa-ancestry populations in continental African populations.

Chapter 2 addresses the first objective of the thesis by investigating the genetic basis of high and low LDL-C levels in continental African populations. Using targeted next-generation sequencing, the chapter explores potential functional variants in genes known to be associated with LDL-C regulation, such as *LDLR*, *APOB*, and *PCSK9*. It evaluates the contribution of rare, likely pathogenic mutations to LDL-C extremes, providing insight into the monogenic architecture of dyslipidaemia in Africa. Chapter 3 focuses on the second objective and evaluates the performance of PGSs derived either European or African cohorts in predicting elevated LDL-C in continental African populations. This chapter compares four published PGS models across Sub-Saharan African regions, assessing their transferability, predictive accuracy, and potential limitations in ancestrally diverse settings. The findings contribute to a better understanding of the utility and challenges of implementing polygenic risk prediction in underrepresented populations.

1.10. References

Adi, D. *et al.* (2020) 'IDOL gene variant is associated with hyperlipidemia in Han population in Xinjiang, China', *Scientific Reports*, 10(1), p. 14280. Available at: https://doi.org/10.1038/s41598-020-71241-1.

Agongo, G. *et al.* (2018) 'The burden of dyslipidaemia and factors associated with lipid levels among adults in rural northern Ghana: An AWI-Gen sub-study.', *PloS one*, 13(11), p. e0206326. Available at: https://doi.org/10.1371/journal.pone.0206326.

Akioyamen, L.E. *et al.* (2017) 'Estimating the prevalence of heterozygous familial hypercholesterolaemia: a systematic review and meta-analysis.', *BMJ open*, 7(9), p. e016461. Available at: https://doi.org/10.1136/bmjopen-2017-016461.

Alhuneafat, L. *et al.* (2024) 'Burden of cardiovascular disease in Sub-Saharan Africa, 1990-2019: An analysis of the Global Burden of Disease Study.', *Current problems in cardiology*, 49(6), p. 102557. Available at: https://doi.org/10.1016/j.cpcardiol.2024.102557.

Ali, S.A. *et al.* (2018) 'Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study.', *Global*

health action, 11(sup2), p. 1507133. Available at: https://doi.org/10.1080/16549716.2018.1507133.

Altmann, S.W. *et al.* (2004) 'Niemann-Pick C1 Like 1 protein is critical for intestinal cholesterol absorption.', *Science (New York, N.Y.)*, 303(5661), pp. 1201–1204. Available at: https://doi.org/10.1126/science.1093131.

Anastos, K. *et al.* (2007) 'Association of serum lipid levels with HIV serostatus, specific antiretroviral agents, and treatment regimens.', *Journal of acquired immune deficiency syndromes (1999)*, 45(1), pp. 34–42. Available at: https://doi.org/10.1097/QAI.0b013e318042d5fe.

Ando, Y. *et al.* (2003) 'A decreased expression of angiopoietin-like 3 is protective against atherosclerosis in apoE-deficient mice.', *Journal of lipid research*, 44(6), pp. 1216–1223. Available at: https://doi.org/10.1194/jlr.M300031-JLR200.

Antonarakis, S.E. and Beckmann, J.S. (2006) 'Mendelian disorders deserve more attention.', *Nature reviews. Genetics*, 7(4), pp. 277–282. Available at: https://doi.org/10.1038/nrg1826.

Arca, M. (2017) 'Old challenges and new opportunities in the clinical management of heterozygous familial hypercholesterolemia (HeFH): The promises of PCSK9 inhibitors.', *Atherosclerosis*, 256, pp. 134–145. Available at: https://doi.org/10.1016/j.atherosclerosis.2016.09.001.

Auclair, N. *et al.* (2023) 'High-fat diet reveals the impact of Sar1b defects on lipid and lipoprotein profile and cholesterol metabolism.', *Journal of lipid research*, 64(9), p. 100423. Available at: https://doi.org/10.1016/j.jlr.2023.100423.

Austin, M.A. *et al.* (2004) 'Genetic causes of monogenic heterozygous familial hypercholesterolemia: a HuGE prevalence review.', *American journal of epidemiology*, 160(5), pp. 407–420. Available at: https://doi.org/10.1093/aje/kwh236.

Beheshti, S.O. *et al.* (2020) 'Worldwide Prevalence of Familial Hypercholesterolemia: Meta-Analyses of 11 Million Subjects.', *Journal of the American College of Cardiology*, 75(20), pp. 2553–2566. Available at: https://doi.org/10.1016/j.jacc.2020.03.057.

Benn, M. *et al.* (2016) 'Mutations causative of familial hypercholesterolaemia: screening of 98 098 individuals from the Copenhagen General Population Study estimated a prevalence of 1 in 217.', *European heart journal*, 37(17), pp. 1384–1394. Available at: https://doi.org/10.1093/eurheartj/ehw028.

Bentley, A.R. and Rotimi, C.N. (2012) 'Interethnic Variation in Lipid Profiles: Implications for Underidentification of African-Americans at risk for Metabolic Disorders.', *Expert review of endocrinology & metabolism*, 7(6), pp. 659–667. Available at: https://doi.org/10.1586/eem.12.55.

Berberich, A.J. and Hegele, R.A. (2019) 'The complex molecular genetics of familial hypercholesterolaemia', *Nature Reviews Cardiology*, 16(1), pp. 9–20. Available at: https://doi.org/10.1038/s41569-018-0052-6.

Berge, K.E., Ose, L. and Leren, T.P. (2006) 'Missense mutations in the PCSK9 gene are associated with hypocholesterolemia and possibly increased response to statin therapy.',

Arteriosclerosis, thrombosis, and vascular biology, 26(5), pp. 1094–1100. Available at: https://doi.org/10.1161/01.ATV.0000204337.81286.1c.

Besseling, J. *et al.* (2014) 'Severe heterozygous familial hypercholesterolemia and risk for cardiovascular disease: a study of a cohort of 14,000 mutation carriers.', *Atherosclerosis*, 233(1), pp. 219–223. Available at: https://doi.org/10.1016/j.atherosclerosis.2013.12.020.

Bonnefont-Rousselot, D. *et al.* (2009) 'Cryptogenic cirrhosis in a patient with familial hypocholesterolemia due to a new truncated form of apolipoprotein B.', *European journal of gastroenterology & hepatology*, 21(1), pp. 104–108. Available at: https://doi.org/10.1097/MEG.0b013e3282ffd9f8.

Borén, J. *et al.* (2001) 'The molecular mechanism for the genetic disorder familial defective apolipoprotein B100.', *The Journal of biological chemistry*, 276(12), pp. 9214–9218. Available at: https://doi.org/10.1074/jbc.M008890200.

Bouhairie, V.E. and Goldberg, A.C. (2015) 'Familial Hypercholesterolemia', *Lipidology*, 33(2), pp. 169–179. Available at: https://doi.org/10.1016/j.ccl.2015.01.001.

Braun, K.V.E. *et al.* (2016) 'The role of DNA methylation in dyslipidaemia: A systematic review.', *Progress in lipid research*, 64, pp. 178–191. Available at: https://doi.org/10.1016/j.plipres.2016.10.002.

Burke, A.C. *et al.* (2017) 'PCSK9: Regulation and Target for Drug Development for Dyslipidemia.', *Annual review of pharmacology and toxicology*, 57, pp. 223–244. Available at: https://doi.org/10.1146/annurev-pharmtox-010716-104944.

Burnett, J.R., Hooper, A.J. and Hegele, R.A. (1993) 'APOB-Related Familial Hypobetalipoproteinemia.', in M.P. Adam et al. (eds) *GeneReviews(®)*. Seattle (WA): University of Washington, Seattle.

Cartier, J.L. and Goldberg, A.C. (2016) 'Familial Hypercholesterolemia: Advances in Recognition and Therapy.', *Progress in cardiovascular diseases*, 59(2), pp. 125–134. Available at: https://doi.org/10.1016/j.pcad.2016.07.006.

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.

Clebak, K.T. and Dambro, A.B. (2020) 'Hyperlipidemia: An Evidence-based Review of Current Guidelines.', *Cureus*, 12(3), p. e7326. Available at: https://doi.org/10.7759/cureus.7326.

Cohen, J. *et al.* (2005) 'Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9.', *Nature genetics*, 37(2), pp. 161–165. Available at: https://doi.org/10.1038/ng1509.

Cohen, Jonathan C. *et al.* (2006) 'Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels.', *Proceedings of the National Academy of Sciences of the United States of America*, 103(6), pp. 1810–1815. Available at: https://doi.org/10.1073/pnas.0508483103.

Cohen, Jonathan C *et al.* (2006) 'Sequence variations in PCSK9, low LDL, and protection against coronary heart disease.', *The New England journal of medicine*, 354(12), pp. 1264–1272. Available at: https://doi.org/10.1056/NEJMoa054013.

Conlon, D.M. (2019) 'Role of sortilin in lipid metabolism.', *Current opinion in lipidology*, 30(3), pp. 198–204. Available at: https://doi.org/10.1097/MOL.0000000000000598.

Coutinho, M.F. *et al.* (2013) 'Sortilina e risco de doença cardiovascular', *Revista Portuguesa de Cardiologia*, 32(10), pp. 793–799. Available at: https://doi.org/10.1016/j.repc.2013.02.006.

Davidson, M.H. and Toth, P.P. (2004) 'Comparative effects of lipid-lowering therapies.', *Progress in cardiovascular diseases*, 47(2), pp. 73–104. Available at: https://doi.org/10.1016/j.pcad.2004.04.007.

Dean, M., Hamon, Y. and Chimini, G. (2001) 'The human ATP-binding cassette (ABC) transporter superfamily.', *Journal of lipid research*, 42(7), pp. 1007–1017.

Do, R. *et al.* (2015) 'Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction.', *Nature*, 518(7537), pp. 102–106. Available at: https://doi.org/10.1038/nature13917.

Dron, J.S. and Hegele, R.A. (2018) 'Polygenic influences on dyslipidemias.', *Current opinion in lipidology* [Preprint]. Available at: https://doi.org/10.1097/MOL.0000000000000482.

Dron, J.S. and Hegele, R.A. (2019) 'The evolution of genetic-based risk scores for lipids and cardiovascular disease.', *Current opinion in lipidology*, 30(2), pp. 71–81. Available at: https://doi.org/10.1097/MOL.0000000000000576.

Duncan, L. *et al.* (2019) 'Analysis of polygenic risk score usage and performance in diverse human populations.', *Nature communications*, 10(1), p. 3328. Available at: https://doi.org/10.1038/s41467-019-11112-0.

Eyrich, T.M. *et al.* (2024) 'Polygenic risk of high LDL cholesterol and ischemic heart disease in the general population', *Atherosclerosis*, 397. Available at: https://doi.org/10.1016/j.atherosclerosis.2024.118574.

Faiz, F. *et al.* (2013) 'Detection of variations and identifying genomic breakpoints for large deletions in the LDLR by Ion Torrent semiconductor sequencing.', *Atherosclerosis*, 230(2), pp. 249–255. Available at: https://doi.org/10.1016/j.atherosclerosis.2013.07.050.

Farhan, M. *et al.* (2025) 'Evaluating the role of PCSK9 inhibitors in reducing cardiovascular events among statin-intolerant patients: a systematic review and meta-analysis.', *Annals of medicine and surgery (2012)*, 87(2), pp. 891–899. Available at: https://doi.org/10.1097/MS9.0000000000002927.

Fellin, R. *et al.* (2015) 'The history of Autosomal Recessive Hypercholesterolemia (ARH). From clinical observations to gene identification.', *Gene*, 555(1), pp. 23–32. Available at: https://doi.org/10.1016/j.gene.2014.09.020.

Ference, B.A. *et al.* (2017) 'Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement

from the European Atherosclerosis Society Consensus Panel.', *European heart journal*, 38(32), pp. 2459–2472. Available at: https://doi.org/10.1093/eurheartj/ehx144.

de Ferranti, S.D. *et al.* (2016) 'Prevalence of Familial Hypercholesterolemia in the 1999 to 2012 United States National Health and Nutrition Examination Surveys (NHANES).', *Circulation*, 133(11), pp. 1067–1072. Available at: https://doi.org/10.1161/CIRCULATIONAHA.115.018791.

Fouchier, S.W. *et al.* (2014) 'Mutations in STAP1 Are Associated With Autosomal Dominant Hypercholesterolemia', *Circulation Research*, 115(6), pp. 552–555. Available at: https://doi.org/10.1161/CIRCRESAHA.115.304660.

Francke, U., Brown, M.S. and Goldstein, J.L. (1984) 'Assignment of the human gene for the low density lipoprotein receptor to chromosome 19: synteny of a receptor, a ligand, and a genetic disease.', *Proceedings of the National Academy of Sciences of the United States of America*, 81(9), pp. 2826–2830. Available at: https://doi.org/10.1073/pnas.81.9.2826.

Futema, M. *et al.* (2015) 'Refinement of variant selection for the LDL cholesterol genetic risk score in the diagnosis of the polygenic form of clinical familial hypercholesterolemia and replication in samples from 6 countries.', *Clinical chemistry*, 61(1), pp. 231–238. Available at: https://doi.org/10.1373/clinchem.2014.231365.

Garcia-Calvo, M. *et al.* (2005) 'The target of ezetimibe is Niemann-Pick C1-Like 1 (NPC1L1).', *Proceedings of the National Academy of Sciences of the United States of America*, 102(23), pp. 8132–8137. Available at: https://doi.org/10.1073/pnas.0500269102.

'Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.' (2020) *Lancet (London, England)*, 396(10258), pp. 1204–1222. Available at: https://doi.org/10.1016/S0140-6736(20)30925-9.

Goldstein, J.L. and Brown, M.S. (2009) 'The LDL receptor.', *Arteriosclerosis, thrombosis, and vascular biology*, 29(4), pp. 431–438. Available at: https://doi.org/10.1161/ATVBAHA.108.179564.

Gomez, F., Hirbo, J. and Tishkoff, S.A. (2014) 'Genetic variation and adaptation in Africa: implications for human evolution and disease.', *Cold Spring Harbor perspectives in biology*, 6(7), p. a008524. Available at: https://doi.org/10.1101/cshperspect.a008524.

Gouda, H.N. *et al.* (2019) 'Burden of non-communicable diseases in sub-Saharan Africa, 1990-2017: results from the Global Burden of Disease Study 2017.', *The Lancet. Global health*, 7(10), pp. e1375–e1387. Available at: https://doi.org/10.1016/S2214-109X(19)30374-2.

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.

GWAS Diversity Monitor (no date). Available at: https://www.gwasdiversitymonitor.com/ (Accessed: 26 June 2025).

Hegele, R.A. *et al.* (2005) 'NPC1L1 haplotype is associated with inter-individual variation in plasma low-density lipoprotein response to ezetimibe.', *Lipids in health and disease*, 4, p. 16. Available at: https://doi.org/10.1186/1476-511X-4-16.

Hegele, R.A. *et al.* (2015) 'Targeted next-generation sequencing in monogenic dyslipidemias.', *Current opinion in lipidology*, 26(2), pp. 103–113. Available at: https://doi.org/10.1097/MOL.0000000000000163.

Hooper, A.J. *et al.* (2007) 'The C679X mutation in PCSK9 is present and lowers blood cholesterol in a Southern African population.', *Atherosclerosis*, 193(2), pp. 445–448. Available at: https://doi.org/10.1016/j.atherosclerosis.2006.08.039.

Horton, J.D., Cohen, J.C. and Hobbs, H.H. (2007) 'Molecular biology of PCSK9: its role in LDL metabolism.', *Trends in biochemical sciences*, 32(2), pp. 71–77. Available at: https://doi.org/10.1016/j.tibs.2006.12.008.

Horton, J.D., Cohen, J.C. and Hobbs, H.H. (2009) 'PCSK9: a convertase that coordinates LDL catabolism.', *Journal of lipid research*, 50 Suppl(Suppl), pp. S172-177. Available at: https://doi.org/10.1194/jlr.R800091-JLR200.

Hu, P. *et al.* (2020) 'Prevalence of Familial Hypercholesterolemia Among the General Population and Patients With Atherosclerotic Cardiovascular Disease', *Circulation*, 141(22), pp. 1742–1759. Available at: https://doi.org/10.1161/CIRCULATIONAHA.119.044795.

Huijgen, R., Kindt, I., *et al.* (2012) 'Cardiovascular risk in relation to functionality of sequence variants in the gene coding for the low-density lipoprotein receptor: a study among 29 365 individuals tested for 64 specific low-density lipoprotein-receptor sequence variants', *European Heart Journal*, 33(18), pp. 2325–2330. Available at: https://doi.org/10.1093/eurheartj/ehs038.

Huijgen, R., Hutten, B.A., *et al.* (2012) 'Discriminative Ability of LDL-Cholesterol to Identify Patients With Familial Hypercholesterolemia', *Circulation: Cardiovascular Genetics*, 5(3), pp. 354–359. Available at: https://doi.org/10.1161/CIRCGENETICS.111.962456.

Iacocca, M.A. *et al.* (2018) 'Whole-Gene Duplication of PCSK9 as a Novel Genetic Mechanism for Severe Familial Hypercholesterolemia.', *The Canadian journal of cardiology*, 34(10), pp. 1316–1324. Available at: https://doi.org/10.1016/j.cjca.2018.07.479.

Ibe, U.K. *et al.* (2017) 'Analysis of mutations causing familial hypercholesterolaemia in black South African patients of different ancestr.', *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde*, 107(2), pp. 145–148.

Ibrahim, S. *et al.* (2024) 'LDLR variant classification for improved cardiovascular risk prediction in familial hypercholesterolemia.', *Atherosclerosis*, 397, p. 117610. Available at: https://doi.org/10.1016/j.atherosclerosis.2024.117610.

Ison, H.E., Clarke, S.L. and Knowles, J.W. (2025) 'Familial Hypercholesterolemia', in *GeneReviews® [Internet]*. University of Washington, Seattle. Available at: https://www.ncbi.nlm.nih.gov/books/NBK174884/ (Accessed: 20 June 2025).

Jeon, H. and Blacklow, S.C. (2005) 'Structure and physiologic function of the low-density lipoprotein receptor.', *Annual review of biochemistry*, 74, pp. 535–562. Available at: https://doi.org/10.1146/annurev.biochem.74.082803.133354.

Kalwick, M. and Roth, M. (2025) 'A Comprehensive Review of the Genetics of Dyslipidemias and Risk of Atherosclerotic Cardiovascular Disease', *Nutrients*, 17(4). Available at: https://doi.org/10.3390/nu17040659.

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.

Kersten, S. (2017) 'Angiopoietin-like 3 in lipoprotein metabolism', *Nature Reviews Endocrinology*, 13(12), pp. 731–739. Available at: https://doi.org/10.1038/nrendo.2017.119.

Khalil, Y.A. *et al.* (2021) 'APOE gene variants in primary dyslipidemia.', *Atherosclerosis*, 328, pp. 11–22. Available at: https://doi.org/10.1016/j.atherosclerosis.2021.05.007.

Khera, A.V. *et al.* (2018) 'Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations', *Nature Genetics*, 50(9), pp. 1219–1224. Available at: https://doi.org/10.1038/s41588-018-0183-z.

Lakoski, S.G. *et al.* (2009) 'Genetic and metabolic determinants of plasma PCSK9 levels.', *The Journal of clinical endocrinology and metabolism*, 94(7), pp. 2537–2543. Available at: https://doi.org/10.1210/jc.2009-0141.

Lamiquiz-Moneo, I. *et al.* (2020) 'Predicted pathogenic mutations in STAP1 are not associated with clinically defined familial hypercholesterolemia', *Atherosclerosis*, 292, pp. 143–151. Available at: https://doi.org/10.1016/j.atherosclerosis.2019.11.025.

Lamiquiz-Moneo, I. *et al.* (2021) 'Diagnostic yield of sequencing familial hypercholesterolemia genes in individuals with primary hypercholesterolemia.', *Revista espanola de cardiologia (English ed.)*, 74(8), pp. 664–673. Available at: https://doi.org/10.1016/j.rec.2020.06.003.

Leigh, S. *et al.* (2017) 'The UCL low-density lipoprotein receptor gene variant database: pathogenicity update.', *Journal of medical genetics*, 54(4), pp. 217–223. Available at: https://doi.org/10.1136/jmedgenet-2016-104054.

Li, F. and Zhang, H. (2019) 'Lysosomal Acid Lipase in Lipid Metabolism and Beyond.', *Arteriosclerosis, thrombosis, and vascular biology*, 39(5), pp. 850–856. Available at: https://doi.org/10.1161/ATVBAHA.119.312136.

Lin, Y. *et al.* (2010) 'Current status and future directions in lipid management: emphasizing low-density lipoproteins, high-density lipoproteins, and triglycerides as targets for therapy', *Vascular Health and Risk Management*, 6, pp. 73–85. Available at: https://doi.org/10.2147/vhrm.s8725.

Lindholm, D., Bornhauser, B.C. and Korhonen, L. (2009) 'Mylip makes an Idol turn into regulation of LDL receptor.', *Cellular and molecular life sciences : CMLS*, 66(21), pp. 3399–3402. Available at: https://doi.org/10.1007/s00018-009-0127-y.

Lonardo, A. *et al.* (1998) 'Familial heterozygous hypobetalipoproteinemia, extrahepatic primary malignancy, and hepatocellular carcinoma.', *Digestive diseases and sciences*, 43(11), pp. 2489–2492. Available at: https://doi.org/10.1023/a:1026646618643.

Marais, A.D. (2019) 'Apolipoprotein E in lipoprotein metabolism, health and cardiovascular disease.', *Pathology*, 51(2), pp. 165–176. Available at: https://doi.org/10.1016/j.pathol.2018.11.002.

Mariano, C. *et al.* (2020) 'The familial hypercholesterolaemia phenotype: Monogenic familial hypercholesterolaemia, polygenic hypercholesterolaemia and other causes', *Clinical Genetics*, 97(3), pp. 457–466. Available at: https://doi.org/10.1111/cge.13697.

Martin, A.R. *et al.* (2018) 'The critical needs and challenges for genetic architecture studies in Africa.', *Current opinion in genetics & development*, 53, pp. 113–120. Available at: https://doi.org/10.1016/j.gde.2018.08.005.

Martin, A.R. *et al.* (2019) 'Clinical use of current polygenic risk scores may exacerbate health disparities.', *Nature genetics*, 51(4), pp. 584–591. Available at: https://doi.org/10.1038/s41588-019-0379-x.

Maxwell, K.N. and Breslow, J.L. (2004) 'Adenoviral-mediated expression of Pcsk9 in mice results in a low-density lipoprotein receptor knockout phenotype.', *Proceedings of the National Academy of Sciences of the United States of America*, 101(18), pp. 7100–7105. Available at: https://doi.org/10.1073/pnas.0402133101.

Mayosi, B.M. *et al.* (2009) 'The burden of non-communicable diseases in South Africa.', *Lancet (London, England)*, 374(9693), pp. 934–947. Available at: https://doi.org/10.1016/S0140-6736(09)61087-4.

McLaren, W. *et al.* (2016) 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17(1), p. 122. Available at: https://doi.org/10.1186/s13059-016-0974-4.

Mensah, G.A. (2013) 'Descriptive epidemiology of cardiovascular risk factors and diabetes in sub-Saharan Africa.', *Progress in cardiovascular diseases*, 56(3), pp. 240–250. Available at: https://doi.org/10.1016/j.pcad.2013.10.014.

Micklesfield, L.K. *et al.* (2023) 'Identifying the prevalence and correlates of multimorbidity in middle-aged men and women: a cross-sectional population-based study in four African countries.', *BMJ open*, 13(3), p. e067788. Available at: https://doi.org/10.1136/bmjopen-2022-067788.

Mills, M.C. and Rahal, C. (2020) 'The GWAS Diversity Monitor tracks diversity by disease in real time', *Nature Genetics*, 52(3), pp. 242–243. Available at: https://doi.org/10.1038/s41588-020-0580-y.

Minicocci, I. *et al.* (2012) 'Mutations in the ANGPTL3 gene and familial combined hypolipidemia: a clinical and biochemical characterization.', *The Journal of clinical endocrinology and metabolism*, 97(7), pp. E1266-1275. Available at: https://doi.org/10.1210/jc.2012-1298.

Minja, N.W. *et al.* (2022) 'Cardiovascular diseases in Africa in the twenty-first century: Gaps and priorities going forward.', *Frontiers in cardiovascular medicine*, 9, p. 1008335. Available at: https://doi.org/10.3389/fcvm.2022.1008335.

Moustafa, B. *et al.* (2024) 'Efficacy and safety of PCSK9 inhibitors for stroke prevention: Systematic review and meta-analysis.', *Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association*, 33(4), p. 107633. Available at: https://doi.org/10.1016/j.jstrokecerebrovasdis.2024.107633.

Musialik, J. *et al.* (2020) 'A Rare Mutation in The APOB Gene Associated with Neurological Manifestations in Familial Hypobetalipoproteinemia.', *International journal of molecular sciences*, 21(4). Available at: https://doi.org/10.3390/ijms21041439.

Natarajan, P. *et al.* (2017) 'Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting', *Circulation*, 135(22), pp. 2091–2101. Available at: https://doi.org/10.1161/CIRCULATIONAHA.116.024436.

National Institute for Health and Care Excellence (NICE). (2023). *Familial hypercholesterolaemia: identification and management (Clinical guideline [CG71]) —* updated guidance. Available at: https://www.nice.org.uk/guidance/cg71
Non communicable diseases (no date). Available at: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (Accessed: 3 July 2025).

Ohanian, M., Otway, R. and Fatkin, D. (2012) 'Heuristic methods for finding pathogenic variants in gene coding sequences.', *Journal of the American Heart Association*, 1(5), p. e002642. Available at: https://doi.org/10.1161/JAHA.112.002642.

Park, S.W., Moon, Y.-A. and Horton, J.D. (2004) 'Post-transcriptional regulation of low density lipoprotein receptor protein by proprotein convertase subtilisin/kexin type 9a in mouse liver.', *The Journal of biological chemistry*, 279(48), pp. 50630–50638. Available at: https://doi.org/10.1074/jbc.M410077200.

Peck, R. *et al.* (2014) 'Preparedness of Tanzanian health facilities for outpatient primary care of hypertension and diabetes: a cross-sectional survey.', *The Lancet. Global health*, 2(5), pp. e285-292. Available at: https://doi.org/10.1016/S2214-109X(14)70033-6.

Peloso, G.M. *et al.* (2019) 'Rare Protein-Truncating Variants in APOB, Lower Low-Density Lipoprotein Cholesterol, and Protection Against Coronary Heart Disease.', *Circulation. Genomic and precision medicine*, 12(5), p. e002376. Available at: https://doi.org/10.1161/CIRCGEN.118.002376.

Raal, F.J., Hovingh, G.K. and Catapano, A.L. (2018) 'Familial hypercholesterolemia treatments: Guidelines and new therapies', *Atherosclerosis*, 277, pp. 483–492. Available at: https://doi.org/10.1016/j.atherosclerosis.2018.06.859.

Raal, F.J. *et al.* (2020) 'Cascade Screening for Familial Hypercholesterolemia in South Africa', *Arteriosclerosis, Thrombosis, and Vascular Biology*, 40(11), pp. 2747–2755. Available at: https://doi.org/10.1161/ATVBAHA.120.315040.

Rahalkar, A.R. and Hegele, R.A. (2008) 'Monogenic pediatric dyslipidemias: classification, genetics and clinical spectrum.', *Molecular genetics and metabolism*, 93(3), pp. 282–294. Available at: https://doi.org/10.1016/j.ymgme.2007.10.007.

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.

Rasmussen, K.L. *et al.* (2016) 'Data on plasma levels of apolipoprotein E, correlations with lipids and lipoproteins stratified by APOE genotype, and risk of ischemic heart disease.', *Data in brief*, 6, pp. 923–932. Available at: https://doi.org/10.1016/j.dib.2016.01.060.

Reeskamp, L.F. *et al.* (2020) 'ABCG5 and ABCG8 genetic variants in familial hypercholesterolemia.', *Journal of clinical lipidology*, 14(2), pp. 207-217.e7. Available at: https://doi.org/10.1016/j.jacl.2020.01.007.

Rogozik, J., Główczyńska, R. and Grabowski, M. (2024) 'Genetic backgrounds and diagnosis of familial hypercholesterolemia.', *Clinical genetics*, 105(1), pp. 3–12. Available at: https://doi.org/10.1111/cge.14435.

Rosenson, R.S. (2021) 'Existing and emerging therapies for the treatment of familial hypercholesterolemia.', *Journal of lipid research*, 62, p. 100060. Available at: https://doi.org/10.1016/j.jlr.2021.100060.

Ruiz-Iruela, C. *et al.* (2020) 'Genetic contribution to lipid target achievement with statin therapy: a prospective study', *The Pharmacogenomics Journal*, 20(3), pp. 494–504. Available at: https://doi.org/10.1038/s41397-019-0136-7.

Santos, R.D. *et al.* (2016) 'Defining severe familial hypercholesterolaemia and the implications for clinical management: a consensus statement from the International Atherosclerosis Society Severe Familial Hypercholesterolemia Panel.', *The lancet. Diabetes & endocrinology*, 4(10), pp. 850–861. Available at: https://doi.org/10.1016/S2213-8587(16)30041-9.

Serrano-Pozo, A., Das, S. and Hyman, B.T. (2021) 'APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches.', *The Lancet. Neurology*, 20(1), pp. 68–80. Available at: https://doi.org/10.1016/S1474-4422(20)30412-9.

Simon, J.S. *et al.* (2005) 'Sequence variation in NPC1L1 and association with improved LDL-cholesterol lowering in response to ezetimibe treatment.', *Genomics*, 86(6), pp. 648–656. Available at: https://doi.org/10.1016/j.ygeno.2005.08.007.

Slimane, M.N. *et al.* (1993) 'Phenotypic expression of familial hypercholesterolaemia in central and southern Tunisia.', *Atherosclerosis*, 104(1–2), pp. 153–158. Available at: https://doi.org/10.1016/0021-9150(93)90186-x.

Smyth, N., Ramsay, M. and Raal, F.J. (2018) 'Population specific genetic heterogeneity of familial hypercholesterolemia in South Africa.', *Current opinion in lipidology*, 29(2), pp. 72–79. Available at: https://doi.org/10.1097/MOL.0000000000000488.

Sun, H. *et al.* (2018) 'PCSK9 deficiency reduces atherosclerosis, apolipoprotein B secretion, and endothelial dysfunction', *Journal of Lipid Research*, 59(2), pp. 207–223. Available at: https://doi.org/10.1194/jlr.M078360.

Takahashi, M. *et al.* (2021) 'Current Diagnosis and Management of Abetalipoproteinemia.', *Journal of atherosclerosis and thrombosis*, 28(10), pp. 1009–1019. Available at: https://doi.org/10.5551/jat.RV17056.

Tarugi, P. and Lonardo, A. (1997) 'Heterozygous familial hypobetalipoproteinemia associated with fatty liver.', *The American journal of gastroenterology*, 92(8), pp. 1400–1402.

'Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report.' (2002) *Circulation*, 106(25), pp. 3143–3421.

Thompson, J.F. *et al.* (2009) 'Comprehensive whole-genome and candidate gene analysis for response to statin therapy in the Treating to New Targets (TNT) cohort.', *Circulation. Cardiovascular genetics*, 2(2), pp. 173–181. Available at: https://doi.org/10.1161/CIRCGENETICS.108.818062.

Totoń-Żurańska, J. *et al.* (2023) 'Targeted sequencing of a gene panel in patients with familial hypercholesterolemia from Southern Poland.', *Polish archives of internal medicine*, 133(6), p. 16417. Available at: https://doi.org/10.20452/pamw.16417.

Viturro, E. *et al.* (2006) 'Cholesterol and saturated fat intake determine the effect of polymorphisms at ABCG5/ABCG8 genes on lipid levels in children', *Genetics in Medicine*, 8(9), pp. 594–599. Available at: https://doi.org/10.1097/01.gim.0000237760.25195.e7.

Vos, A. *et al.* (2017) 'HIV and risk of cardiovascular disease in sub-Saharan Africa: Rationale and design of the Ndlovu Cohort Study.', *European journal of preventive cardiology*, 24(10), pp. 1043–1050. Available at: https://doi.org/10.1177/2047487317702039.

Wakabayashi, T. *et al.* (2024) 'Current Diagnosis and Management of Familial Hypobetalipoproteinemia 1.', *Journal of atherosclerosis and thrombosis*, 31(7), pp. 1005–1023. Available at: https://doi.org/10.5551/jat.RV22018.

Wang, J., Williams, C.M. and Hegele, R.A. (2005) 'Compound heterozygosity for two non-synonymous polymorphisms in NPC1L1 in a non-responder to ezetimibe.', *Clinical genetics*, 67(2), pp. 175–177. Available at: https://doi.org/10.1111/j.1399-0004.2004.00388.x.

Watts, G.F. *et al.* (2014) 'Integrated guidance on the care of familial hypercholesterolaemia from the International FH Foundation', *International Journal of Cardiology*, 171(3), pp. 309–325. Available at: https://doi.org/10.1016/j.ijcard.2013.11.025.

Weiss, L.A. *et al.* (2006) 'The sex-specific genetic architecture of quantitative traits in humans.', *Nature genetics*, 38(2), pp. 218–222. Available at: https://doi.org/10.1038/ng1726.

Wetterau, J.R. and Zilversmit, D.B. (1984) 'A triglyceride and cholesteryl ester transfer protein associated with liver microsomes.', *The Journal of biological chemistry*, 259(17), pp. 10863–10866.

Whitfield, A.J. *et al.* (2004) 'Lipid disorders and mutations in the APOB gene.', *Clinical chemistry*, 50(10), pp. 1725–1732. Available at: https://doi.org/10.1373/clinchem.2004.038026.

WHO (2017) *Cardiovascular diseases (CVDs)*, *Cardiovascular diseases (CVDs)*.

Wijers, M., Kuivenhoven, J.A. and van de Sluis, B. (2015) 'The life cycle of the low-density lipoprotein receptor: insights from cellular and in-vivo studies.', *Current opinion in lipidology*, 26(2), pp. 82–87. Available at: https://doi.org/10.1097/MOL.0000000000000157.

Willer, C.J. *et al.* (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.

Wright, M.L., Housman, D. and Taylor, J.Y. (2016) 'A perspective for sequencing familial hypercholesterolaemia in African Americans.', *NPJ genomic medicine*, 1, p. 16012. Available at: https://doi.org/10.1038/npjgenmed.2016.12.

Yazıcı, H. *et al.* (2024) 'Rapid lipid-lowering response in two cases of autosomal recessive hypercholesterolemia', *Journal of Clinical Lipidology* [Preprint]. Available at: https://doi.org/10.1016/j.jacl.2024.09.003.

Yoo, E.-G. (2016) 'Sitosterolemia: a review and update of pathophysiology, clinical spectrum, diagnosis, and management.', *Annals of pediatric endocrinology & metabolism*, 21(1), pp. 7–14. Available at: https://doi.org/10.6065/apem.2016.21.1.7.

Yu, L. *et al.* (2002) 'Disruption of Abcg5 and Abcg8 in mice reveals their crucial role in biliary cholesterol secretion.', *Proceedings of the National Academy of Sciences of the United States of America*, 99(25), pp. 16237–16242. Available at: https://doi.org/10.1073/pnas.252582399.

Chapter 2 (Manuscript 1): Targeted gene-panel sequencing reveals known and novel variants in Low Density Lipoprotein Cholesterol associated genes in continental African populations

Authors:

Natalie Smyth[1,2], Jian Wang[3], Adam D. McIntyre[3], Robert A. Hegele[3,4,5], Frederick Raal[6], Dhriti Sengupta[1], Michèle Ramsay[1]

Affiliations:

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

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

[3] Robarts Research Institute, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada.

[4] Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada

[5] Department of Medicine, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada.

[6] Division of Endocrinology & Metabolism, Faculty of Health Sciences, University of the Witwatersrand, South Africa

Corresponding author: Michèle Ramsay (michele.ramsay@wits.ac.za; +2711 717 6635)

Funding Sources: National Research Foundation, Research Chair in Bioinformatics or African Populations, grant number 89646. The AWI-Gen Study was funded by the National Institutes of Health (NIH) under award number U54HG006938 and its supplements, as part of the H3Africa Consortium.

2.1. Abstract

Cardiovascular diseases (CVD) are rapidly increasing in prevalence across Africa, with dyslipidaemia, particularly elevated low-density lipoprotein cholesterol (LDL-C), being a significant risk factor. However, there is limited understanding of the genetics of lipid metabolism in African populations. This study, nested within the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen), aims to investigate the genetic variants associated with high and low LDL-C levels in 191 individuals from West, East, and South Africa. Using a targeted next-generation sequencing (NGS) panel, we focused on 16 genes previously associated with the LDL-C metabolism, including *LDLR*, *APOB*, and *PCSK9*. We identified 356 unique variants, including both known pathogenic variants and novel variants predicted to affect protein function. The majority of variants were classified as variants of uncertain significance (VUS), with notable differences in variant distribution across genes, and 28 were known variants associated with phenotypic effects in ClinVar or the Human Gene Mutation Database. Although many variants had high in silico pathogenicity scores, their individual contribution to LDL-C levels was challenging to assess due to epistatic effects and the lack of family studies. Our findings highlight the genetic complexity of LDL-C disorders in African populations and highlight the need for functional studies, family-based screening, and improved classification of genetic variants. This is the first study to use targeted gene-panel sequencing to investigate dyslipidaemia in an African population and emphasizes the importance of considering population-specific genetic factors in CVD risk assessment and management.

Keywords: Low-Density Lipoprotein Cholesterol, Familial Hypercholesterolaemia, DNA Sequencing, variant interpretation, Africa

1.2. Introduction:

The prevalence of cardiovascular diseases (CVD) is rapidly rising throughout Africa and is the leading cause of non-communicable diseases and mortality worldwide (WHO, 2017). Lack of early diagnosis and inadequate access to health care is linked to at least 75% of CVD related deaths in Africa and other low and middle income regions ('Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report.', 2002). Elevated low-density lipoprotein cholesterol (LDL-C) levels have been shown to be a primary causal factor for CVD (Ference *et al.*, 2017).

Dyslipidaemia is defined as abnormal (either elevated or decreased) plasma lipid levels (Braun *et al.*, 2016). According to the World Health Organization, standard healthy LDL-C levels range from 0.5 to 3.0 mmol/L. These levels are regulated by a complex network of genetic and non-genetic factors, classifying dyslipidaemia as a multifactorial disorder. Rare genetic variants with large phenotypic effects cause monogenic forms of dyslipidaemia, however, the majority of cases with dyslipidaemia are the result of many common genetic variants, each with small to moderate phenotypic effects, that cumulatively underlie polygenic forms of the disorder (Dron and Hegele, 2018). Mendelian or monogenic disorders that result in elevated LDL-C levels are caused by a single highly penetrant variant in an individual or family, making the identification of a causal variant less complex than unravelling the intricate genetic architecture of multifactorial disorders (Antonarakis and Beckmann, 2006). Mendelian forms of dyslipidaemia are characterised by locus heterogeneity, with different individuals having different genes mutated with different effects on their lipid profiles. These disorders can be autosomal dominant, co-dominant or recessive traits. There are approximately 27 monogenic disorders of lipid metabolism caused by rare variants in 25 different genes affecting lipid metabolism (Rahalkar and Hegele, 2008).

Familial hypercholesterolaemia (FH), characterized by an elevated blood serum LDL-C, is one of the most common monogenic diseases in the world (Dron and Hegele, 2018). Worldwide, the majority of pathogenic FH mutations are found in three genes, namely those encoding the low-density lipoprotein receptor (*LDLR*), apolipoprotein B (*APOB*) and proprotein convertase subtilisin/kexin type 9 *(PCSK9*). In most populations, more than 90% are attributable to mutations in the *LDLR* gene (Leigh *et al.*, 2017).

The estimated world prevalence of FH is approximately 1 in 300, but the assumption of a relatively even distribution across all populations has not considered adequate data form many geographic regions to assess inter- and intra-continental heterogeneity (Beheshti *et al.*, 2020). Many countries, especially in poorer regions like Africa, do not have accurate FH registries and criteria for FH diagnosis may differ from country to country. Studies among African Americans have demonstrated that in spite of exhibiting healthier lipid profiles (characterized by lower average LDL-C, higher average high-density lipoprotein cholesterol (HDL-C) and lower triglyceride levels) compared to non-African ancestry individuals, they have an increased risk of CVD mortality. These interethnic differences can be attributed to differences in genetic architecture alongside the impact of physiological and environmental factors (Bentley and Rotimi, 2012).

There is a paucity of data and research regarding the genetics of LDL-C levels in Africa, presenting an opportunity to delve into this field and to understand the genetic contributions that may affect LDL-C levels in African populations. This study is nested in the Africa Wits-INDEPTH partnership for Genomic studies (AWI-Gen), leveraging infrastructure, base-line data, an experienced team of investigators and field workers, and stored DNA samples from over 12 000 individuals (Ramsay *et al.*, 2016). A targeted NGS gene panel of 16 genes previously associated with Mendelian forms of dyslipidaemia, including *LDLR, APOB* and *PCSK9*, was developed to investigate genetic variation in genes involved in LDL-C levels. The aim was to identify functional variants in the exons and exon/intron junctions of each gene, and to assess their association with high or low LDL-cholesterol levels in individuals. The assessment of a broad spectrum of genes that potentially underlie variation in lipid profiles, including both pathogenic and protective variants, could contribute to the understanding of clinical questions about penetrance and prognosis of dyslipidaemias in African populations.

2.3. Methods and Materials

2.3.1. AWI-Gen Cohort:

Participants with extreme LDL-C levels were selected from the H3Africa AWI-Gen Collaborative Center (Africa Wits-INDEPTH Partnership for Genomic studies). The AWI-Gen research platform is a population cross-sectional study that includes over 12 000 participants (males and females; predominantly 40 to 60 years) from Burkina Faso (Nanoro), Ghana (Navrongo), Kenya (Nairobi) and South Africa (Agincourt, Dikgale/DIMAMO and Soweto). The study was established to investigate genomic and environmental factors that influence body composition, body fat distribution and cardiometabolic disease (CMD) risk. Biospecimens, including fasting blood samples were collected from each participant and biochemical tests were performed, including serum fasting lipid profiles (total cholesterol, triglycerides, measured and calculated LDL-cholesterol and HDL-cholesterol). Comprehensive questionnaires were administered alongside the collection of anthropometric measurements, yielding extensive data on demographics, health history, environmental exposures, behaviours, and biomarkers (Ramsay *et al.*, 2016; Ali *et al.*, 2018) which was leveraged when selecting the participants for our research. The AWI-Gen study was approved by the Human Research Ethics Committee (HREC) (Medical) of the University of the Witwatersrand (Wits) and renewed in 2017 and 2022 (protocol numbers M121029, M16021, M170880, M2210108).

2.3.2. Participant selection:

The participants were selected from a population cohort (AWI-Gen) of mid-life adults (40 to 60 years) based on their LDL-C levels, and individuals with high LDL-C were not clinically diagnosed with FH. Participants were selected on the basis of having high or low measured LDL-C levels at baseline data collection. We identified and selected 10 participants with the highest and 10 participants with the lowest fasting LDL-C levels from each of the six study centres. An additional 36 participants with the highest and 36 with the lowest fasting LDL-C, irrespective of centres, were also selected. A total of 192 participants were chosen for sequencing; 96 with high LDL-C levels and 96 with low LDL-C levels. Demographic data are shown in Table 1.

The following exclusion criteria were applied to enhance the chance of detecting genetic causes of high and low LDL-C by excluding non-genetic causes that could affect serum lipid levels: HIV positive status; hypothyroidism, diabetes mellitus and/or calculated BMI > 40 kg/m^2; problematic alcohol use; high triglyceride levels (>4.5 mmol/L) (as FH patients rarely have high triglyceride levels); and participants on treatment for high cholesterol.

2.3.3. High-throughput targeted gene Sequencing Panel:

A targeted gene panel was designed, capturing exonic regions of 16 genes (Supplementary Table 1) known to be involved in lipid metabolism, namely, *LDLR*, *APOB*, *PCSK9*, *STAP1*, *APOE*, *LDLRAP1*, *LIPA*, *ABCG5*, *ABCG8*, *SORT1*, *MYLIP*, *NPC1L1*, *MTTP*, *SAR1B*, *ANGPTL3* and *GATB*. Probes were designed to capture regions of interest (exons and 50bp flanking regions) with a total size of 41 815 bp and 99.7% coverage of all sequences. Library preparation was performed using the Nextera Flex for Enrichment Library Prep from Illumina according to manufacturer's instructions. Pooled libraries were loaded on to a standard flow-cell on the Illumina iSeq sequencer using 2 × 101 bp paired-end chemistry according to the manufacturer's instructions. A 1% PhiX control was spiked in as a positive control to monitor sequencing performance. Sequencing quality was assessed using the Illumina Sequencing Analysis Viewer. FASTQ files were then downloaded and processed individually for each sample. All runs had a minimum of 400x coverage with a %Q30 above 90% per read.

2.3.4. Data Analysis

2.3.4.1.Sequence alignment and variant calling:

The Local Run Manager DNA Enrichment analysis module of Illumina was utilized for read alignment and variant calling. Burrow-Wheeler Aligner (BWA-MEM) (Li and Durbin, 2009)

was used to map reads to the complete human genome reference GRCh37/hg19 build. Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010) was used to call raw variants for each sample, analyze variants against known variants, and then calculate a false discovery rate for each variant. PCR duplicates were flagged in the BAM files and not used for variant calling. Regions containing indels were locally realigned to minimize the number of mismatches and generate VCF files for subsequent filtering.

2.3.4.2.Variant Annotation, Filtering and Classification:

Variants were annotated using both ANNOVAR (Wang, Li and Hakonarson, 2010) and VEP(McLaren *et al.*, 2016). Resulting VCFs were compared to ensure concordance. Synonymous and intronic variants were excluded from further analysis, unless they were considered to be splice site mutations. We then removed variants with a CADD score (Rentzsch *et al.*, 2019) lower than 14 in order to exclude variants that have overall low pathogenicity scores and are unlikely to have a significant functional effect on the protein. Next, common variants were filtered out by selecting those with a minor allele frequency (MAF) greater that 0.05 in the gnomAD database African-ancestry sub-set (Chen *et al.*, 2024). Remaining variants were classified into two categories utilizing an amended classification system initially developed by Haas *et al*. (Haas *et al.*, 2015) and further modified by Pan-Lizcano et al. (Pan-Lizcano *et al.*, 2022) in order to determine the likelihood of the variant being phenotype altering (Figure 1). The first category (I) included variants that passed filtering and have either been described in literature (HGMD) (Stenson *et al.*, 2020) or have been reported in ClinVar to be disease-associated. The subcategory (Ib) lists variants that are known to be associated with either high or low LDL-C levels. The second category (II) included the remaining variants that have not previously been associated with pathology in literature. The subcategory (IIb) variants were selected according to prediction scores. Three in silico prediction tools were utilized to assign functional impact, namely SIFT (https://sift.bii.a-star.edu.sg/), PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/) and MutationTaster (https://www.mutationtaster.org/). Variants that had 2 or more pathogenic predictions for the 3 tools were considered potential function altering mutations.

The classification of variants in this study followed the guidelines established by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) (Richards *et al*., 2015). These guidelines provide a standardized framework for evaluating the clinical significance of variants based on multiple criteria, including population frequency, *in silico* predictions, functional studies, segregation data, and previously

reported clinical associations. Variants are classified into five different categories: pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, and benign.

2.3.4.3.Validation:

Validation of 11 selected potentially functional variants was done using Sanger Sequencing through Inqaba Biotechnical Industries (Pretoria, South Africa). Variants were selected as they were identified as being potentially novel based on the outcome of the analysis of the sequence data..

2.3.4.4.Copy Number Variant Analysis:

FASTQ files from 96 High LDL and 94 Low LDL samples were used for variant calling (output VCF files) and targeted region coverage statistics (output BAM files) using CLC Bio genomics workbench (CLC bio version 21, QIAGEN Digital Insights). VCF and BAM files were subsequently imported into VarSeq CNV Caller (VarSeq v2.2.4, Golden Helix, Bozeman, MT) for copy number variation (CNV) analysis.

Depth of coverage (DOC) of targeted regions was normalized and compared between samples of interest and a group of references without CNV. Ratio was calculated as the sample DOC divided by the mean DOC of the references. Z-score is standard deviation of sample DOC from the mean DOC of the references. CNV state is assigned based on ratio and z-score on each targeted region: Diploid (normal) state: ratio = 1, z-score = 0; Heterozygous deletion: ratio = 0.5, z-score $\leq$ 5; Homozygous deletion: ratio = 0, z-score $\leq$ 5; Duplication: ratio $\geq$ 1.5, z-score $\geq$ 5.

There is no confirmed "normal or positive control" for this 16 gene targeted panel for this population and therefore the reference group was generated by adding all samples to the reference group in multiple rounds of CNV analysis and removing samples with suspected CNV or possible change of DOC. Confirmation tests were performed when increasing number of references and excluding samples being tested from the references. Samples with confirmed CNV also served as positive controls in multiple confirmation analysis.

2.4. Results:

2.4.1. Characteristics of the study participants

The participants were divided into two groups. The first group was selected on the basis of having high levels of LDL-C. It included a total of 96 individuals selected from the six different study sites with an average age of 51 years, and average BMI of 26.8 kg/m^2 and an LDL-C

level of 6.31 mmol/L. The second group includes 95 individuals with low LDL-C levels, also selected from the six different study sites and average age was 49.4 years, average BMI was 22.3 kg/m^2 and average LDL-C level was 0.52 mmol/L. The demographic data of the selected participants in the two groups is displayed in Table 1.

Table 1: Characteristics of participants with high and low levels of LDL-C across six sites

	DIMAMO (South Africa)	Agincourt (South Africa)	Soweto (South Africa)	Nanoro (Burkina Faso)	Navrongo (Ghana)	Nairobi (Kenya)	All Sites
High LDL-C							
n	13	19	14	20	10	20	96
Age (yr)	54.5 ± 7.5	56.9 ± 7.0	48.6 ± 6.5	46.5 ± 5.2	53.3 ± 6.1	48.3 ± 5.1	51 ± 7.2
Sex (% male)	0.8	68.4	92.9	100	70	60	68.8
BMI (kg/m^2)	28.8 ± 5.1	29.4 ± 4.3	25 ± 5.0	26.4 ± 4.1	21.5 ± 3.6	27.1± 4.4	26.8 ± 4.9*
Total Cholesterol (mmol/L)	6.11 ± 1.55	5.99 ± 0.9	6.48 ± 0.99	6.02 ± 0.68	5.43 ± 0.57	6.32 ± 0.9	6.09 ± 0.98
HDL-C (mmol/L)	1.32 ± 0.36	1.19 ± 0.21	1.48 ± 0.31	1.16 ± 0.21	1.34 ± 0.21	1.45 ± 0.54	1.31 ± 0.36
LDL-C (mmol/L)	6.19 ± 0.45	6.15 ± 0.58	6.21 ± 0.52	5.81 ± 0.57	5.35 ± 0.54	6.6 ± 0.61	6.31 ± 0.68
Triglycerides (mmol/L)	1.78 ± 0.91	1.14 ± 0.6	1.28 ± 0.52	1.25 ± 0.85	0.89 ± 0.42	1.62 ± 0.85	1.34 ± 0.77
Low LDL-C							
n	20	17	10	17	19	12	95
Age (yr)	49.6 ± 8.4	49.4 ± 7.3	48.4 ± 7.5	47.8 ± 5.7	49.9 ± 5.9	51.3 ± 7.2	49.4 ± 6.9

Sex, (% male)	50	23.5	90	41.2	68.4	75	54.7
BMI (kg/m^2)	23.9 ± 5.1	24.6 ± 5.3	23.4 ± 5.3	19.3 ± 2.2	20.2 ± 2.3	23.3 ± 4.8	22.3 ± 0.46*
Total Cholesterol (mmol/L)	2.27 ± 0.91	2.67 ± 0.52	2.73 ± 0.56	1.78 ± 0.43	1.76 ± 0.89	2.34 ± 0.56	2.21 ± 0.78
HDL-C (mmol/L)	1.02 ± 0.82	1.14 ± 0.35	1.05 ± 0.53	0.92 ± 0.48	0.78 ± 0.46	1.14 ± 0.56	1 ± 0.53
LDL-C (mmol/L)	0.54 ± 0.09	0.63 ± 0.09	0.57 ± 0.1	0.46 ± 0.05	0.41 ± 0.01	0.62 ± 0.12	0.52 ± 0.11
Triglycerides (mmol/L)	0.7 ± 0.46	0.52 ± 0.16	0.66 ± 0.27	0.58 ± 0.35	0.4 ± 0.26	0.99 ± 0.71	0.62 ± 0.42

Values are means ± SD except sex

*Statistically significant difference (p-value < 0.05) in BMI between high and low groups (Wilcoxon rank-sum test)

2.4.2. Classification of Variants

Overall, 938 variants were detected in 191 participants with an average of 5 variants per participant. A total of 356 unique variants were identified in the exonic regions and the 50bp flanking regions of genes involved in the LDL-C metabolism. Nonsynonymous variants were most common (55%), followed by synonymous variants (38%) and the remaining variants were either stopgain (>1%), startloss (>1%), nonframeshift insertions (>1%), nonframeshift deletion (>1%) or frameshift deletion variants (>1%).

After filtering according to CADD score and minor allele frequency (MAF) as described in the methods, 81 unique variants remained (Figure 1). From the 81 variants, 28 were grouped into category I, Table 2 lists these along with their associated phenotype. Of the 28 variants, 19 are associated with extreme LDL-C levels (category Ib), and all were identified in *LDLR*, *APOB* or *PCSK9*.

The remaining 53 variants were grouped into category II. The potential impact on the protein was evaluated using in silico prediction tools and 36 variants were classified as potentially protein altering and thus grouped into category IIb (Table 3).

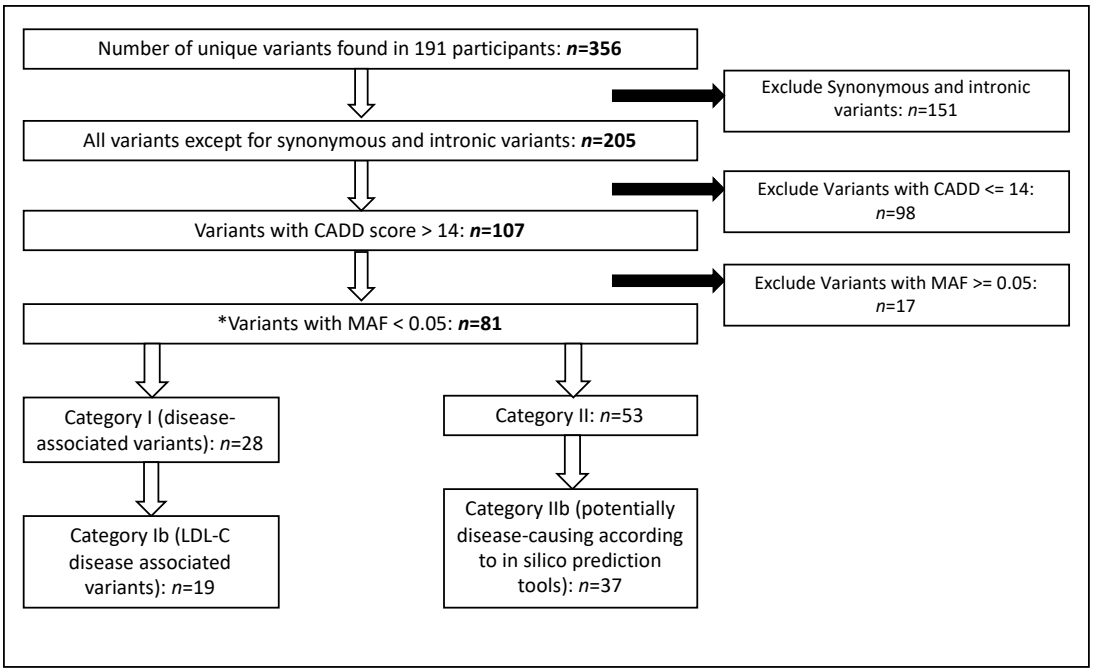


Figure 1. Stepwise filtering and classification of variants, according the amended classification system initially developed by Haas et al. (Haas et al., 2015) and further modified by Pan-Lizcano et al. (Pan-Lizcano et al., 2022), identified in participants with high or low levels of LDL-C.

Of the total 81 variants, 41 were found in *LDLR*, *APOB* and *PCSK9*. The numbers of variants per phenotypic group in each gene are shown in Figure 2. Over 42% of the variants were detected in individuals from both groups, 21% only in individuals with low LDL-C and 35% with high LDL-C. *APOB* presented with the highest total number of variants (23) followed by *PCSK9* (12), *ABCG5* encoding TP-binding cassette sub-family G member 5 (9) and *NPC1L1* encoding Niemann-Pick C1-like 1 protein (7). Variants in *APOB* and *PCSK9* were detected in both phenotypic groups. Pathogenic variants in LDLR were all found in the high LDL-C group and one variant (Benign) was detected in both groups.

These 81 variants were assigned to three categories: variants detected exclusively found in the group of participants with elevated levels of LDL-C (35.8%); variants detected exclusively in the low LDL-C group (22.0%); and variants present in both high LDL-C and low LDL-C groups (42.2%). The number of variants per category in each gene is shown in Figure 2. *APOB* presented with the highest number of variants (*n*=23) followed by *PCSK9* (*n*=12), the variants detected in these two genes were distributed across all three categories. All the six variants detected in *LDLR* are grouped in Catagory1b. Variants in *GATB, ABCG8, NPC1L1, STAP1*

and *LDLRAP1* also showed a similar profile i.e. absent in the exclusive low LDL-C group. The variants detected in *ANGPTL3* on the other hand were detected in low LDL-C group only.

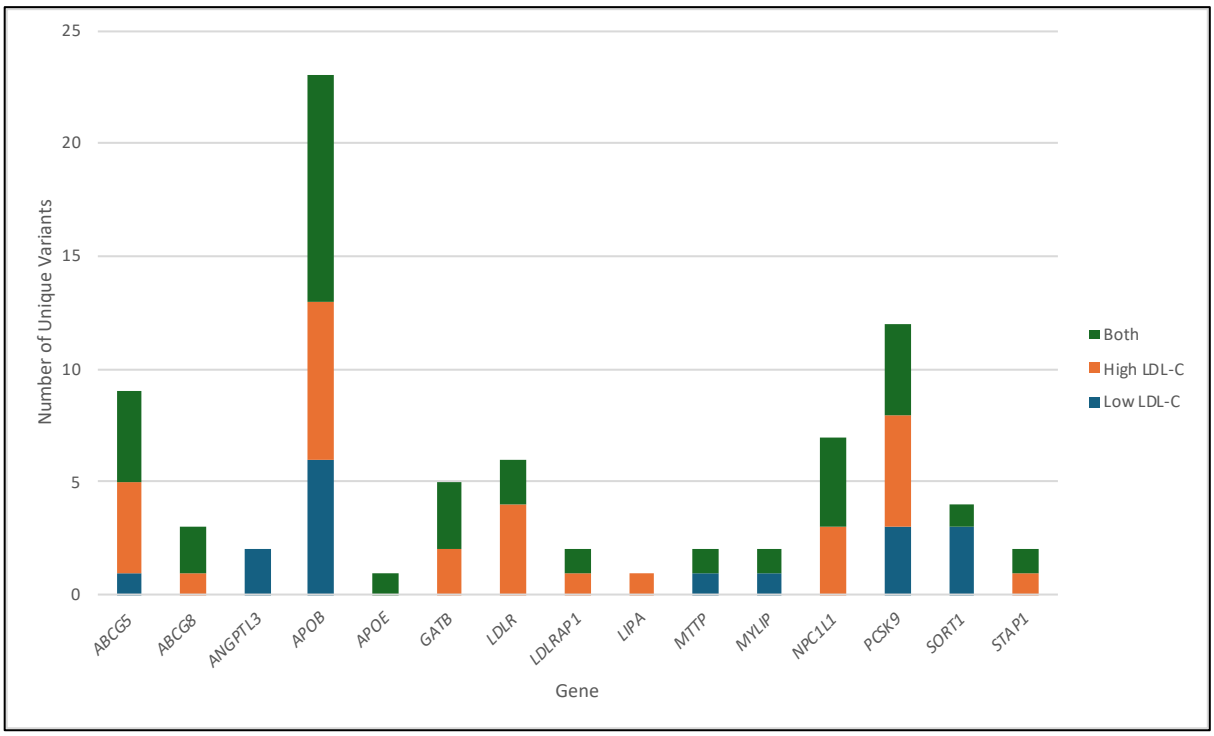


Figure 2. Number of variants per gene after filtering separated according to phenotype group.

According the ACMG guidelines, only 6 variants in category I are classified as 'pathogenic'. Twelve are classified as 'variant of unknown/uncertain significance' (VUS) and the remaining 10 are 'benign' (Table 2). In the category IIb group, only 3 variants are classified as pathogenic whereas 26 are VUS and 8 are benign (Table 3). Only 19 individuals out of the 191 presented with potentially pathogenic variants in the 3 main genes, namely *LDLR*, *APOB* and *PCSK9*. Figure 3 depicts the numbers of each classification of variants separated by phenotypic group. More than 50% of the variants are classified as VUS. Four of the 'pathogenic' variants were detected in the group with high LDL-C levels. Three pathogenic variants were detected in *LDLR* (p.F202S, p.G343S and p.V797M) that have known associations to hypercholesterolemia. We also detected one pathogenic (Category IIb) variant in *ANGPTL3* (Frameshift Deletion p.Q192Rfs*4) in an individual with a low LDL-C. Approximately 50% of the variants detected in *APOB* are classified as VUS.

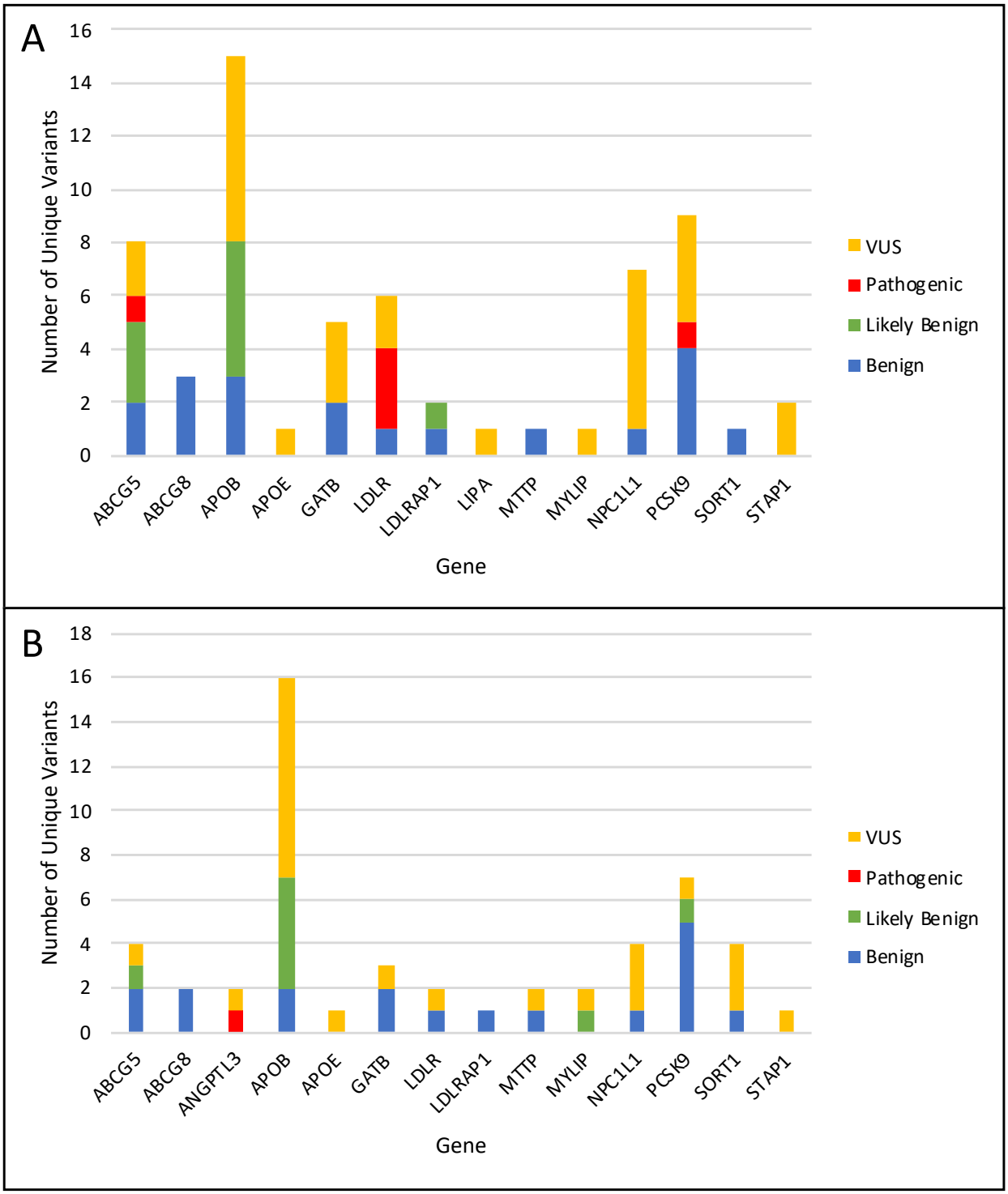


Figure 3: Number variants per gene shown by ACMG classification among Categories I and IIb in the high group (A) and low group (B).

Two variants detected in *PCSK9*, namely, p.Y142X and p.C679X are both potentially loss of function variants classified as benign despite high CADD scores (37 and 38 respectively). Table 4 lists the Minor Allele Frequencies from different databases and ethnic groups showing that both variants are more common in African populations. A single deletion of exons 6-10

was identified in *NPC1L1* in an individual with low LDL-C. A duplication of the 5'UTR and exon 1 regions in *MYLIP* was detected in an individual with high LDL-C. We also detected four potentially novel missense variants in two genes-*APOB* (p.M566T and p.G852E) and *NPC1L1* (p.G1255D and p.A1259T) that were classified as VUS by the in silico prediction tools. These novel variants were further validated using Sanger Sequencing for technical validation and quality control (Table 5). All selected variants were successfully validated.

No CNVs were detected in the three major genes: *LDLR*, *PCSK9* and *APOB*.

Table 2. Variants with described pathology in ClinVar or HGMD (Category I) identified in the gene panel

Gene	Position (GRCh37/hg19 build)	dbSNP Code	Nucleotide Change	Amino Acid Change	High n^{Ψ}	Low n^{Ψ}	ACMG Classification	Reference	Described Pathology in HGMD and ClinVar
ABCG5	2:44059195	rs145164937	c.C293G	p.A98G	1	0	Likely Benign	(Ali *et al.*, 2016)	Macrothrombocytopaenia
ABCG5	2:44053568	rs119479066	c.C727T	p.R243X	1	0	Pathogenic	(Lee *et al.*, 2001)	Sitosterolaemia
ABCG5	2:44041634	rs144973796	c.G1744A	p.G582R	1	0	Likely Benign		Sitosterolaemia*
ABCG5	2:44058978	rs376797531	c.T431C	p.V144A	0	1	VUS		Sitosterolaemia*
ABCG8	2:44078891	rs74358901	c.G491A	p.R164Q	3	0	Benign		Sitosterolaemia*
ABCG8	2:44079559	rs9282574	c.G628A	p.V210M	15	7	Benign		Sitosterolaemia*
APOB	2:21255355	rs12714225**	c.T1223C	p.I408T	3	1	Benign	(Zeissig *et al.*, 2010)	Hypobetalipoproteinaemia
APOB	2:21255268	rs142114415**	c.G1310A	p.R437H	1	2	Likely Benign		Hypercholesterolaemia
APOB	2:21252534	rs13306194	c.C1594T	p.R532W	2	0	Benign	(Lye *et al.*, 2013)	Increased HDL cholesterol levels, in hypercholesterolaemia
APOB	2:21237450	rs72653078**	c.C3712A	p.L1238I	1	2	Likely Benign		Hypercholesterolaemia / Hypobetalipoproteinaemia*
APOB	2:21237422	rs61741164**	c.A3740G	p.Y1247C	1	0	VUS	(Blanco-Vaca *et al.*, 2019)	Hypobetalipoproteinaemia
APOB	2:21236070	rs143282164**	c.C4178T	p.A1393V	0	2	Likely Benign		Hypercholesterolaemia / Hypobetalipoproteinaemia*
APOB	2:21234674	rs151009667	c.G5066A	p.R1689H	1	0	Likely Benign	(Johansen *et al.*, 2010)	Hypertriglyceridaemia
APOB	2:21232845	rs12713681**	c.G6895C	p.D2299H	5	2	Benign	(Zeissig *et al.*, 2010)	Hypobetalipoproteinaemia
APOB	2:21228960	rs61744288**	c.T10780C	p.W3594R	0	3	Likely Benign		Hypercholesterolaemia
APOE	19:45412040	rs769455	c.C487T	p.R163C	2	6	VUS	(Emi *et al.*, 1988)	Apolipoprotein E deficiency

LDLR	19:11216187	rs879254590**	c.T605C	p.F202S	1	0	Pathogenic	(Fan *et al.*, 2015)	Hypercholesterolaemia
LDLR	19:11221357	rs72658860**	c.G970A	p.G324S	3	4	Benign	(Arca and Jokinen, 1998)	Hypercholesterolaemia
LDLR	19:11221414	rs730882096**	c.G1027A	p.G343S	1	0	Pathogenic	(Hobbs, Brown and Goldstein, 1992)	Hypercholesterolaemia
LDLR	19:11224398	rs141673997**	c.G1546A	p.G516S	1	0	VUS	(Hooper *et al.*, 2012)	Hypercholesterolaemia
LDLR	19:11238761	rs750518671**	c.G2389A	p.V797M	1	0	Pathogenic	(Pereira *et al.*, 1995)	Hypercholesterolaemia
LDLR	19:11240240	rs5928**	c.G2441A	p.R814Q	3	1	VUS	(Thiart *et al.*, 2000)	Hypercholesterolaemia
PCSK9	1:55505668	rs11583680**	c.C158T	p.A53V	2	1	Benign		Association with LDL-cholesterol levels
PCSK9	1:55512222	rs67608943**	c.C426G	p.Y142X	0	1	Likely Benign	(Cohen *et al.*, 2005)	Low LDL cholesterol
PCSK9	1:55518043	rs753857795**	c.G616A	p.E206K	1	0	VUS		Hypercholesterolaemia*
PCSK9	1:55521701	rs72646509**	c.C835A	p.P279T	2	1	Benign		Hypercholesterolaemia*
PCSK9	1:55524222	rs141502002**	c.C1405T	p.R469W	4	3	Benign	(Pisciotta *et al.*, 2006)	Hypercholesterolaemia
PCSK9	1:55529215	rs28362286**	c.C2037A	p.C679X	2	14	Benign	(Cohen *et al.*, 2005)	Low LDL cholesterol

*Reported in ClinVar, but not in HGMD

**Category Ib, variants known to be associated with extreme LDL-C levels.

ψ Number of Individuals in either the high LDL-C group or low LDL-C group

Table 3. Potential functional mutations predicted using in silico prediction tools (Category IIb) identified in the gene panel

Gene	Position (GRCh37/hg 19 build)	dbSNP Code	Nucleotide Change	Type	Amino Acid Change	SIFT	Polyphen2 HVAR	Mutation Taster	CADD	GERP++ Score	ACMG Classification	High *n*	Low *n*
ABCG5	2:44065680	rs72542426	c.G139T	Missense	p.V47F	D (0.003)	P (0.48)	D (0.810)	23.5	-0.287	Likely Benign	4	5
ABCG5	2:44051148	rs115590670	c.A1228C	Missense	p.N410H	T (0.124)	D (0.924)	D (0.392)	22.4	3.05	Benign	4	4
ABCG5	2:44047117	rs986503686	c.T1586C	Missense	p.I529T	D (0.018)	B (0.314)	D (0.504)	26.3	4.79	VUS	1	0
ABCG8	2:44071736	rs142250628	c.C154G	Missense	p.L52V	T (0.17)	D (0.957)	D (0.588)	24.2	5.27	Benign	1	2
ANGPTL3	1:63064446	rs577189349	c.575delA	Frameshift Deletion	p.Q192Rfs*4				.	.	Pathogenic (Familial Hyperbetalipoproteinaemia)	0	1
APOB	2:21251331	-	c.T1697C*	Missense	p.M566T	D (0.001)	D (0.955)	D (0.548)	25.3	5.69	VUS	1	0
APOB	2:21250792	rs1026800482	c.C1975T	Missense	p.L659F	D (0.006)	D (0.919)	D (0.416)	27.2	4.85	VUS	1	0
APOB	2:21246446	-	c.G2555A*	Missense	p.G852E	D (0.0)	D (1.0)	D (0.810)	26.1	5.35	VUS	0	1
APOB	2:21245884	rs564030306	c.G2635A	Missense	p.V879M	T (0.084)	P (0.626)	D (0.385)	23.1	4.64	VUS	0	1
APOB	2:21238368	rs72653076	c.C3382T	Missense	p.R1128C	D (0.004)	P (0.877)	D (0.472)	34	5.43	VUS	1	0
APOB	2:21236404	rs781366272	c.G3844A	Missense	p.D1282N	T (0.096)	P (0.518)	D (0.530)	25.3	5.29	VUS	1	1
APOB	2:21236088	rs1389459768	c.C4160A	Missense	p.A1387D	D (0.001)	P (0.873)	N (0.090)	23.5	4.43	VUS	0	1
APOB	2:21235365	rs140877474	c.A4375G	Missense	p.S1459G	D (0.001)	P (0.449)	D (0.366)	23.4	5.62	Likely Benign	1	0
APOB	2:21231095	rs113399155	c.C8645T	Missense	p.T2882I	D (0.01)	P (0.564)	N (0.090)	18.54	4.62	VUS	3	3

APOB	2:21229803	rs146687604	c.C9937A	Missense	p.L3313I	D (0.032)	P (0.566)	N (0.191)	19.27	2.16	Likely Benign	1	1
APOB	2:21228479	rs139599466	c.C11261T	Missense	p.T3754I	D (0.002)	P (0.834)	N (0.314)	25.5	4.29	VUS	0	1
APOB	2:21224624	rs146687572	c.G13670C	Missense	p.G4557A	D (0.005)	P (0.73)	D (0.345)	22.8	2.97	VUS	1	1
GATB	4:152638226	rs116634238	c.G442T	Missense	p.A148S	T (0.149)	D (0.966)	D (0.810)	25.0	5.84	VUS	1	0
GATB	4:152638153	rs114812418	c.G515T	Missense	p.G172V	D (0.025)	D (0.968)	D (0.810)	23.7	6.08	VUS	2	0
GATB	4:152629221	rs140995129	c.G796A	Missense	p.V266M	D (0.002)	D (0.979)	D (0.810)	34	5.71	VUS	1	1
GATB	4:152626418	rs79384184	c.A881G	Missense	p.Y294C	D (0.003)	P (0.819)	D (0.588)	26.0	2.77	Benign	3	2
LDLRAP 1	1:25890161	rs141522360	c.C626T	Missense	p.T209I	D (0.021)	B (0.408)	D (0.548)	27.7	5.14	Likely Benign	2	0
LIPA	10:90988153	rs903291028	c.C232A	Missense	p.P78T	D (0.041)	B (0.046)	D (0.422)	23.0	4.57	VUS	1	0
MTTP	4:100496109	rs371979566	c.T43C	Missense	p.Y15H	T (0.097)	D (0.942)	D (0.810)	18.24	5.85	VUS	0	1
MTTP	4:100512442	rs144315111	c.A552T	Missense	p.K184N	D (0.013)	B (0.052)	D (0.365)	23.3	3.56	Benign	6	2
NPC1L1	7:44579814	rs114102795	c.C182T	Missense	p.T61M	D (0.008)	P (0.46)	D (0.323)	23.9	3.95	VUS	1	1
NPC1L1	7:44579394	rs145094484	c.G602A	Missense	p.R201H	D (0.029)	P (0.867)	D (0.588)	24.2	4.15	VUS	1	0
NPC1L1	7:44555515	-	c.G3764A *	Missense	p.G1255D	D (0.014)	P (0.836)	D (0.491)	24.6	3.61	VUS	1	0
NPC1L1	7:44555504	-	c.G3775A *	Missense	p.A1259T	D (0.041)	B (0.016)	D (0.445)	27.7	4.6	VUS	1	0
NPC1L1	7:44556834	rs79519744	c.G3340C	Missense	p.D1114H	T (0.065)	D (0.935)	D (0.363)	28.8	5.33	VUS	3	8
PCSK9	1:55518079	rs1278078294	c.A652G	Missense	p.R218G	T (0.103)	P (0.778)	D (0.810)	24.1	2.89	Pathogenic (FH)	1	0
PCSK9	1:55518422	rs72646508	c.C757T	Missense	p.L253F	D (0.0)	D (0.999)	D (0.810)	29.3	4.02	Benign	0	1
PCSK9	1:55524214	rs72646517	c.G1397A	Missense	p.G466E	D (0.011)	D (0.96)	D (0.362)	22.0	3.56	VUS	1	0

SORT1	1:109940304	-	c.C211T	Missense	p.R71C	D (0.025)	P (0.231)	D (0.333)	27.4	1.79	VUS	0	1
SORT1	1:109940295	-	c.C220T	Missense	p.R74C	D (0.009)	P (0.832)	D (0.497)	33	3.9	VUS	0	1
STAP1	4:68424595	rs376738460	c.C68T	Missense	p.A23V	T (0.006)	B (0.076)	D (0.388)	24.8	5.36	VUS	1	0
STAP1	4:68456649	-	c.A707T	Missense	p.Y236F	T (0.119)	D (0.998)	D (0.418)	23.8	5.56	VUS	1	1

*Novel

Table 4: Minor Allele Frequencies of the two loss-of-function *PCSK9* variants across the databases

Database	p.Y142X	p.C679X
1000 Genomes Worldwide	0.0014	0.0022
1000 Genomes African	0.0045	0.0083
gnomAD Exome Worldwide	0.0002	0.0006
gnomAD Exome African	0.0027	0.0084
gnomAD Exome Admixed America	8.933e-05	0.0002
gnomAD Exome Ashkenazi Jewish	0	0
gnomAD Exome East Asian	0	0
gnomAD Exome Finnish	0	0
gnomAD Exome Non-Finish European	0	1.823e-05
gnomAD Exome Other	0	0.0004
gnomAD Exome South Asian	0	3.25e-05
gnomAD Genome All	0.0005	0.0023
gnomAD Genome African	0.0017	0.0080
gnomAD Genome Admixed American	0	0
gnomAD Genome Ashkenazi Jewish	0	0
gnomAD Genome East Asian	0	0
gnomAD Genome Finnish	0	0

gnomAD Genome Non-Finish European	0	0
gnomAD Genome Other	0	0

Table 5: Variants selected for Sanger Sequencing Validation

Variant	Protein change	Exon	Gene	Number of participants	High or Low LDL-C Group
c.C1975T	p.L659F	14	*APOB*	1	High
c.G2555A	p.G852E	17	*APOB*	1	Low
c.C4160A	p.A1387D	25	*APOB*	1	Low
c.A12211G	p.K4071E	29	*APOB*	1	Low
c.T1697C	p.M566T	13	*APOB*	1	High
c.A652G	p.R218G	4	*PCSK9*	1	High
c.G3764A	p.G1255D	19	*NPC1L1*	1	High
c.G3775A	p.1259T	19	*NPC1L1*	1	High
c.T1586C	p.I529T	11	*ABCG5*	1	High
c.C232A	p.P78T	4	*LIPA*	1	High
c.A707T	p.Y236F	7	*STAP1*	2	High and Low

2.5. Discussion:

Africa is burdened with an increasing prevalence of CVDs and lack of data for the specific risk factors such as LDL-C levels and their contributing genetic factors. In this study we assessed the presence of potential functional variants in 16 genes involved in the LDL-C metabolism in 191 continental African participants from West, East and South Africa.

A substantial number of variants were detected, highlighting the unique genetic variability found in African populations. Disorders of lipid metabolism, including FH, have a significant heritable component, however the majority of studies focus on non-African populations (Bentley, Callier and Rotimi, 2020). The challenges of genetic studies in Africa include the variability among the populations and the paucity of data which results in ambiguous classification of variants. This highlights the gaps along with the challenge of classification of variants as the majority of SNPs detected in this study were classified as variants of uncertain significance. More than half of these variants present with relatively high *in silico* predicted pathogenicity scores (with CADD scores above 20) and low minor allele frequencies in both African and global populations. The majority of these variants occur in *APOB*. Unfortunately, it is difficult to assess whether variants are individually associated with high or low LDL-C, as the cholesterol levels are modulated by epistatic effects of several genes and environmental factors. There are a few instances where 'possible cause' can be proposed. p.R1689H in *APOB*, for example, occurs in an individual with high LDL-C, with no other known pathogenic variants, and could be the cause of the high LDL-C phenotype despite being classified as likely benign. However, the majority of variants observed occur in the presence of additional variants classified as VUS. Due to a scarcity of data on variants, it is also difficult to derive conclusions based on ACMG classifications. If prediction tools and minor allele frequencies infer a pathogenic or likely pathogenic outcome but according to ACMG are classified as benign or likely benign, is a variant excluded from being potentially disease-associated? This conundrum can be illustrated with the *PCSK9* variants, p.Y142X and p.C679X. Both variants were detected in this cohort, are classified as benign, but have been previously shown to be associated with low LDL-C levels. These variants are probably loss-of-function variants which result in increased levels of *LDLR*, and thus low LDL-C serum levels (Cohen *et al.*, 2005). Both are also more common in African populations compared to non-African populations (Table 4).

Individuals of African origin have on average about 1 million more variants, compared to the reference genome, than those of European or Asian origin (Auton *et al.*, 2015). It is therefore not surprising that we have observed many more variants per individual in this study and often

observed multiple potentially functional variants in a single individual. Since some variants are likely to be associated with high LDL-C levels and others with low levels, epistatic effects could result in a lipid profile in the normal range, despite the presence of a known pathogenic variant. In genes such as *PCSK9* and *APOB*, multiple variants with opposing effects of the abundance of LDL-C can be detected in a single individual. The extreme phenotype in the presence of only one of the variants can provide suggestive evidence as to which variant plays what role. The variant, p.C679X in *PCSK9*, usually associated with low LDL-C levels, was detected 16 times, 2 of which were present in individuals with high LDL-C. Of these two individuals one has a variant in *LDLR*, namely p.R814Q (ClinVar ID 161278) which was previously classified as pathogenic, however, recently the ClinGen Familial Hypercholesterolemia Variant Curation Expert Panel (FH VCEP) have reclassified it as VUS. The second individual also has a variant in *LDLR*, namely p.G324S (ClinVar ID 161263) which was previously classified as VUS but subsequently has been reclassified as benign by the FH VCEP. When investigating the unfiltered list of variants, it is notable that there is a 'benign' *APOB* variant present (p.P145S, rs6752026) with high MAF (>0.05) and high *in silico* pathogenicity scores that may have an effect on the phenotype. It may contribute to the multifactorial and polygenic determination of lipid levels and exert its effect on modulating the phenotype.

There were two potentially novel variants detected in *APOB* which have relatively high pathogenicity scores (c.T1697C and c.G2555A). The former occurs in an individual with high LDL-C levels with no other potential disease-causing variants. Family cascade screening is required to infer conclusions regarding novel variants but in this cohort we did not have information on family members.

Previous studies in populations of predominantly non-African ancestry have detected CNVs in approximately 10% of FH patients, predominantly in *LDLR* (Iacocca *et al.*, 2018), however no CNVs were detected in our study in the three major FH genes. A deletion in *NPC1L1* was detected. This protein transports dietary cholesterol from the gut lumen into intestinal enterocytes, and is expressed in both the liver and small intestine (Davis *et al.*, 2004). One of the most commonly prescribed drugs for lowering plasma levels of LDL-C, ezetimibe, inhibits *NPC1L1* and thus blocks sterol absorption (Garcia-Calvo *et al.*, 2005). Variants that inhibit NPC1L1 function has been associated with lower levels of LDL-C and thus a reduced risk of CAD. The individual carrying this alteration has low levels of LDL-C, and although there is

insufficient evidence that there is a direct link between the genotype and phenotype, the CNV may contribute to the lower levels of LDL-C. Functional studies would be required to confirm.

The identification of both known and potentially novel pathogenic variants represents an important step toward improving the understanding and clinical management of dyslipidaemia in African populations. Known variants can inform clinical care through early diagnosis, targeted treatment, and family cascade screening, contributing to the prevention of premature cardiovascular disease. The detection of novel variants, although requiring functional validation and segregation studies in families, provides valuable insights into population-specific genetic mechanisms influencing LDL-C regulation. Functional studies and family cascade screening are required to verify the pathogenicity of the novel variants, as well as the potentially functional variants in African populations. An accurate prevalence of FH needs to be calculated in different continental African regions, as evidence from this study suggests that it may be significantly lower than the worldwide prevalence. A larger African study including family members would be required to determine a more accurate FH prevalence. Our findings highlight the urgent need for more comprehensive genetic studies in African populations to better characterize the risk factors for CVDs and refine variant classification. This study has several limitations. They include the use of non-African reference panels for variant annotation and differences in allele frequencies and linkage disequilibrium patterns across populations, could potentially lead to misclassification of variants. Furthermore, the modest sample size and the imbalance in sex distribution within the cohort—where females were overrepresented—may have influenced lipid level distributions and limited the generalizability of findings. Despite these limitations, this study contributes to the growing body of knowledge on lipid metabolism genetics and emphasizes the importance of considering population-specific factors when assessing genetic risk for cardiovascular diseases.

Author Contributions

N.S. was involved in research conceptualisation, experimental laboratory work, data analysis and wrote the main manuscript text and created tables and figures. J.W., A.D.M. and R.H. assisted with panel design, performed the CNV analyses and assisted with writing of manuscript. D.S. assisted with data analyses and writing of manuscript. F.R. and M.R. were involved in conceptualisation and project design and assisted with writing of manuscript. All authors reviewed the manuscript.

Data Availability

AWI-Gen Phenotype data is available on EGA (EGAD00001006425) and raw sequence data will be uploaded on to EGA.

Ethics Declaration

Ethics was granted through the University of the Witwatersrand Human Research Ethics Council (protocol numbers M121029, M16021, M170880, M2210108)**.** All procedures were conducted in accordance with applicable guidelines and regulations. Informed consent was obtained from all participants prior to their enrolment in the study.

List of Figures

List of Tables:

2.6. References

Ali, S. *et al.* (2016) 'Congenital macrothrombocytopenia is a heterogeneous disorder in India.', *Haemophilia : the official journal of the World Federation of Hemophilia*, 22(4), pp. 570–582. Available at: https://doi.org/10.1111/hae.12917.

Ali, S.A. *et al.* (2018) 'Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study.', *Global health action*, 11(sup2), p. 1507133. Available at: https://doi.org/10.1080/16549716.2018.1507133.

Antonarakis, S.E. and Beckmann, J.S. (2006) 'Mendelian disorders deserve more attention.', *Nature reviews. Genetics*, 7(4), pp. 277–282. Available at: https://doi.org/10.1038/nrg1826.

Arca, M. and Jokinen, E. (1998) 'Low density lipoprotein receptor mutations in a selected population of individuals with moderate hypercholesterolemia.', *Atherosclerosis*, 136(1), pp. 187–194. Available at: https://doi.org/10.1016/s0021-9150(97)00210-4.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation.', *Nature*, 526(7571), pp. 68–74. Available at: https://doi.org/10.1038/nature15393.

Beheshti, S.O. *et al.* (2020) 'Worldwide Prevalence of Familial Hypercholesterolemia: Meta-Analyses of 11 Million Subjects.', *Journal of the American College of Cardiology*, 75(20), pp. 2553–2566. Available at: https://doi.org/10.1016/j.jacc.2020.03.057.

Bentley, A.R., Callier, S.L. and Rotimi, C.N. (2020) 'Evaluating the promise of inclusion of African ancestry populations in genomics.', *NPJ genomic medicine*, 5, p. 5. Available at: https://doi.org/10.1038/s41525-019-0111-x.

Bentley, A.R. and Rotimi, C.N. (2012) 'Interethnic Variation in Lipid Profiles: Implications for Underidentification of African-Americans at risk for Metabolic Disorders.', *Expert review of endocrinology & metabolism*, 7(6), pp. 659–667. Available at: https://doi.org/10.1586/eem.12.55.

Blanco-Vaca, F. *et al.* (2019) 'Molecular analysis of APOB, SAR1B, ANGPTL3, and MTTP in patients with primary hypocholesterolemia in a clinical laboratory setting: Evidence supporting polygenicity in mutation-negative patients.', *Atherosclerosis*, 283, pp. 52–60. Available at: https://doi.org/10.1016/j.atherosclerosis.2019.01.036.

Braun, K.V.E. *et al.* (2016) 'The role of DNA methylation in dyslipidaemia: A systematic review.', *Progress in lipid research*, 64, pp. 178–191. Available at: https://doi.org/10.1016/j.plipres.2016.10.002.

Chen, S. *et al.* (2024) 'A genomic mutational constraint map using variation in 76,156 human genomes', *Nature*, 625(7993), pp. 92–100. Available at: https://doi.org/10.1038/s41586-023-06045-0.

Cohen, J. *et al.* (2005) 'Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9.', *Nature genetics*, 37(2), pp. 161–165. Available at: https://doi.org/10.1038/ng1509.

Davis, H.R.J. *et al.* (2004) 'Niemann-Pick C1 Like 1 (NPC1L1) is the intestinal phytosterol and cholesterol transporter and a key modulator of whole-body cholesterol homeostasis.', *The Journal of biological chemistry*, 279(32), pp. 33586–33592. Available at: https://doi.org/10.1074/jbc.M405817200.

Dron, J.S. and Hegele, R.A. (2018) 'Polygenic influences on dyslipidemias.', *Current opinion in lipidology* [Preprint]. Available at: https://doi.org/10.1097/MOL.0000000000000482.

Emi, M. *et al.* (1988) 'Genotyping and sequence analysis of apolipoprotein E isoforms.', *Genomics*, 3(4), pp. 373–379. Available at: https://doi.org/10.1016/0888-7543(88)90130-9.

Fan, L. *et al.* (2015) 'Novel mutations of low-density lipoprotein receptor gene in China patients with familial hypercholesterolemia.', *Applied biochemistry and biotechnology*, 176(1), pp. 101–109. Available at: https://doi.org/10.1007/s12010-015-1554-x.

Ference, B.A. *et al.* (2017) 'Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel.', *European heart journal*, 38(32), pp. 2459–2472. Available at: https://doi.org/10.1093/eurheartj/ehx144.

Garcia-Calvo, M. *et al.* (2005) 'The target of ezetimibe is Niemann-Pick C1-Like 1 (NPC1L1).', *Proceedings of the National Academy of Sciences of the United States of America*, 102(23), pp. 8132–8137. Available at: https://doi.org/10.1073/pnas.0500269102.

Haas, J. *et al.* (2015) 'Atlas of the clinical genetics of human dilated cardiomyopathy.', *European heart journal*, 36(18), pp. 1123–35a. Available at: https://doi.org/10.1093/eurheartj/ehu301.

Hobbs, H.H., Brown, M.S. and Goldstein, J.L. (1992) 'Molecular genetics of the LDL receptor gene in familial hypercholesterolemia.', *Human mutation*, 1(6), pp. 445–466. Available at: https://doi.org/10.1002/humu.1380010602.

Hooper, A.J. *et al.* (2012) 'Genetic analysis of familial hypercholesterolaemia in Western Australia.', *Atherosclerosis*, 224(2), pp. 430–434. Available at: https://doi.org/10.1016/j.atherosclerosis.2012.07.030.

Iacocca, M.A. *et al.* (2018) 'Whole-Gene Duplication of PCSK9 as a Novel Genetic Mechanism for Severe Familial Hypercholesterolemia.', *The Canadian journal of cardiology*, 34(10), pp. 1316–1324. Available at: https://doi.org/10.1016/j.cjca.2018.07.479.

Johansen, C.T. *et al.* (2010) 'Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia.', *Nature genetics*, 42(8), pp. 684–687. Available at: https://doi.org/10.1038/ng.628.

Lee, M.H. *et al.* (2001) 'Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption.', *Nature genetics*, 27(1), pp. 79–83. Available at: https://doi.org/10.1038/83799.

Leigh, S. *et al.* (2017) 'The UCL low-density lipoprotein receptor gene variant database: pathogenicity update.', *Journal of medical genetics*, 54(4), pp. 217–223. Available at: https://doi.org/10.1136/jmedgenet-2016-104054.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform.', *Bioinformatics (Oxford, England)*, 25(14), pp. 1754–1760. Available at: https://doi.org/10.1093/bioinformatics/btp324.

Lye, S.-H. *et al.* (2013) 'Genetic polymorphisms in LDLR, APOB, PCSK9 and other lipid related genes associated with familial hypercholesterolemia in Malaysia.', *PloS one*, 8(4), p. e60729. Available at: https://doi.org/10.1371/journal.pone.0060729.

McKenna, A. *et al.* (2010) 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.', *Genome research*, 20(9), pp. 1297–1303. Available at: https://doi.org/10.1101/gr.107524.110.

McLaren, W. *et al.* (2016) 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17(1), p. 122. Available at: https://doi.org/10.1186/s13059-016-0974-4.

Pan-Lizcano, R. *et al.* (2022) 'Rare Variants in Genes of the Cholesterol Pathway Are Present in 60% of Patients with Acute Myocardial Infarction.', *International journal of molecular sciences*, 23(24). Available at: https://doi.org/10.3390/ijms232416127.

Pereira, E. *et al.* (1995) 'Recurrent and novel LDL receptor gene mutations causing heterozygous familial hypercholesterolemia in La Habana.', *Human genetics*, 96(3), pp. 319–322. Available at: https://doi.org/10.1007/BF00210415.

Pisciotta, L. *et al.* (2006) 'Additive effect of mutations in LDLR and PCSK9 genes on the phenotype of familial hypercholesterolemia.', *Atherosclerosis*, 186(2), pp. 433–440. Available at: https://doi.org/10.1016/j.atherosclerosis.2005.08.015.

Rahalkar, A.R. and Hegele, R.A. (2008) 'Monogenic pediatric dyslipidemias: classification, genetics and clinical spectrum.', *Molecular genetics and metabolism*, 93(3), pp. 282–294. Available at: https://doi.org/10.1016/j.ymgme.2007.10.007.

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.

Rentzsch, P. *et al.* (2019) 'CADD: predicting the deleteriousness of variants throughout the human genome', *Nucleic Acids Research*, 47(D1), pp. D886–D894. Available at: https://doi.org/10.1093/nar/gky1016.

Richards, S. *et al*. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine,* 17(5), 405–424. https://doi.org/10.1038/gim.2015.30

Stenson, P.D. *et al.* (2020) 'The Human Gene Mutation Database (HGMD(®)): optimizing its use in a clinical diagnostic or research setting.', *Human genetics*, 139(10), pp. 1197–1207. Available at: https://doi.org/10.1007/s00439-020-02199-3.

Thiart, R. *et al.* (2000) 'Mutation analysis in a small cohort of New Zealand patients originating from the United Kingdom demonstrates genetic heterogeneity in familial

hypercholesterolemia.', *Molecular and cellular probes*, 14(5), pp. 299–304. Available at: https://doi.org/10.1006/mcpr.2000.0318.

'Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report.' (2002) *Circulation*, 106(25), pp. 3143–3421.

Wang, K., Li, M. and Hakonarson, H. (2010) 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.', *Nucleic acids research*, 38(16), p. e164. Available at: https://doi.org/10.1093/nar/gkq603.

WHO (2017) *Cardiovascular diseases (CVDs)*, *Cardiovascular diseases (CVDs)*.

Zeissig, S. *et al.* (2010) 'Primary deficiency of microsomal triglyceride transfer protein in human abetalipoproteinemia is associated with loss of CD1 function.', *The Journal of clinical investigation*, 120(8), pp. 2889–2899. Available at: https://doi.org/10.1172/JCI42703.

Chapter 3 (Manuscript 2): Evaluation of Polygenic Scores for LDL-C in Sub-Saharan African Populations

3.1. Abstract

Polygenic scores (PGS) are used to determine the genetic contribution to complex traits like LDL cholesterol (LDL-C), but most PGS are developed in populations of European ancestry, limiting their applicability in diverse populations. This study evaluates the predictive performance of four published LDL-C PGS, developed in European, multi-ancestry and African ancestry-enriched cohorts, in target Sub-Saharan African populations. Using data from the AWI-Gen cohort (n=10 603) we assessed the variance explained (R^2) by the four PGS (Graham/African ancestry, Trinder/~20% African ancestry, Vanhoye/European ancestry, and Zhang/African ancestry) in predicting LDL-C across Sub-Saharan African regions in South, East and West Africa. Linear regression models were adjusted for age, sex, and principal components to account for population structure. The R^2 values ranged from 0.03 to 0.205, with the Graham and Zhang scores demonstrating the highest overall predictive accuracy in the full AWI-Gen sample (R^2=0.205 and R^2=0.196), but performance varied substantially by geographic region. Counterintuitively, since the African ancestry PGS were developed in predominantly African-Americans with mostly West African ancestry, scores tended to perform better in South and East Africa compared to West Africa. Current PGS for LDL-C have limited predictive power in African populations and show regional variability within the continent. These findings highlight the need for ancestry-diverse discovery cohorts and region-specific validation to improve the utility of PGS in African settings.

3.2. Introduction

Dyslipidaemia is defined as abnormal (either high or low) lipid levels (Braun *et al.*, 2016). Lipid levels are regulated by a complex interaction between genetic and non-genetic factors, classifying dyslipidaemia as a multifactorial disorder. High Low-Density Lipoprotein Cholesterol (LDL-C) levels are exacerbated by the consuming fatty, processed foods and low levels of adequate physical activity, in the context of predisposing genetic factors, and can lead to cardiovascular diseases. Rare genetic variants with large phenotypic effects cause monogenic forms of dyslipidaemia, for example, FH, but the majority of cases with high LDL-C are the result of many common genetic variants, each with small to moderate phenotypic effects, that cumulatively underlie polygenic forms of the disorder (Dron and Hegele, 2018). The heritability of LDL-C levels is estimated to be between 40 and 60% (Weiss *et al.*, 2006).

The European Atherosclerosis Society states that high LDL-C is a causal factor for atherosclerotic cardiovascular disease (ASCVD) based on genetic studies, prospective epidemiologic cohort studies, and Mendelian randomization studies, and have been validated in randomized trials. The evidence demonstrates a consistent relationship between LDL-C levels and ASCVD risk (Ference *et al.*, 2017).

Polygenic scores (also referred to as polygenic risk scores, polygenic trait scores, SNP scores or genetic risk scores) are designed to measure the genetic contribution (usually the sum of weighted associated single nucleotide polymorphism (SNP) alleles) to a particular phenotype. Here we assess a series of PGS for LDL-C developed to measure the overall burden of lipid-related genetic variants identified through genome-wide association studies (GWASs). A higher score reflects a greater accumulation of risk-associated SNP alleles. These scores can be assessed for their statistical correlation with quantitative traits such as LDL-C levels in individuals. PGSs have been suggested as a tool to stratify the population to identify high risk individuals and to target them for interventions, such as treatment for elevated cholesterol levels and for the prevention of coronary heart disease (Abramowitz *et al.*, 2025).

Despite the potential value of a PGS to identify individuals with high genetic susceptibility for diseases, the majority of PGSs are developed and validated in individuals of European ancestry. This European discovery bias can reduce the predictive performance in non-European populations, especially in those of African descent (Duncan *et al.*, 2019; Martin *et al.*, 2019). Differences in allele frequencies, linkage disequilibrium patterns, and genetic architecture between populations contribute to this reduced accuracy (Cavazos and Witte, 2021). Recent studies have shown that PGSs for cardiometabolic traits derived from European datasets have limited predictive power in African and South Asian populations (Wojcik *et al.*, 2019; Topriceanu *et al.*, 2023). These findings highlight the urgent need for greater diversity in GWASs to improve the global applicability and equity of PGS-based approaches.

When it comes to lipid traits, GWASs have identified over 700 independent associated SNPs in European and Asian populations (Willer *et al.*, 2008; Chasman *et al.*, 2009; Asselbergs *et al.*, 2012; Liu *et al.*, 2017; Hoffmann *et al.*, 2018). Recently, there has been a shift in focus from European studies to multi-ethnic studies (including African ancestry populations) to assess genetic associations and develop PGS for lipid traits (Choudhury *et al.*, 2022). A study utilizing the summary statistics from the Million Veteran's Program showed that PGSs that were generated from African Americans strengthened the prediction of different lipid traits in Sub-Saharan Africa when compared to scores generated from European and even multi-

ancestry populations, however it was also shown in another study that prediction accuracy varied within different African populations such as Ugandan and South African Zulu individuals (Kamiza *et al.*, 2022).

Africa plays a vital role in genetic research due to the high levels of genetic diversity in African populations, making them important for discovering novel disease-associated variants (Oni-Orisan Akinyemi *et al.*, 2021). Despite this, they remain vastly underrepresented in genomic studies, leading to missed opportunities for discovery and perpetuating health disparities (Popejoy and Fullerton, 2016). This lack of inclusion has contributed to the limited accuracy and applicability of PGSs in African populations.

Initiatives such as H3Africa contribute to bridging this gap, aiming to increase genomic data from African cohorts. The inclusion of African-ancestry data in genomic analyses has been shown to improve the accuracy and transferability of PGSs across populations, highlighting the importance of increasing diversity in global genomic research (Cavazos and Witte, 2021).

The aim of this study is to assess the transferability of recently developed PGSs for LDL-C in targeted African populations. The data from the Africa Wits INDEPTH Partnership for genomics studies (AWI-Gen) cohort participants from four African countries are used for this study (Ramsay *et al.*, 2016; Ali *et al.*, 2018). Specifically we aim to: (1) compare the predictive utility of selected published PGSs in AWI-Gen; (2) evaluate regional variation in PGS performance in populations spanning three regions in Sub-Saharan Africa (East, West, and South); (3) assess specific PGSs in individuals with extreme (both high and low) LDL-C levels; and (4) assess the contribution of LDL-C PGS to dyslipidaemia and CVD.

3.3. Methods

3.3.1. PGS Selection

Using the PGS Catalogue (https://www.pgscatalog.org/), we selected PGS models for LDL-C levels published after 2020 to ensure inclusion of most up to date scores. We chose the PGS models with a GWAS cutoff of $p<10$. Preference was given to PGS models that included African ancestry participants to enable evaluation of their predictive performance across three Sub-Saharan African regions and in the combined Sub-Saharan African cohort. In addition, one PGS developed exclusively in a European ancestry population was included to assess the transferability and comparative performance of European-derived scores in African populations.

3.3.2. AWI-Gen Target Cohort

The H3Africa AWI-Gen study is a population cross-sectional study that includes over 12 000 participants (males and females; predominantly 40 to 60 years) from Burkina Faso (Nanoro), Ghana (Navrongo), Kenya (Nairobi) and South Africa (Agincourt, Dikgale/DIMAMO and Soweto). The study was established to investigate genomic and environmental factors that influence body composition, body fat distribution and cardiometabolic disease (CMD) risk. Biospecimens, including fasting whole blood samples were collected from each participant and biochemical tests were performed, including serum fasting lipid profiles (total cholesterol, triglycerides, measured and calculated LDL-cholesterol and HDL-cholesterol). Detailed questionnaires were administered alongside the collection of anthropometric measurements, yielding extensive data on demographics, health history, environmental exposures, behaviours, and biomarkers (Ramsay *et al.*, 2016; Ali *et al.*, 2018). The AWI-Gen study was approved by the Human Research Ethics Committee (HREC) (Medical) of the University of the Witwatersrand (Wits) (protocol number M121029) and renewed in 2017 and 2022 (protocol numbers M170880 and M2210108).

Genotyping of individuals from the AWI-Gen cohort was carried out using the 2.3M SNP H3Africa array at Illumina® FastTrack™ Microarray Services (Illumina, San Diego, USA). The Illumina pipeline was utilized for genotype calling, followed by quality control procedures. Imputation was conducted using the African Genome Resources reference panel via the Sanger Imputation Server (https://imputation.sanger.ac.uk/). Further quality control after impouation included the exclusion of indels, rare variants (minor allele frequency ≤ 0.01), and poorly imputed SNPs (Info score ≤ 0.6). The final dataset comprised 10,603 participants and 13.98 million SNPs, as described in further detail (Choudhury *et al.*, 2022).

3.3.3. Predictive utility of PGSs developed to predict LDL-C levels in Sub-Saharan Africa, and stratified for East Africa, South Africa and West Africa

One of the four selected PGSs was originally based on the GRCh38 genome build. To ensure compatibility with the AWI-Gen genetic data, these PGSs were converted to GRCh37 using the LiftOver tool (Genovese *et al.*, 2024). For each PGS, only SNPs present in the AWI-Gen imputed dataset with matching alleles were retained for downstream analysis. PGSs were calculated for each AWI-Gen participant using the --*score* function in PLINK (v1.9). To assess the predictive performance of each PGS-derived score for LDL-C levels, linear regression models were fitted using StataMP 15.1, adjusting for age, sex, and 10 principal components

(PCs) for all the AWI-Gen participants. This analysis was then repeated with participants stratified by region-East Africa (Kenya), South Africa and West Africa (Ghana and Burkina Faso). In the region-stratified models, the adjustment included age, sex, and the first five PCs to avoid overcorrection for population structure.

3.3.4. Assessing PGSs developed to predict LDL-C levels in individuals with extreme LDL-C levels

Individuals with extreme LDL-C levels we chose on the basis of having dyslipidaemia. A total of 192 participants were selected across the different AWI-Gen centres; 96 with high LDL-C levels (>4 mmol/L) and 96 with low LDL-C levels (<1 mmol/L) DNA from each of the individuals was sequenced using a targeted next-generation sequencing panel of the exons of 16 genes, as previously described (Chapter 2, to be submitted for publication). Using the sequence data, SNPs known or predicted to be associated with altered LDL-C plasma levels were used to identify individuals with potential monogenic causes of either phenotype. After removing the individuals with monogenic variants previously shown to be associated with altered LDL-C levels, PGSs were calculated in the remaining individuals to evaluate their potential contribution to the phenotypes.

3.3.5. Assessing LDL-C PGSs prediction accuracy of Dyslipidaemia and Cardiovascular Disease (CVD) Status

CVD status and dyslipidemia were defined as described by Micklesfield *et al.* (Micklesfield *et al.*, 2023). CVD was considered present if a participant self-reported a history of heart attack, stroke, or transient ischaemic attack, or had a prior diagnosis of congestive heart failure or angina. Due to missing data on angina and heart attack prevalence for men in Soweto, CVD prevalence estimates were calculated using data from the remaining sites. Dyslipidaemia was defined as the presence of one or more of the following lipid abnormalities measured in fasting blood: Total cholesterol (TC) $\geq$ 5.0 mmol/L, High-density lipoprotein cholesterol (HDL-C) < 1.0 mmol/L for men and < 1.3 mmol/L for women and Low-density lipoprotein cholesterol (LDL-C) $\geq$ 3.0 mmol/L.

Logistic regression analyses were performed to assess the correlation between (1) LDL-C levels and CVD status, (2) PGS for LDL-C and dyslipidaemia status and (3) PGS for LDL-C and CVD status. The percentage of individuals with dyslipidaemia and CVD were calculated per PGS decile.

3.4. Results

3.4.1. Cohort Description

Table 1 describes the number of participants included in the analyses and their mean age, sex distribution and mean LDL-C levels, which were similar across the regions and LDL-C distributions are shown in Supplemental Figure 1.

Table 1: Demographic data of AWI-Gen cohort and per region

	All Regions	East Africa	South Africa	West Africa
n	10376	1732	4937	3707
Age (yr)	51,8±8,2	48,9±5,3	53,9±9,7	50,4±5,9
% Males	45,2	45,4	42,8	48,5
Mean LDL-C (mmol/L)	2,3±0,9	2,6±0,9	2,5±0,9	1,9±0,7

3.4.2. PGS Selection

Four PGSs developed to predict LDL-C levels were selected for evaluation and are described in Table 2, showing the number and percentage of SNPs available in AWI-Gen data for each of the scores (ranging from ~80 to over 90% of SNPs). Two of these scores, developed by Graham *et al*. (2021) and Zhang *et al*. (2024), were derived entirely from cohorts of African ancestry participants (mostly admixed individuals from the African diaspora). The PGS by Trinder *et a*l. (2020) was based on a discovery cohort that included approximately 20% individuals of African descent. The fourth score, developed by Vanhoye *et al*. (2023), was derived from a European ancestry cohort and was included in this study to serve as a comparator for assessing predictive performance against scores developed using African ancestry populations.

Table 2: Selected Polygenic Scores for LDL-C levels from recent studies (PGS catalog, accessed April 2025)

Reference	PGS Catalog ID	% African ancestry Discovery participants	Cohort Size	Number of SNPs	% SNPs in AWI-Gen (n)
Graham SE *et al.* Nature (2021)	PGS000887	100	99 432	295	76,9 (227)
Trinder M *et al.* JAMA Cardiol (2020)	PGS000115	19.3	48 741	223	81,2 (181)
Zhang J *et al.* Nat Commun (2024)	PGS004637	100	87 760	173	94,2 (163)
Vanhoye X *et al.* Transl Res (2023)	PGS003405	0	427 099	169	71,0 (120)

Graham *et al.* (2021) carried out a comprehensive GWAS meta-analysis which included over 99 000 individuals of African descent (African-American) (Graham *et al.*, 2021). The method employed to generate the African specific PGS was Pruning and Thresholding, whereby independent SNPs were selected according to linkage disequilibrium patterns and significance thresholds. PLINK was used for the pruning and thresholding, METAL for meta-analyses and Python for statistical analyses. The PGS that was developed for LDL-C for African populations demonstrated adjusted R^2 values ranging from 0.10 to 0.16 in different African cohorts.

Trinder *et al.* (2020) (Trinder, Francis and Brunham, 2020) constructed a PGS using independent SNPs originally identified by the phase 1 Global Lipid Genetics Consortium (GLGC) (Pubmed ID: 24097068, 29083407) and was developed for LDL-C prediction. Approximately 19% of the discovery cohort was of African descent, specifically African-

American or Afro-Caribbean. A PGS R^2 value of 0.04 was obtained in individuals of African descent compared to 0.09 in Europeans.

Zhang *et al.* (2024) utilized a novel method, namely Polygenic Risk scOres based on ensemble of Penalized Regression models (PROSPER), which consolidates GWAS summary statistics from populations of different ethnicities to develop ancestry specific PGSs to predict LDL-C levels (Zhang *et al.*, 2024). This method uses both lasso and ridge penalized regression techniques to model effect sizes within each different population. The LDL-C R^2 value (0.141) was shown to be higher than those of other methods, including PGS-CS, the method utilized by Graham *et al*. in 2021.

Vanhoye *et al.* (2022) developed a PGS to predict LDL-C levels utilizing an entirely European cohort (Vanhoye *et al.*, 2023). SNPs were selected using a meta-analysis from 3 different GWASs from the GLGC. PLINK and PGSice-2 were employed for PGS construction, with a LDL-C R^2=0.19.

There was an overlap between the three scores with regards to African discovery cohorts. Graham (2021) utilized GLGC (with individuals of African-American ancestry). Zhang (2024) also utilized data from the GLGC. Trinder (2020) utilized data from the UK Biobank. There are only 2 SNPs that were present in all the PGS (Figure 1), 30 SNPs were in both the PGS of Graham (2021) and Zhang (2024), and 62 SNPs in both the PGS of Trinder (2020) and Vanhoye (2023).

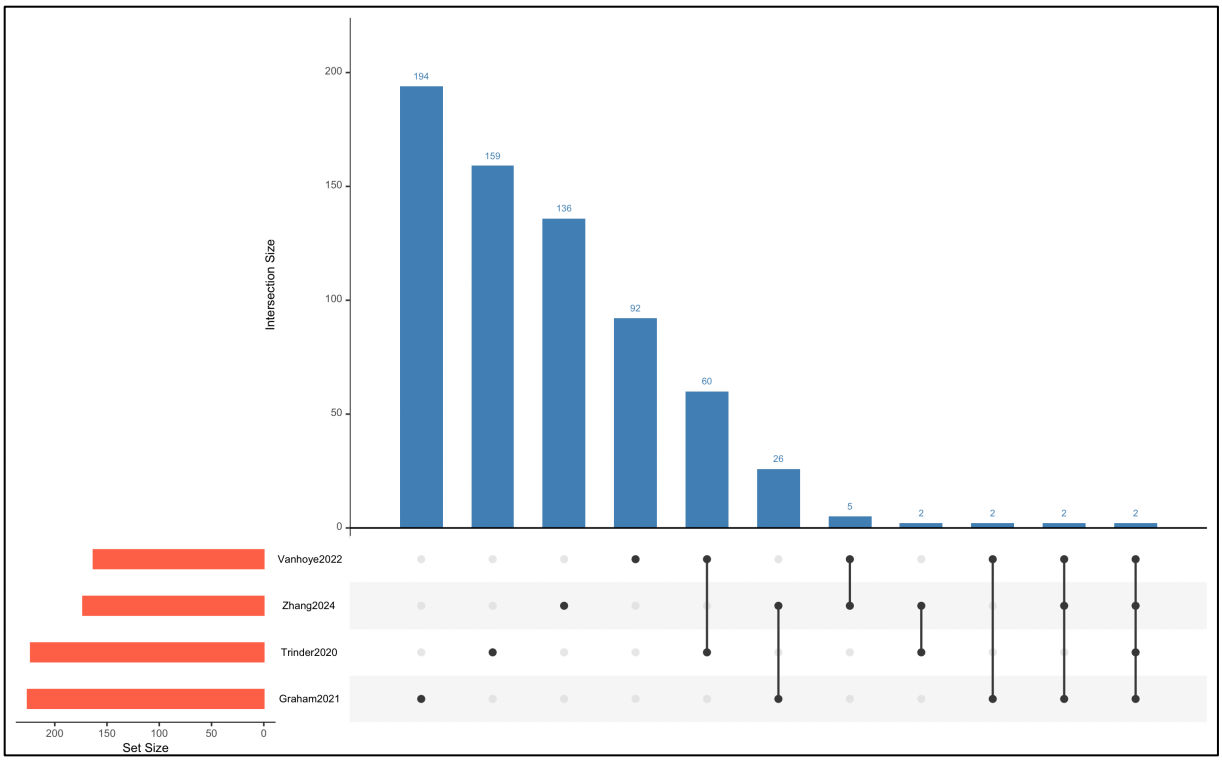


Figure 1: Upset Plot showing overlap of SNPs across the four PGSs. The majority of SNPs were present only in one of the PGSs and overlaps were minimal.

3.4.3. Predictability of different PGSs for LDL-C in AWI-Gen (Sub-Saharan Africa) and per region

Linear regression analyses were performed using all four PGSs to assess their predictability of variation in LDL-C levels. Figure 2 shows the adjusted R-squared values for each score. and Root Mean Squared Error (RMSE) values for each score are shown in Supplementary Figure 3. As expected, the PGSs by Graham and Zhang demonstrated the highest predictive performance, explaining 20.5% and 19.6% of the variance in LDL-C levels, respectively. The RMSE values were highest for the Trinder and Vanhoye PGSs suggesting greater average prediction error. Figures 3A-D illustrate the average LDL-C level increase with each PGS decile for each of the four scores and Figure 3E shows the trend of median LDL-C within each decile. Overall, the Graham and Zhang PGSs showed better correlation with increased LDL-C levels across PGS decile.

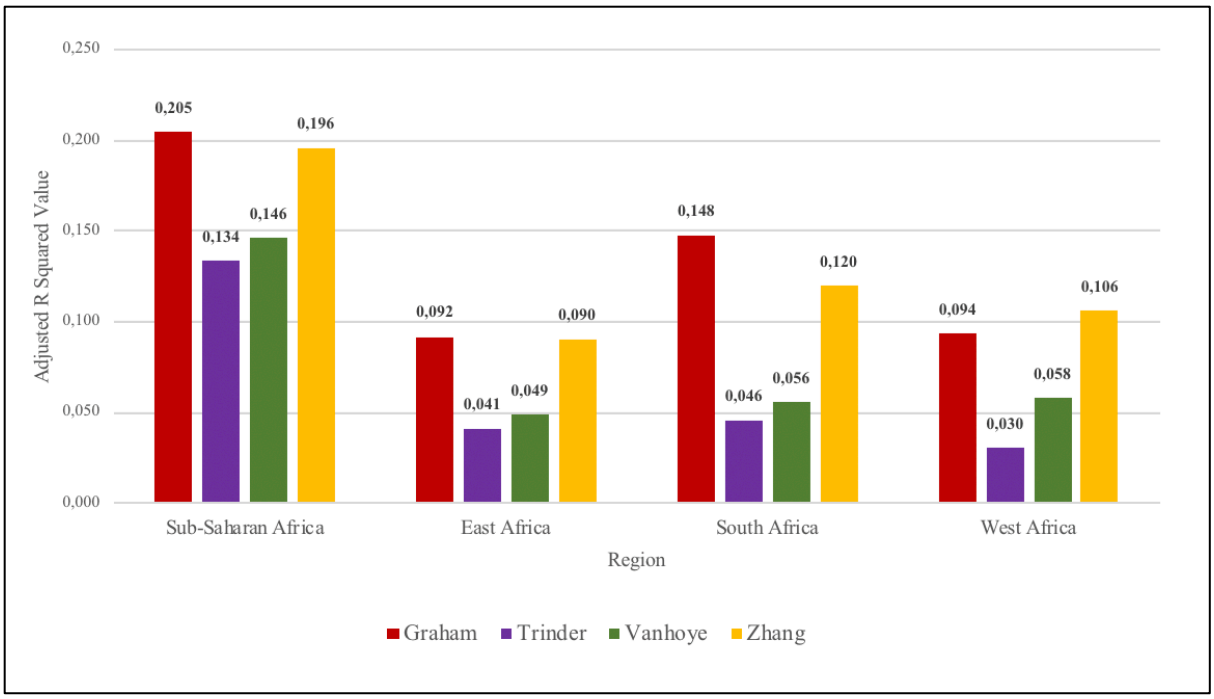


Figure 2: Assessing the performance (adjusted R^2 value) of the four PGSs (Graham, Tinder, Vanhoye and Zhang) for LDL-C in the full Sub-Saharan African AWI-Gen target sample (n=10 376), East African region (n=1 732), South African Region (n=4 937) and West Africa (n=3 707)

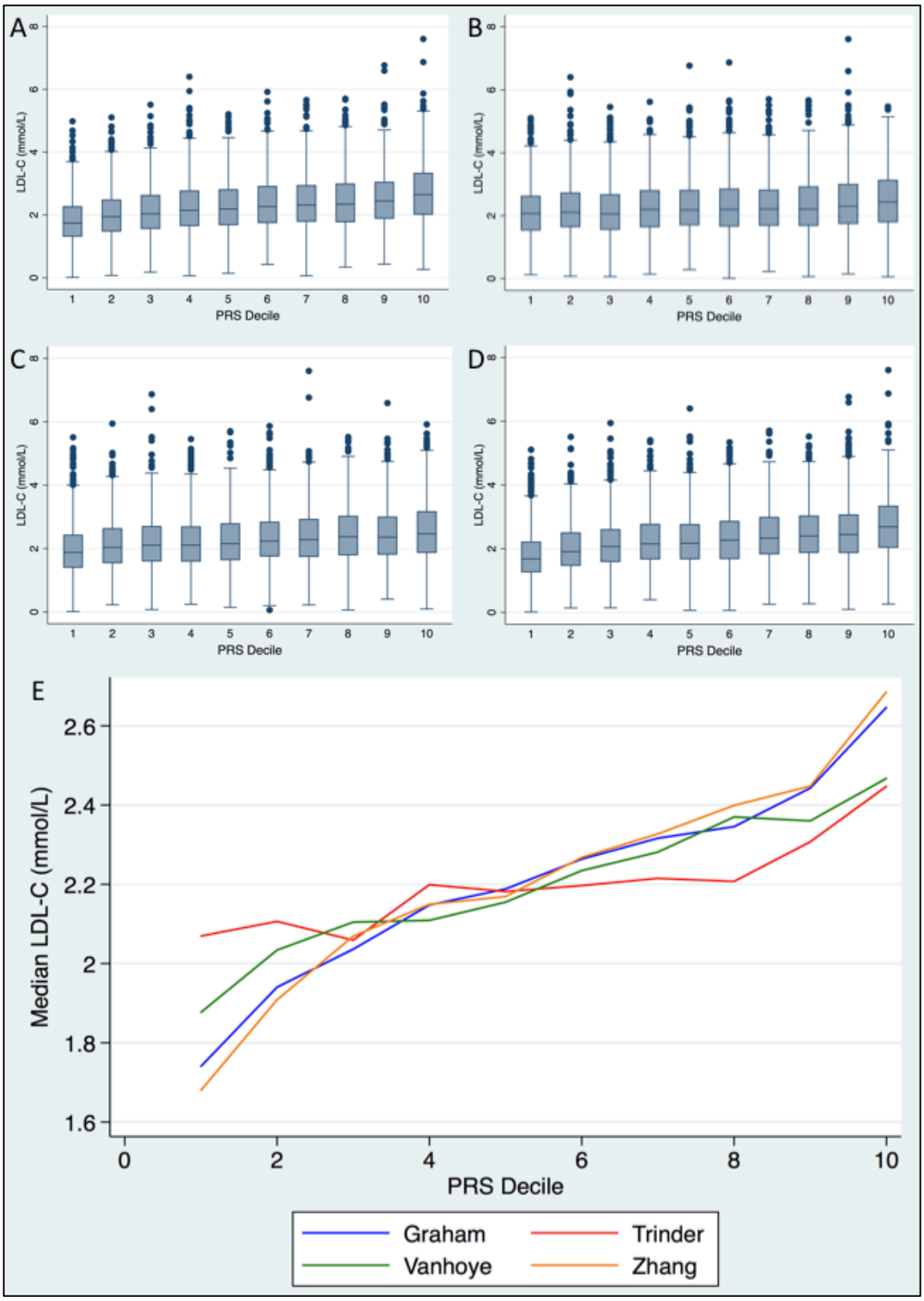


Figure 3: Assessing the distributions of LDL-C levels per PGS decile for A: Graham Score, B: Trinder Score, C: Vanhoye Score and D: Zhang Score and E: Comparing the median LDL-C levels per decile of each of the scores showing the better performance of the Graham and Zhang Scores.

3.4.4. Predictability of different PGSs of LDL-C in different African regions

When stratified by region, performance of all four PGS was lower relative to the full cohort, though scores derived from the two African ancestry cohorts (Graham and Zhang) had consistently better prediction accuracy (Figure 2). The Graham score demonstrated the highest regional predictive power, with adjusted R^2 values of 0.15 in South Africa (RMSE = 0.83), 0.09 in East Africa (RMSE = 0.83) and 0.09 in West Africa (RMSE = 0.71). The Zhang score also showed good performance, particularly in South Africa (R^2 = 0.12; RMSE = 0.84), followed by West Africa (R^2 = 0.11; RMSE = 0.70) and East Africa (R^2 = 0.09; RMSE = 0.83). In contrast, the Trinder score with only ~20% African ancestry participants showed the lowest prediction accuracy across regions, with adjusted R^2 values of just 0.05 in South Africa (RMSE = 0.88), 0.04 in East Africa (RMSE = 0.85), and 0.03 in West Africa (RMSE = 0.73). The Vanhoye score, derived entirely from a European ancestry discovery cohort, also performed poorly, with regional R^2 values of 0.056 in South Africa (RMSE = 0.87), 0.05 in East Africa (RMSE = 0.85), and 0.06 in West Africa (RMSE = 0.72). Despite not showing the highest adjusted R^2 values, West Africa showed the lowest Root MSE indicating the lowest prediction error rate, which may be because African-Americans were included in the development of the scores, and they have more West Africa genetic ancestry. This could indicate that despite being on par with East Africa, the model may be a better fit over all. South Africa, with the largest sample size, has the highest adjusted R^2 values however also had the highest Root MSE.

Figure 4 depicts median LDL-C levels across PGS deciles for the Graham and Zhang scores respectively. The difference in mean LDL-C levels between the lowest and highest decile was 0.811 mmol/l in East, 1.044 mmol/l in South and 0.671 mmol/l in West Africa when applying the Graham PGS compared to 0.856 mmol/L, 0.887 mmol/L and 0.743 mmol/L for the Zhang PGS. Figure 5 shows the effect on LDL-C levels per standard deviation (SD) score coefficient with the 95% confidence levels for each score in all regions, including the Sub-Saharan Africa. Graham and Zhang scores show the highest effects, however the Zhang score is more consistent across the regions. The models of all four scores in East Africa, despite not being the worst performing region, show the largest confidence levels, indicating a less accurate fit.

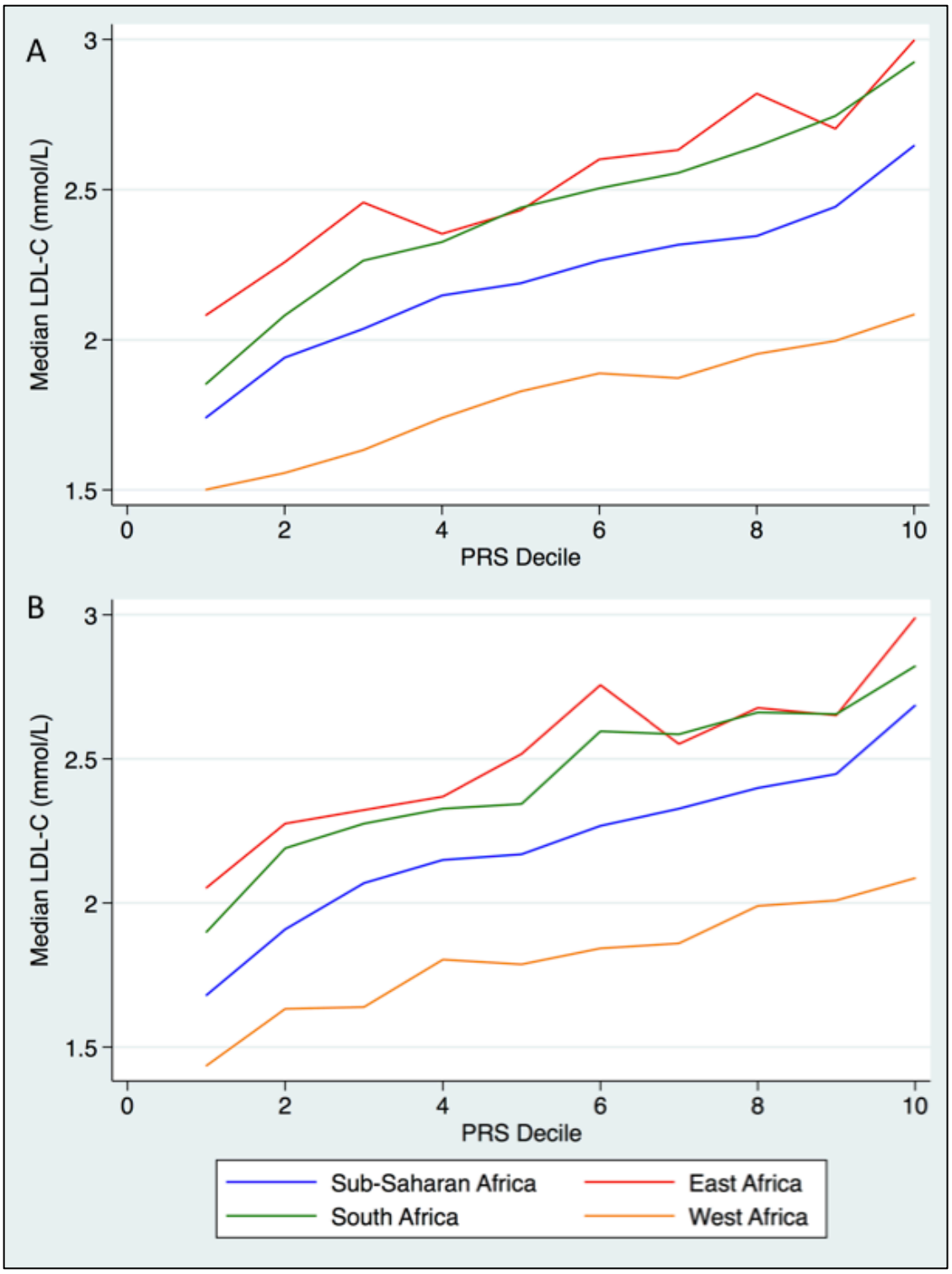


Figure 4: Performance of the A: Graham PGS score and B: Zhang PGS Score for Median LDL-C levels across risk deciles in Sub-Saharan Africa (n=10 376), East Africa (n=1 732), South Africa (n=4 937) and West Africa (n=3 707).

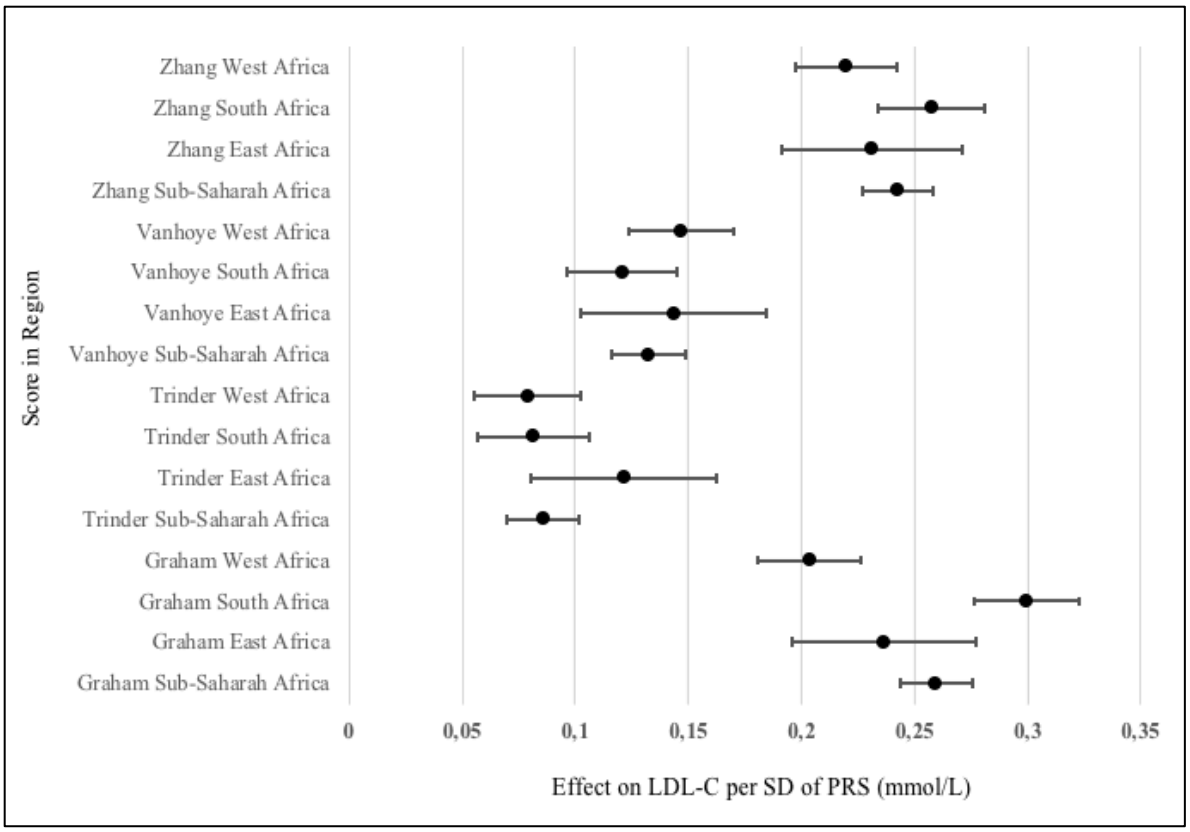


Figure 5: Forrest Plot illustrating the Standard Deviation Score Coefficient and 95% confidence intervals for each score in each region.

3.4.5. Assessing PGSs in individuals with extreme LDL-C levels

Of the 191 participants selected according to extreme LDL-C levels for a candidate genes study (described in chapter 2), 19 were excluded due to having a known monogenic variant in either *LDLR* (LDL-C increasing) or *PCSK9* (loss of function and LDL-C lowering) and 25 were excluded as genotype data was not available, leaving 147 participants for this analysis (83 with high LDL-C levels and 64 with low LDL-C levels). Figure 7 shows the number of participants per decile of the 2 best performing scores. Both scores show the expected trends, supporting the underlying polygenic aetiology in this sub-group with extreme phenotypes. Two individuals with high LDL-C levels (5,5 and 6,5 mmol/L) who did not have potentially functional variants in *LDLR*, *PCSK9* or *APOB*, fall in the 10th decile for the Graham score and 8-10th deciles for the other three scores. Interestingly, one individual with a low level of LDL-C and no potentially functional variant in the three abovementioned genes, was in the highest decile for both the Graham and Zhang scores.

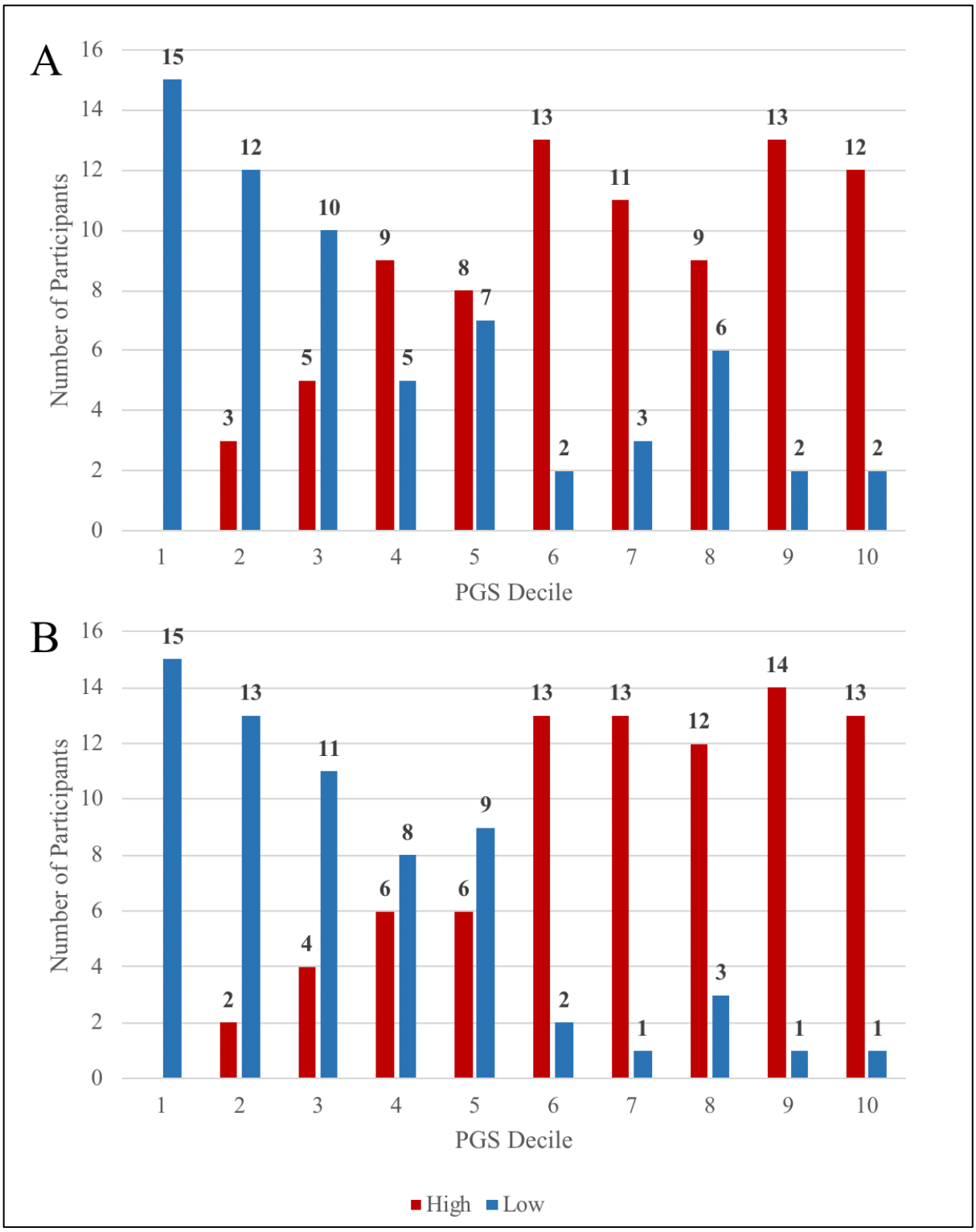


Figure 6: The number of participants with either high or low LDL-C levels in each A: Graham score decile and B: Zhang score decile

3.4.6. Evaluating the Accuracy of PGS in Predicting Dyslipidaemia and CVD

A total of 302 out of 8 446 individuals with genotype data and CVD status information presented with CVD. Logistic regression analysis was performed to determine the correlation

between LDL-C levels and CVD status. LDL-C levels were significantly associated with CVD status ($p<0.001$) and for every 1 mmol/L increase in LDL-C, the log odds of having CVD increased by 0.24, with an odds ratio of 1.28. Logistic regression analysis was then performed to determine if the LDL-C PGS scores from Graham and Zhang were associated with dyslipidaemia and then CVD status. Out of the 10 376 individuals with genotype data and dyslipidaemia status, 7 478 individuals were classified as having dyslipidaemia according to the criteria outlined by Micklesfield *et al*. (2023) (Micklesfield *et al.*, 2023). Both Graham and Zhang scores are significantly associated with dyslipidaemia status ($p<0.001$). Figure 7 illustrates the increasing percentage of individuals found per increasing PGS decile.

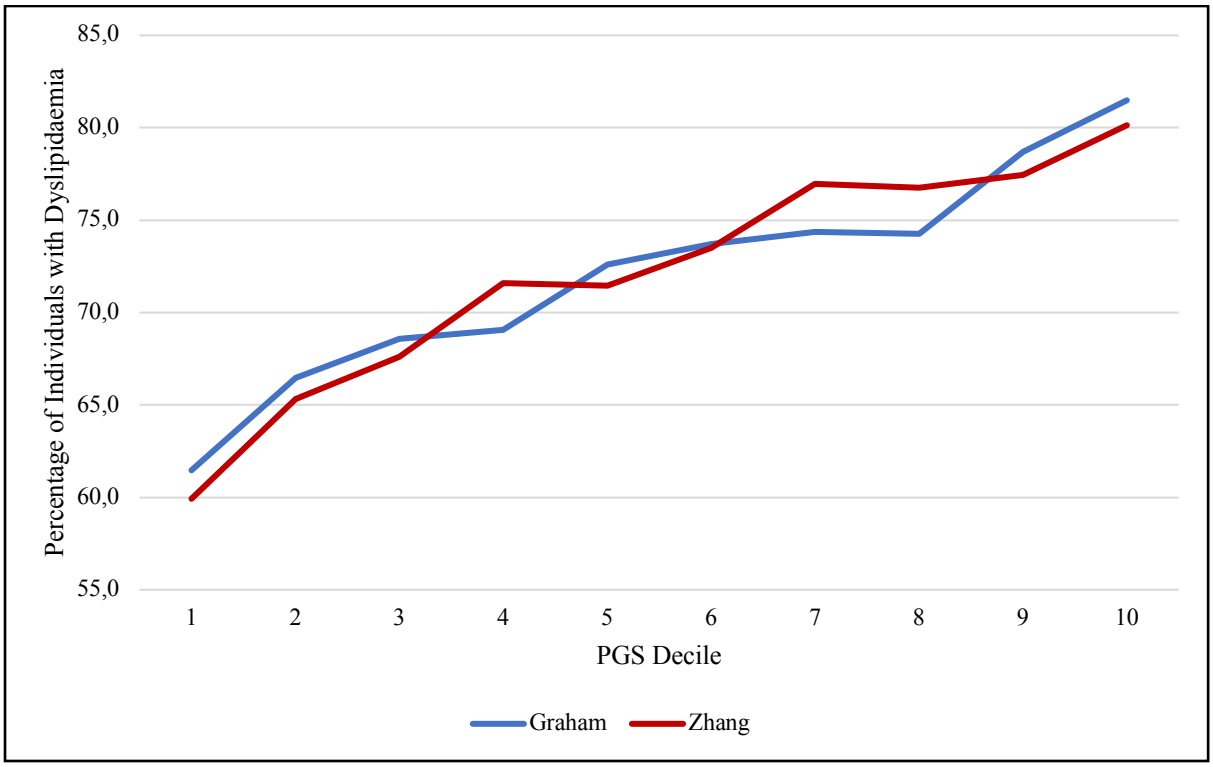


Figure 7: Comparing the percentage of individuals with dyslipidaemia within each decile of the Graham and Zhang PGSs

There was no observed association between CVD status and either scores ($p=0.365$ and $p=0.601$). Table 3 provides the numbers of individuals with CVD per decile of each score.

Table 3: The numbers and percentages of individuals with CVD within each PGS decile

	Graham Score				Zhang Score			
PGS Decile	CVD	No CVD	Total	Percentage with CVD (%)	CVD	No CVD	Total	Percentage with CVD (%)
1	33	779	812	**4,1**	33	809	842	**3,9**
2	26	767	793	**3,3**	31	804	835	**3,7**
3	34	775	809	**4,2**	35	787	822	**4,3**
4	31	799	830	**3,7**	26	810	836	**3,1**
5	37	823	860	**4,3**	25	810	835	**3,0**
6	26	829	855	**3,0**	30	829	859	**3,5**
7	26	834	860	**3,0**	30	821	851	**3,5**
8	27	845	872	**3,1**	35	828	863	**4,1**
9	32	851	883	**3,6**	27	835	862	**3,1**
10	30	842	872	**3,4**	30	811	841	**3,6**

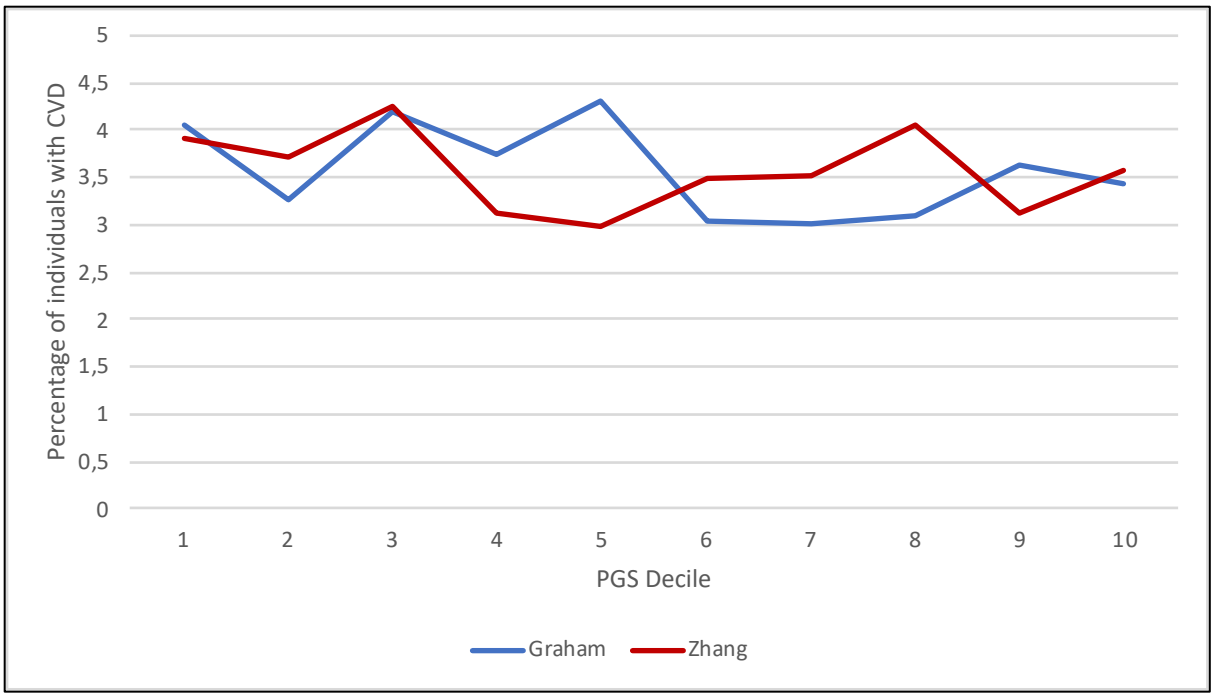


Figure 8: Comparing the percentage of individuals with CVD within each decile of the Graham and Zhang PGSs

3.5. Discussion

PGSs are a useful tool to stratify populations according to relative genetic risk and to target individuals in high-risk groups for specific interventions to reduce the risk of developing disease. In this study we assessed different LDL-C PGSs in target populations in Sub-Saharan Africa with a view to understanding the genetic contribution to the variability in LDL-C. We also examined the proportion of individuals with dyslipidemia and CVD across risk deciles for increased LDL-C levels for the two best performing PGSs.

Four PGSs were selected from the PRS online catalogue, published after 2020 and developed using different methods. Two of the scores, Graham *et al*. (2021) and Zhang *et al*. (2024), were developed using entirely African-ancestry discovery cohorts, specifically African-America, from various studies both including the GLGC data. There were only 26 common SNPs between the two scores, but additional SNPs mapping to shared loci, but with different chosen SNPs due to different models. The two other scores selected, Trinder et al. (2020) and Vanhoye et al. (2023), one discovery cohorts with ~20% African-ancestry and the other with only European ancestry participants. The Graham and Zhang scores, outperformed the Trinder and Vanhoye scores in the target Sub-Saharan African populations, with 20.5% and 19.6% of variance in LDL-C explained respectively in the full cohort. This was also observed when stratifying by geographic region (East, South and West Africa). The R^2 values are higher than

in previous studies, notably that of Choudhury *et al*. (2022) where authors showed only 7% explained variance of a different PGS for LDL-C in this target SSA cohort (Choudhury *et al.*, 2022). This 43 SNP PGS was developed from a Ugandan cohort and despite performing better than scores developed from European and multi-ancestry populations, performed less accurately than a score developed from a wider African-ancestry cohort, likely due to vast genetic variability within Africa (Gurdasani *et al.*, 2019). Kamiza *et al.* (2022) similarly showed that 8.14% of variance was explained by a PGS for LDL-C developed from an African American cohort in a target South African Zulu cohort. This particular study showed that including local ancestry information can improve PGS prediction accuracy within African cohorts, further highlighting the need for regionally generated models rather than a generalized African approach (Kamiza *et al.*, 2022). Our study underscores this by showing vast differences in predicted variance between the regions, where in South Africa the Graham and Zhang scores predict 14.8% and 12.0% of the variance respectively, compared to 9.4% and 10.6% in West Africa. This also aligns with recent studies that show that genetic diversity within Africa can be so varied that PGS accuracy between African regions may exceed differences observed between African and European populations (Majara *et al.*, 2023). Studies on other traits, such as body mass index (BMI), similarly report that polygenic prediction performance varies across African regions, not only in terms of genetic risk, but also genetic and environmental interactions (Chikowore *et al.*, 2024). The African components of existing PGSs are primarily derived from African- American populations rather than continental Africans. Given the substantial West African genetic contribution to African American populations due to the transatlantic slave trade, it is plausible that PGSs derived from African-American cohorts yield better predictive performance in West African populations than in other African populations. Interestingly, our study shows the South African region exhibited higher R^2 values across all four scores. However, the RMSE was lowest in the West African cohort, indicating a lower prediction error. The relatively higher RMSE values in South Africa, despite larger sample sizes, highlight the potential need to interpret R^2 and RMSE together.

These findings are consistent with previous research, which has shown that PGSs developed using European discovery cohorts are often less accurate when applied individuals of African descent (Martin *et al.*, 2019). In contrast, PGSs that are developed from African (or admixed African-ancestry) discovery cohorts demonstrate higher predictive accuracy (Kamiza *et al.*, 2022). Previous studies have also emphasized the challenges of using PGSs across different populations due to intra-continental genetic variability (Fatumo *et al.*, 2023). The varied

statistical results encountered in this study may be due to several factors such as the ancestry of the African discovery cohorts (African American or continental African). Overall, these findings reinforce the importance of conducting larger, regionally diverse genomic studies across the African continent to inform the development of region-specific PGSs, thereby improving their clinical utility and equity.

Extreme LDL-C levels can be caused by monogenic and/or polygenic contributors. While monogenic causes are often presumed to underlie extreme lipid phenotypes, our data demonstrate that polygenic influences may be even more prevalent in the aetiology of LDL-C in African populations. Among the 191 individuals selected based on extreme LDL-C levels, only 19 were found to have potentially pathogenic variants in the three major LDL-C–associated genes (*LDLR*, *PCSK9*, and *APOB*). In contrast, polygenic risk score distributions (Figures 6) clearly demonstrate enrichment patterns suggestive of underlying polygenic contributions in the remaining individuals, highlighting the complexity of genetic risk assessment for dyslipidaemia.

In a study by Khera *et al.* (2019) where whole genome sequencing was performed on over 2000 individuals hospitalized with early onset myocardial infarction, only 1.7% presented with a monogenic variant for FH whereas 17% presented with a high PGS (Khera *et al.*, 2019). This is pertinent to African settings, where high genetic diversity further challenges the application of standard models and highlights the need for regionally modified genomic tools and assays informed by local data.

The prevalence of CVD is rising globally, with a particularly notable increase in Africa. Circulating lipid levels and their underlying genetic determinants are closely associated with the risk of developing CVD, including conditions such as stroke and CAD (Siewert and Voight, 2018). A previous study demonstrated that PGSs can predict the genetic risk for elevated levels of LDL-C and showed an association with risk of ASCVD (Wu *et al.*, 2021), one of the more severe end-points of CVD. In this study we showed that LDL-C is significantly associated with ASCVD status, but then we curiously found that neither score (Graham or Zhang) was associated with ASCVD status, possibly because the study was underpowered. It may also suggest that the genetic contribution to elevated LDL-C levels is only one of many health risk factors that contribute to ASCVD, including obesity, hypertension, diabetes as well as lifestyle and socio-economic status. Therefore, ASCVD risk assessment would benefit from integrated scores taking into consideration all known risk factors.

These results demonstrate the potential of integrating LDL-C PGS together with other risk factors into CVD risk stratification frameworks, but will require additional evidence and models with increased predictivity before being ready for clinical implementation. For PGS to be meaningful in African populations, they must be developed and validated using regionally representative data.

3.6. References

Abramowitz, S.A. *et al.* (2025) 'Evaluating Performance and Agreement of Coronary Heart Disease Polygenic Risk Scores.', *JAMA*, 333(1), pp. 60–70. Available at: https://doi.org/10.1001/jama.2024.23784.

Ali, S.A. *et al.* (2018) 'Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study.', *Global health action*, 11(sup2), p. 1507133. Available at: https://doi.org/10.1080/16549716.2018.1507133.

Asselbergs, F.W. *et al.* (2012) 'Large-Scale Gene-Centric Meta-analysis across 32 Studies Identifies Multiple Lipid Loci', *The American Journal of Human Genetics*, 91(5), pp. 823–838. Available at: https://doi.org/10.1016/j.ajhg.2012.08.032.

Braun, K.V.E. *et al.* (2016) 'The role of DNA methylation in dyslipidaemia: A systematic review.', *Progress in lipid research*, 64, pp. 178–191. Available at: https://doi.org/10.1016/j.plipres.2016.10.002.

Cavazos, T.B. and Witte, J.S. (2021) 'Inclusion of variants discovered from diverse populations improves polygenic risk score transferability.', *HGG advances*, 2(1), p. 100017. Available at: https://doi.org/10.1016/j.xhgg.2020.100017.

Chasman, D.I. *et al.* (2009) 'Forty-Three Loci Associated with Plasma Lipoprotein Size, Concentration, and Cholesterol Content in Genome-Wide Analysis', *PLOS Genetics*, 5(11), p. e1000730. Available at: https://doi.org/10.1371/journal.pgen.1000730.

Chikowore, T. *et al.* (2024) 'Variability of polygenic prediction for body mass index in Africa.', *Genome medicine*, 16(1), p. 74. Available at: https://doi.org/10.1186/s13073-024-01348-x.

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.

Dron, J.S. and Hegele, R.A. (2018) 'Polygenic influences on dyslipidemias.', *Current opinion in lipidology* [Preprint]. Available at: https://doi.org/10.1097/MOL.0000000000000482.

Duncan, L. *et al.* (2019) 'Analysis of polygenic risk score usage and performance in diverse human populations.', *Nature communications*, 10(1), p. 3328. Available at: https://doi.org/10.1038/s41467-019-11112-0.

Fatumo, S. *et al.* (2023) 'Polygenic risk scores for disease risk prediction in Africa: current challenges and future directions.', *Genome medicine*, 15(1), p. 87. Available at: https://doi.org/10.1186/s13073-023-01245-9.

Ference, B.A. *et al.* (2017) 'Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel.', *European heart journal*, 38(32), pp. 2459–2472. Available at: https://doi.org/10.1093/eurheartj/ehx144.

Genovese, G. *et al.* (2024) 'BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies', *Bioinformatics*, 40(2), p. btae038. Available at: https://doi.org/10.1093/bioinformatics/btae038.

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.

Hoffmann, T.J. *et al.* (2018) 'A large electronic-health-record-based genome-wide study of serum lipids', *Nature Genetics*, 50(3), pp. 401–413. Available at: https://doi.org/10.1038/s41588-018-0064-5.

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.

Khera, A.V. *et al.* (2019) 'Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction', *Circulation*, 139(13), pp. 1593–1602. Available at: https://doi.org/10.1161/CIRCULATIONAHA.118.035658.

Liu, D.J. *et al.* (2017) 'Exome-wide association study of plasma lipids in >300,000 individuals', *Nature Genetics*, 49(12), pp. 1758–1766. Available at: https://doi.org/10.1038/ng.3977.

Lu, X. *et al.* (2017) 'Exome chip meta-analysis identifies novel loci and East Asian–specific coding variants that contribute to lipid levels and coronary artery disease', *Nature Genetics*, 49(12), pp. 1722–1730. Available at: https://doi.org/10.1038/ng.3978.

Majara, L. *et al.* (2023) 'Low and differential polygenic score generalizability among African populations due largely to genetic diversity.', *HGG advances*, 4(2), p. 100184. Available at: https://doi.org/10.1016/j.xhgg.2023.100184.

Martin, A.R. *et al.* (2019) 'Clinical use of current polygenic risk scores may exacerbate health disparities.', *Nature genetics*, 51(4), pp. 584–591. Available at: https://doi.org/10.1038/s41588-019-0379-x.

Micklesfield, L.K. *et al.* (2023) 'Identifying the prevalence and correlates of multimorbidity in middle-aged men and women: a cross-sectional population-based study in four African

countries.', *BMJ open*, 13(3), p. e067788. Available at: https://doi.org/10.1136/bmjopen-2022-067788.

Oni-Orisan Akinyemi *et al.* (2021) 'Embracing Genetic Diversity to Improve Black Health', *New England Journal of Medicine*, 384(12), pp. 1163–1167. Available at: https://doi.org/10.1056/NEJMms2031080.

Popejoy, A.B. and Fullerton, S.M. (2016) 'Genomics is failing on diversity.', *Nature*, 538(7624), pp. 161–164. Available at: https://doi.org/10.1038/538161a.

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.

Siewert, K.M. and Voight, B.F. (2018) 'Bivariate Genome-Wide Association Scan Identifies 6 Novel Loci Associated With Lipid Levels and Coronary Artery Disease.', *Circulation. Genomic and precision medicine*, 11(12), p. e002239. Available at: https://doi.org/10.1161/CIRCGEN.118.002239.

Topriceanu, C.-C. *et al.* (2023) 'Validity of European-centric derived cardiometabolic polygenic risk scores in multi-ancestry populations', *medRxiv*, p. 2023.03.20.23287475. Available at: https://doi.org/10.1101/2023.03.20.23287475.

Trinder, M., Francis, G.A. and Brunham, L.R. (2020) 'Association of Monogenic vs Polygenic Hypercholesterolemia With Risk of Atherosclerotic Cardiovascular Disease', *JAMA Cardiology*, 5(4), pp. 390–399. Available at: https://doi.org/10.1001/jamacardio.2019.5954.

Vanhoye, X. *et al.* (2023) 'A new 165-SNP low-density lipoprotein cholesterol polygenic risk score based on next generation sequencing outperforms previously published scores in routine diagnostics of familial hypercholesterolemia.', *Translational research : the journal of laboratory and clinical medicine*, 255, pp. 119–127. Available at: https://doi.org/10.1016/j.trsl.2022.12.002.

Weiss, L.A. *et al.* (2006) 'The sex-specific genetic architecture of quantitative traits in humans.', *Nature genetics*, 38(2), pp. 218–222. Available at: https://doi.org/10.1038/ng1726.

Willer, C.J. *et al.* (2008) 'Newly identified loci that influence lipid concentrations and risk of coronary artery disease', *Nature Genetics*, 40(2), pp. 161–169. Available at: https://doi.org/10.1038/ng.76.

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.

Wu, H. *et al.* (2021) 'Polygenic Risk Score for Low-Density Lipoprotein Cholesterol Is Associated With Risk of Ischemic Heart Disease and Enriches for Individuals With Familial Hypercholesterolemia.', *Circulation. Genomic and precision medicine*, 14(1), p. e003106. Available at: https://doi.org/10.1161/CIRCGEN.120.003106.

Zhang, J. *et al.* (2024) ‘An Ensemble Penalized Regression Method for Multi-ancestry Polygenic Risk Prediction.’ United States. Available at: https://doi.org/10.1101/2023.03.15.532652.

Chapter 4: Discussion

This dissertation investigates the genetic architecture of LDL cholesterol (LDL-C) in continental African populations using the AWI-Gen cohort. The research focused on two main objectives: (1) identifying monogenic contributors to extreme levels of LDL-C and (2) evaluating the performance of previously published polygenic scores (PGSs) in African populations. The findings provide novel insights into the contributions of both rare and common variants to dyslipidaemia in the African population, highlighting the importance of developing ancestry-appropriate genomic tools for equitable clinical utility.

4.1. Design

The AWI-Gen research platform is a cross-sectional population study comprising over 12,000 participants, both male and female, primarily aged between 40 and 60 years. The cohort spans six sites across four African countries: Nanoro (Burkina Faso), Navrongo (Ghana), Nairobi (Kenya), and three sites in South Africa (Agincourt, Dikgale/DIMAMO, and Soweto). AWI-Gen was established to explore the genomic and environmental factors contributing to body composition, fat distribution, and cardiometabolic disease (CMD) risk. As part of the study, biospecimens were collected, including fasting whole blood samples, which were used to perform a range of biochemical analyses such as serum fasting lipid profiles (total cholesterol, triglycerides, LDL-C, and HDL-C). Participants also completed detailed questionnaires and underwent comprehensive anthropometric assessments. This resulted in a rich dataset encompassing demographic, clinical, behavioural, environmental, and biomarker information, which was leveraged for this dissertation.

Firstly, participants from AWI-Gen were selected based on their measured LDL-C levels, selecting individuals with extremes of high and low levels, while excluding those with known confounding features that directly affect LDL-C levels. A targeted Next-Generation Sequencing (NGS) panel was designed, incorporating the exons of 16 genes involved in LDL-C metabolism. Sequencing was performed on the Illumina iSeq100 using the Nextera Flex for Enrichment Library preparation kit. Non-synonymous variants were filtered based on minor allele frequency (MAF) and CADD score to identify those most likely to be high-effect monogenic contributors. Secondly, four previously published polygenic risk scores for LDL-C, were chosen and assessed for prediction accuracy in the AWI-Gen cohort. All four scores were calculated in approximately 10,460 AWI-Gen individuals, for whom genotype data were available. Linear regression analyses were performed to determine the accuracy of LDL-C

prediction in the entire cohort, as well as within each geographic region of AWI-Gen, specifically East, South, and West Africa.

4.2. Familial Hypercholesterolaemia (FH) in African populations

Elevated LDL-C levels in African populations reflect a complex genetic landscape, shaped not only by monogenic pathogenic variants but also by polygenic risk factors, epistatic interactions, and environmental influences. FH is primarily associated with pathogenic variants in three genes, namely *LDLR*, *APOB*, and *PCSK9*; however, studies have consistently shown that 20–40% of individuals with a clinical diagnosis of FH do not carry detectable monogenic variants in these genes, suggesting a significant contribution from polygenic or oligogenic inheritance to hypercholesterolaemia (Talmud *et al.*, 2013; Wang *et al.*, 2016). This complexity is particularly relevant in African populations, where the genetic diversity is highest globally, yet remains understudied.

Currently, data on FH prevalence and genetics across continental African populations are sparse, with most existing research focusing on populations of European or mixed ancestry. However, large-scale initiatives, such as the H3Africa AWI-Gen study, have begun to fill this gap. For instance, AWI-Gen data show that LDL-C levels vary regionally, being significantly higher in South and East African sites (Agincourt, Dikgale, Soweto, Nairobi) compared to West African sites (Nanoro, Navrongo), emphasising the geographical heterogeneity of lipid profiles across the continent. Furthermore, obesity has been strongly associated with elevated LDL-C levels, underscoring the interplay between genetic susceptibility and lifestyle-related risk factors (Nonterah *et al.*, 2025). In this study, individuals with high BMI (>=40 kg/m^2) were removed as to avoid including individuals with high LDL-C due to confounding variables and to limit the cohort to individuals with likely genetic influences.

In South Africa, population-specific founder effects have led to a higher FH prevalence among Afrikaner, Indian, and Jewish groups, where specific mutations in *LDLR* are well-characterised (Smyth, Ramsay and Raal, 2018). However, as in much of Africa, data on Black South African populations remain limited (Smyth, Ramsay and Raal, 2018). The FIND FH program, launched in 2016, aimed to address this gap by systematically identifying individuals with FH across South Africa and genetically characterising FH-causing variants across different ethnic groups. The program successfully identified Black African individuals carrying pathogenic FH-causing variants, although in relatively small numbers. This suggests either a lower prevalence of FH

in these populations or, more likely, under diagnosis due to reduced clinical awareness, limited access to lipid testing, or less-sensitive diagnostic criteria (Raal *et al.*, 2020).

A significant gap in understanding FH in Africa is the lack of accurate prevalence estimates across most regions. Existing estimates are either extrapolated from European cohorts or derived from small, unrepresentative African samples. This is problematic, as population-specific genetic and environmental factors can dramatically influence both the clinical presentation and genetic basis of FH. The lower-than-expected prevalence of FH may also be attributed to the high genetic variation and epistatic interactions with lipid lowering variants in African populations. As shown in this PhD study, the presence of variants that lower LDL-C levels may mask the effect of those associated with higher LDL-C levels and FH. This conundrum plays a significant role in the complexity of the genetics of dyslipidaemia. Accurately determining prevalence requires the use of systematic population screening, validated clinical diagnostic criteria adapted to local contexts, and ideally, genetic confirmation of suspected cases. Additionally, family cascade screening, a key method for identifying individuals with undiagnosed conditions, is rarely implemented due to limited healthcare infrastructure. Without such comprehensive approaches, FH will remain underdiagnosed and undertreated, perpetuating health disparities in cardiovascular outcomes.

Ninety-six individuals with extreme high LDL-C levels were selected after excluding those with confounding factors such as elevated BMI or triglycerides, and being HIV positive, aiming to enrich for individuals with a likely genetic cause. Despite this stringent phenotypic filtering, the NGS panel yielded only three known FH pathogenic variants, all in *LDLR*. Several variants of uncertain significance (VUS) were identified, which could potentially contribute to disease in a monogenic or oligogenic manner, though their clinical relevance requires further functional validation. This lower-than-expected frequency of confirmed FH-causing variants may be explained by population-specific genetic architectures or the influence of environmental modifiers, such as diet and socioeconomic status, which can mask or exaggerate genetic effects. A substantial number of these variants were classified as VUS (49%) according to ACMG guidelines, despite having high pathogenicity scores and low MAFs, indicating that insufficient information is available on these variants. This raises the question: How can we assess the pathogenicity of variants detected in underrepresented populations? There are concerns regarding the classification of variants in African populations. Evidence provided by Manrai *et al.* in 2016 illustrated this concern by showing variants that were incorrectly classified due to differing frequencies in European datasets (Manrai Arjun K. *et al.*, no date).

Bope *et al.* also discussed the misinterpretation of variants in African populations, highlighting that in silico prediction tools are trained on European datasets and may therefore lead to either false positives or negatives when applied to African data (Bope *et al.*, 2019). There are multiple examples of studies where variants have undergone reclassification due to different MAF in non-European populations such as the study done by Rosamilia *et al.* (2024) where over 30% of variants initially classified as pathogenic for cardiomyopathy were reclassified as VUS due to high MAF in non-European populations (Rosamilia Michael B. *et al.*, 2024). Ultimately, functional validation of novel variants in African populations remains a critical gap. While some variants may appear pathogenic based on computational scores, confirmation through family segregation analyses or laboratory-based functional analyses is essential. However, such studies are resource-intensive and rarely possible in low-resource regions such as Africa, highlighting the need for increased investment in local genomics research.

4.3. Generally Lower LDL-C Levels in African populations

Multiple studies have indicated that ethnically admixed individuals, such as African Americans, with a greater proportion of African genetic ancestry tend to have lower LDL-C levels, suggesting a protective genetic signature. The Jackson Heart Study, which included over 4,400 African American participants, revealed that individuals in the highest quintiles of proportion of African ancestry had significantly lower LDL-C levels ($p = 0.027$) compared to those in the lowest quintiles (Deo *et al.*, 2009). The Heart of Soweto Study assessed lipid profiles in a cohort comprised of Indian, Caucasian, Mixed Ancestry and Black African individuals from South Africa with heart disease, and showed that individuals of African descent had the lowest LDL-C levels in the cohort (Sliwa *et al.*, 2012). A study by Bentley and Rotimi (2012) observed that each 10% increase in African ancestry correlated with approximately 1% decrease in serum triglyceride levels and a 0.7 mg/dL (0.039 mmol/L) increase in HDL-C, reinforcing evidence for a lipid profile potentially influenced by ancestry (Bentley and Rotimi, 2012).

Lower LDL-C levels in African populations may be influenced by environmental, lifestyle, and genetic factors. Two loss-of-function variants in *PCSK9,* p.Y142X and p.C679X, initially discovered in individuals of African ancestry, are more common in African populations (Cohen *et al.*, 2006; Hooper *et al.*, 2007). In this study, these variants were detected in 15 individuals with low LDL-C levels and only 2 individuals with high LDL-C levels.

Some variants in *APOB* have also been shown to be associated with lower LDL-C levels, causing a disorder called Familial Hypobetalipoproteinaemia (Burnett, Hooper and Hegele, 1993). Potential candidate *APOB* variants were detected in this study; however, the majority of likely pathogenic *APOB* variants were identified in both the high and low LDL-C phenotypic groups. Nevertheless, the high genetic variability and evidence of LDL-C lowering variants in African populations may suggest that genetics is an important contributing factor to the lower LDL-C levels on the continent.

Despite the generally lower levels of LDL-C, CVD continues to be a leading cause of death in Africa, with approximately 13% of deaths being attributed to CVD (Yuyun *et al.*, 2020). In Sub-Saharan Africa, cases of CVD have increased by 131.7% from 1990 to 2019 (Alhuneafat *et al.*, 2024), and it is expected that by 2030, CVD will be the leading cause of death on the continent (WHO, 2017). Several known risk factors contribute to CVD risk, including tobacco use, diabetes, obesity, hypertension and dyslipidaemia. With evidence showing lower levels of LDL-C and higher levels of HDL-C in Africa, it is also possible that other risk factors may play a larger role in CVD than dyslipidaemia. It has been demonstrated that the prevalence of each of the risk factors differs in Africa compared to high-income countries in Western Europe and North America. It is estimated that approximately 30% of individuals in SSA have hypertension (~40% in urban settings and ~20% in rural settings) in comparison to 20% in high-income countries. Smoking, obesity and diabetes also have a higher prevalence in SSA when compared to high-income countries; however, the opposite is shown with dyslipidaemia, with 25% in SSA and up to 40% in high-income countries (Yuyun *et al.*, 2020).

The differences in the prevalence of cardiovascular risk factors between African and non-African regions, such as Europe and North America, underscore the limitations of applying universally accepted CVD risk prediction models to African populations. These models often rely on assumptions and data that may not reflect the unique genetic, environmental, and lifestyle determinants present in African settings. The ongoing lack of robust, region-specific data on CVD and its risk factors further underscores the need for targeted research, particularly studies that aim to understand the relative contributions of both genetic and lifestyle factors to cardiovascular disease in African populations.

4.4. Precision Medicine in Africa

Precision medicine, which broadly involves tailoring prevention, diagnosis, and treatment strategies to an individual's genetic, environmental, and lifestyle characteristics, has rapidly

gained momentum in medicine globally. In the context of dyslipidaemia and FH, precision approaches offer the potential to improve identification of at-risk individuals, enhance stratification for treatment, and reduce CVD burden through early targeted interventions.

Diagnosis of FH typically follows well-validated clinical frameworks, such as the Dutch Lipid Clinic Network (DLCN), Simon Broome, or MEDPED criteria, which combine family history, physical examination, and LDL-C thresholds (Cartier and Goldberg, 2016). Additionally, clinical guidelines now include genetic confirmation via targeted gene panels or genotyping, which identifies pathogenic variants in 60–80% of cases, supporting cascade screening and early intervention (Martin *et al.*, 2016; Totoń-Żurańska *et al.*, 2023). A well-known target gene panel, LipidSeq, developed by the Robarts Institute of Research in Ontario, Canada, targets 69 genes and 185 single-nucleotide polymorphisms, and allows the evaluation of monogenic forms of FH (both SNPs and CNVs) as well as polygenic causes of hypercholesterolaemia (Dron *et al.*, 2020). Next-generation sequencing of whole exomes or genomes may be more appropriate in African populations than genotyping for a complex disorder such as FH, as novel variants can be detected, and the scope of detection includes CNVs; however, the cost per sample is likely significantly higher. While clinical criteria remain essential, especially when a pathogenic gene variant is absent, confirmatory genetic testing enhances diagnostic precision and aligns with international recommendations.

In low-resource settings like Africa, a definitive diagnosis of FH is often more challenging. In South Africa, clinical programs such as the Wits FIND-FH have begun to integrate both clinical and genetic screening. Among 700 participants, mostly of European ancestry, 68% met clinical FH criteria, and 60% of these had a pathogenic FH-causing variant confirmed, revealing three homozygous cases, several compound heterozygous cases, but it only included four Black African patients, highlighting both the program's success and the ongoing underdiagnosis in Black communities (Raal *et al.*, 2020). The four gene targeted sequencing panel (*LDLR*, *APOB*, *PCSK9* and *LDLRAP1*) employed in the FIND-FH study, as well as the 16-gene panel utilised in this study, although smaller and likely more cost-effective than broader panels like LipidSeq, represent a valuable step toward identifying novel variants in African populations. However, to comprehensively characterise the genetic architecture underlying both high and low LDL-C levels in these populations, more comprehensive approaches such as whole-exome sequencing (WES) or whole-genome sequencing (WGS) may be more appropriate. These methods allow for unbiased discovery of variants beyond the known lipid metabolism genes, capturing novel loci that may be overlooked by targeted gene panels limited to previously characterised

pathways. Effective treatment of FH relies heavily on timely and accurate diagnosis. Without the appropriate identification of individuals with FH, whether through clinical criteria, lipid profiling, or genetic testing, opportunities for early intervention are often missed. In many African settings, underdiagnosis of dyslipidaemia remains a significant barrier, resulting in patients presenting only after the onset of cardiovascular disease. Bridging the gap between diagnosis and treatment requires not only increased clinical awareness and infrastructure for lipid screening but also integration of genetic tools that can inform therapeutic decisions. As precision medicine advances globally, treatment strategies are becoming increasingly tailored to the genetic profiles of individual patients. This personalised approach holds significant promise for improving outcomes in African populations, where unique genetic diversity may influence drug response and disease risk. However, to fully realise the benefits of precision treatment in Africa, broader diagnostic coverage and affordable access to therapies must be prioritised.

Currently, statins are the first-line treatment for high LDL-C and thus CVD prevention, and have been shown to reduce the risk of CVD by 20% for every 1 mmol/L reduction in LDL-C (Yang *et al.*, 2022). In 2020, the global use of statins was measured at 68 defined daily doses (DDD) per 1,000 people, with 192 DDD per 1,000 people in high-income countries and only 28 DDD per 1,000 in low-income countries. The markedly lower use of statins in low-income settings, including many African countries, may be attributed to issues related to cost, limited access to healthcare, and a lack of awareness or diagnosis (Guadamuz, Shooshtari and Qato, 2022). Although statins are among the most widely prescribed medications globally, challenges with adherence and ongoing use remain significant (Banach *et al.*, 2016). One of the leading reasons patients discontinue statin therapy is the occurrence of perceived statin-associated muscle symptoms (SAMS), including myopathy (Toth *et al.*, 2018; Rosenson, 2021). Additional reported side effects include cognitive disturbances, liver toxicity, haemorrhagic stroke, and kidney impairment (Stroes *et al.*, 2015; Banach *et al.*, 2016). While these adverse events can contribute to discontinuation, strong evidence of a causal relationship exists only for SAMS, transient increases in alanine aminotransferase levels, and a higher incidence of newly diagnosed diabetes (Stroes *et al.*, 2015). The International Lipid Expert Panel (ILEP) defines statin intolerance (SI) as the inability to tolerate a statin dosage adequate for achieving necessary cardiovascular risk reduction (Banach *et al.*, 2016). Although the prevalence of SI is only 9% globally, a meta-analysis found that African American ethnicity was significantly associated with SI (Bytyçi *et al.*, 2022). Pharmacogenomic testing (PGx) is utilised to prevent

adverse drug reactions and has recently been shown to be relevant with statin treatment (Boudreau, Yu and Johnson, 2010; Miller and Martin, 2016). Statin-related myopathy is estimated to occur in 10-15% of patients, and approximately 30% discontinue use of statins as a result (Canestaro, Austin and Thummel, 2014; Ganga, Slim and Thompson, 2014; Vrablik *et al.*, 2014; Hubáček *et al.*, 2015). For individuals unable to tolerate statins or those with insufficient response, alternative lipid-lowering therapies are increasingly important. These include ezetimibe and *PCSK9* inhibitors such as alirocumab and evolocumab, as well as the siRNA inclisiran. Ezetimibe is a lipid-lowering agent that works by selectively inhibiting the Niemann-Pick C1-Like 1 (NPC1L1) protein, a transporter responsible for cholesterol uptake in the small intestine and is usually administered alongside statins, showing up to 39% decrease in LDL-C levels (Pisciotta *et al.*, 2006; Hamilton-Craig *et al.*, 2010). *PCSK9* inhibitors, have shown a profound reduction in LDL-C and improved cardiovascular outcomes, especially in individuals with FH (Sabatine *et al.*, 2017), however, their high cost limits accessibility in many African settings.

Precision approaches may enable clinicians to identify patients who are more likely to benefit from specific therapies based on their genetic, metabolic, and lifestyle profiles. For example, genotyping for polymorphisms in genes such as *SLCO1B1* can predict the risk of statin-related myopathy, allowing tailored statin prescriptions by adjusting dose or utilising alternative therapy (Jansen *et al.*, 2023). Additionally, *APOE* genotyping may help stratify individuals based on their lipid response to diet or medication, and it has been shown that ε2 carriers generally show more pronounced reductions in lipid levels, indicating greater responsiveness to treatment, while ε4 carriers show a comparatively weaker response (Asiimwe *et al.*, 2024). Precision medicine also supports PGS-guided interventions that help predict CVD risk, improving early intervention strategies (Li *et al.*, 2024).

In Africa, the integration of precision medicine remains in its early stages. Limitations include under-representation in genomic datasets, limited access to sequencing technologies, and lack of population-specific clinical guidelines. The differences in genomics between Europeans and Africans can be illustrated with multiple examples. The variant, *SLCO1B1*5* (rs4149056), has been extensively studied and is associated with myopathy; however, it has been shown that in Africa, the variant rs59502379 is more common and is associated with a 2.9-fold increase in statin-induced myopathy (Haldar *et al.*, 2025). The allele frequency of the *APOE* ε2 variant is higher in African populations than in European populations (0.11 vs. 0.08), with a similar pattern observed for the ε4 variant (Chen *et al.*, 2024). This study (Chapter 3) has shown that

PGSs, for example, developed in European populations are not as accurate at predicting disease risk, specifically higher LDL-C levels, than those developed in studies with African ancestry participants. It was also demonstrated that there are geographic region differences in the predictive accuracy of PGS within Africa. The R^2 values observed in this study are notably higher (as high as 20% variance explained) than those reported in previous work. For example, Choudhury *et al*. (2022) found that their PGS for LDL-C explained only 7% of the variance in a Sub-Saharan African cohort, including AWI-Gen. Their 43-SNP score, developed from a Ugandan population, outperformed scores derived from European and multi-ancestry cohorts, yet still underperformed compared to a PGS derived from a broader African-ancestry population—likely due to the extensive genetic diversity across the continent (Choudhury *et al.*, 2022). Similarly, Kamiza *et al*. (2022) reported that a PGS developed from an African American cohort accounted for just 8.14% of the variance in a South African Zulu cohort and only 2.6% in an Ugandan cohort. That study also demonstrated that incorporating local ancestry information improved predictive power, emphasizing the importance of developing region-specific models rather than relying on a single pan-African approach (Kamiza *et al.*, 2022).

Local initiatives and cohort studies, such as AWI-Gen, provide a foundation for applying PGS and pharmacogenomic tools tailored to African genetic diversity. For precision medicine to be successful, more research and data are required, specifically for the implementation of PGS as predictive tools. As the cost of sequencing continues to fall and bioinformatics capacity expands, there is significant potential to close the treatment gap through genomic-informed, equitable lipid care.

4.5. Translating Genomic Research into Clinical Practice in Africa

While genomic medicine and precision healthcare are rapidly advancing in high-income settings, their clinical translation in low- and middle-income countries (LMICs), notably across Africa, remains limited. The continent's genetic diversity offers promise for enhancing disease risk stratification and tailored therapies; yet African populations remain underrepresented in genomic research, which may slow implementation (Owolabi, Adam and Adebiyi, 2023). In LMICs, implementation faces multiple challenges including high costs of sequencing and infrastructure, shortage of both bioinformatics and clinical genetics expertise, limited workforce training, and absence of supportive regulatory frameworks such as national polices that govern the use of genetic data (Sirisena and Dissanayake, 2019; Kamp *et al.*, 2022; Onywera *et al.*, 2024). As a result, most genetic services, such as screening for FH or PGx, are

not available outside research institutions, and precision public health remains an unattainable goal (B. Tata, A. Ambele and S. Pepper, 2020).

Research consortia, such as the Human, Heredity and Health in Africa (H3Africa) and the African Centre of Excellence for Genomics of Infectious Diseases (ACEGID) now known as the Institute of Genomics and Global Health, are leading efforts to build capacity, establish biobanks, and collect locally relevant genomic data in Africa. The H3Africa consortium continues to build capacity for research in Africa by supporting biobanks and research laboratories, and supporting bioinformatics in multiple countries (The H3Africa Consortium *et al.*, 2014; Mulder *et al.*, 2018; Tindana *et al.*, 2019). The ACEGID group, although focused on infectious diseases and pathogen identification, is building sequencing capacity and training laboratory scientists in Africa (Folarin, Happi and Happi, 2014).

However, the actual impact of these research initiatives will only be realised when genomic discoveries are translated into practical, affordable, and equitable clinical solutions. Achieving this requires robust national policies and laws that regulate the ethical integration of genetic information into healthcare, as well as collaborations among academic institutions, healthcare systems, and governments. Additionally, sustained investment in infrastructure, workforce development, and public education is necessary. Without these, the insights gained from African genomic research risk remaining confined to academic settings, rather than improving the diagnosis, prevention, and treatment of diseases for African populations as a whole. A strategic, long-term commitment from both African governments and international stakeholders will be essential to bridge the gap between research and real-world clinical benefits.

4.6. Limitations and Future Work

The limitations of this work include the small sample size for the NGS gene panel sequencing group and the relatively modest AWI-Gen study cohort from four African countries. A larger and more diverse cohort, including more African countries, would create an opportunity for further variant discovery and knowledge generation. The participants were also cohort based, not individuals clinically diagnosed with FH. Prevalence of FH in African populations could not be accurately estimated due to the small, enriched group of participants with high and low LDL-C levels for the NGS gene panel study; however, future work should aim to calculate an accurate prevalence of FH in Africa and perhaps even within different African regions. The sequencing panel can also be more inclusive of genes involved in lipid metabolism, not

necessarily just those affecting LDL-C levels. For African populations whole-genome sequencing approaches would be ideal, as it allows for the detection of novel potentially pathogenic variants and potential variants in other genes that may affect lipid metabolism and has the added advantage of more accurately assessing structural variants. Family cascade screening, as well as functional studies, should be performed to determine the effects of novel variants. Missense variants may also be further assessed to determine effect on protein structure utilizing *in silico* tools such as Missense 3D (Ittisoponpisan *et al.,* 2019). A limitation for the second section of this dissertation is that PGSs were calculated using PLINK, a relatively basic tool, and it may be possible that using more advanced PGS building methods could further improve predictive performance (Vilhjálmsson *et al.*, 2015). Common tools include PRSice-2, which automates clumping and thresholding and LDpred which uses a Bayesian approach that utilizes linkage disequilibrium information (Choi and O'Reilly, 2019).

PGSs are increasingly being explored as tools for improving disease prediction and stratified prevention strategies. Internationally, PGS have shown promise in identifying individuals at elevated risk for coronary artery disease (CAD), with one study demonstrating that individuals in the highest percentiles of a PGS distribution can have up to a threefold increased risk of disease—equivalent to monogenic risk levels (Khera *et al.*, 2018). However, the transferability of PGS across populations remains a major challenge. Most existing PGS are developed using European-ancestry datasets, limiting their predictive accuracy in non-European populations attributed to different allele frequencies, linkage disequilibrium structures, and environmental modifiers. In Africa, where genomic diversity is highest globally, PGS must be locally validated or constructed using African-specific discovery datasets. Efforts to improve the equity of genomic medicine, such as the Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium, are aiming to address this gap by improving methodological frameworks and increasing the representation of global populations in PGS development (Kullo *et al.*, 2024). As bioinformatics capacity and whole-genome sequencing efforts increase across Africa, there is significant potential for the development of accurate, ancestry-informed PGS that can aid in precision prevention of dyslipidaemia and cardiovascular disease. However, their implementation must be guided by ethical frameworks, local validation, and integration into appropriate health systems infrastructure.

4.7. Conclusion

This dissertation provides an in-depth investigation into the genetic architecture of LDL-C in continental African populations, revealing both the promise and complexity of translating

genomic insights into clinical benefit. Integrating rare variant discovery with polygenic risk score evaluation across the diverse AWI-Gen cohort, this study demonstrates that the genetic underpinnings of dyslipidaemia in Africa are distinct and multifactorial, shaped by ancestry, regional variation, and environmental modifiers. The findings highlight that tools and thresholds derived from non-African populations often fall short when applied in African contexts, highlighting the urgent need for locally validated, ancestry-informed approaches.

Despite identifying a limited number of pathogenic monogenic FH-causing variants in individuals with extreme LDL-C levels, the presence of numerous variants of uncertain significance underscores a broader challenge: current variant interpretation frameworks are ill-equipped to assess genetic variation in African populations. This reinforces the necessity for functional studies, inclusive reference datasets, and equitable classification systems. Similarly, the modest performance of European-derived PGSs in predicting LDL-C levels, particularly when stratified by geographic region, suggests the need for the development of African-centric scores to enhance predictive utility and clinical relevance.

The lower average LDL-C levels observed in African populations contrast sharply with the high and rising burden of cardiovascular disease, suggesting that traditional lipid-based risk models may underperform in African settings. Precision medicine presents an opportunity to address these disparities through personalised risk assessment and targeted treatment. However, systemic barriers, including limited access to genetic testing, under diagnosis of FH, and inequities in treatment availability, must be overcome. The broader implementation of genomic medicine in Africa will require sustained investment in research infrastructure, workforce training, policy development, and the establishment of ethical frameworks.

In conclusion, this dissertation underscores the importance of centring African genomic data in global precision medicine efforts. As sequencing becomes more affordable and bioinformatics capacity expands, Africa is well-positioned to lead the next phase of equitable, genomics-driven healthcare, provided that research findings are translated into affordable and accessible, real-world interventions. To realise this vision, future work must include expanded genomic screening, variant functional validation, and integration of genetic tools into public health strategies. Only then can the promise of precision medicine be fully realised for African populations.

4.8. References

Alhuneafat, L. *et afacl.* (2024) 'Burden of cardiovascular disease in Sub-Saharan Africa, 1990-2019: An analysis of the Global Burden of Disease Study.', *Current problems in cardiology*, 49(6), p. 102557. Available at: https://doi.org/10.1016/j.cpcardiol.2024.102557.

Asiimwe, I. *et al.* (2024) *APOE genotype and the effect of statins: a systematic review and meta-analysis*. Available at: https://doi.org/10.1101/2024.12.13.24318973.

B. Tata, E., A. Ambele, M. and S. Pepper, M. (2020) 'Barriers to Implementing Clinical Pharmacogenetics Testing in Sub-Saharan Africa. A Critical Review', *Pharmaceutics*, 12(9). Available at: https://doi.org/10.3390/pharmaceutics12090809.

Banach, M. *et al.* (2016) 'Statin non-adherence and residual cardiovascular risk: There is need for substantial improvement.', *International journal of cardiology*, 225, pp. 184–196. Available at: https://doi.org/10.1016/j.ijcard.2016.09.075.

Bentley, A.R. and Rotimi, C.N. (2012) 'Interethnic Variation in Lipid Profiles: Implications for Underidentification of African-Americans at risk for Metabolic Disorders.', *Expert review of endocrinology & metabolism*, 7(6), pp. 659–667. Available at: https://doi.org/10.1586/eem.12.55.

Bope, C.D. *et al.* (2019) 'Dissecting in silico Mutation Prediction of Variants in African Genomes: Challenges and Perspectives.', *Frontiers in genetics*, 10, p. 601. Available at: https://doi.org/10.3389/fgene.2019.00601.

Boudreau, D.M., Yu, O. and Johnson, J. (2010) 'Statin use and cancer risk: a comprehensive review.', *Expert opinion on drug safety*, 9(4), pp. 603–621. Available at: https://doi.org/10.1517/14740331003662620.

Burnett, J.R., Hooper, A.J. and Hegele, R.A. (1993) 'APOB-Related Familial Hypobetalipoproteinemia.', in M.P. Adam et al. (eds) *GeneReviews(®)*. Seattle (WA): University of Washington, Seattle.

Bytyçi, I. *et al.* (2022) 'Prevalence of statin intolerance: a meta-analysis.', *European heart journal*, 43(34), pp. 3213–3223. Available at: https://doi.org/10.1093/eurheartj/ehac015.

Canestaro, W.J., Austin, M.A. and Thummel, K.E. (2014) 'Genetic factors affecting statin concentrations and subsequent myopathy: a HuGENet systematic review', *Genetics in Medicine*, 16(11), pp. 810–819. Available at: https://doi.org/10.1038/gim.2014.41.

Cartier, J.L. and Goldberg, A.C. (2016) 'Familial Hypercholesterolemia: Advances in Recognition and Therapy.', *Progress in cardiovascular diseases*, 59(2), pp. 125–134. Available at: https://doi.org/10.1016/j.pcad.2016.07.006.

Chen, S. *et al.* (2024) 'A genomic mutational constraint map using variation in 76,156 human genomes', *Nature*, 625(7993), pp. 92–100. Available at: https://doi.org/10.1038/s41586-023-06045-0.

Choi, S.W. and O'Reilly, P.F. (2019) 'PRSice-2: Polygenic Risk Score software for biobank-scale data', *GigaScience*, 8(7), p. giz082. Available at: https://doi.org/10.1093/gigascience/giz082.

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.

Cohen, J.C. *et al.* (2006) 'Sequence variations in PCSK9, low LDL, and protection against coronary heart disease.', *The New England journal of medicine*, 354(12), pp. 1264–1272. Available at: https://doi.org/10.1056/NEJMoa054013.

Deo, R.C. *et al.* (2009) 'Genetic Differences between the Determinants of Lipid Profile Phenotypes in African and European Americans: The Jackson Heart Study', *PLOS Genetics*, 5(1), p. e1000342. Available at: https://doi.org/10.1371/journal.pgen.1000342.

Dron, J.S. *et al.* (2020) 'Six years' experience with LipidSeq: clinical and research learnings from a hybrid, targeted sequencing panel for dyslipidemias', *BMC Medical Genomics*, 13(1), p. 23. Available at: https://doi.org/10.1186/s12920-020-0669-2.

Folarin, O.A., Happi, A.N. and Happi, C.T. (2014) 'Empowering African genomics for infectious disease control.', *Genome biology*, 15(11), p. 515. Available at: https://doi.org/10.1186/s13059-014-0515-y.

Ganga, H.V., Slim, H.B. and Thompson, P.D. (2014) 'A systematic review of statin-induced muscle problems in clinical trials', *American Heart Journal*, 168(1), pp. 6–15. Available at: https://doi.org/10.1016/j.ahj.2014.03.019.

Guadamuz, J.S., Shooshtari, A. and Qato, D.M. (2022) 'Global, regional and national trends in statin utilisation in high-income and low/middle-income countries, 2015-2020.', *BMJ open*, 12(9), p. e061350. Available at: https://doi.org/10.1136/bmjopen-2022-061350.

Haldar, T. *et al.* (2025) 'SLCO1B1 Functional Variants, Bilirubin, Statin-Induced Myotoxicity, and Recent Sub-Saharan African Ancestry: A Precision Medicine Health Equity Study.', *Clinical pharmacology and therapeutics*, 117(6), pp. 1696–1705. Available at: https://doi.org/10.1002/cpt.3624.

Hamilton-Craig, I. *et al.* (2010) 'Combination therapy of statin and ezetimibe for the treatment of familial hypercholesterolemia.', *Vascular health and risk management*, 6, pp. 1023–1037. Available at: https://doi.org/10.2147/VHRM.S13496.

Hooper, A.J. *et al.* (2007) 'The C679X mutation in PCSK9 is present and lowers blood cholesterol in a Southern African population.', *Atherosclerosis*, 193(2), pp. 445–448. Available at: https://doi.org/10.1016/j.atherosclerosis.2006.08.039.

Hubáček, J.A. *et al.* (2015) 'SLCO1B1 polymorphism is not associated with risk of statin-induced myalgia/myopathy in a Czech population.', *Medical science monitor : international medical journal of experimental and clinical research*, 21, pp. 1454–1459. Available at: https://doi.org/10.12659/MSM.893007.

Ittisoponpisan, S., *et al*. (2019). Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? Journal of Molecular Biology, 431(11), 2197-2212. doi:10.1016/j.jmb.2019.04.009.

Jansen, M.E. *et al.* (2023) 'Predictive Value of SLCO1B1 c.521T>C Polymorphism on Observed Changes in the Treatment of 1136 Statin-Users.', *Genes*, 14(2). Available at: https://doi.org/10.3390/genes14020456.

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.

Kamp, M. *et al.* (2022) 'Clinicians' Perceptions towards Precision Medicine Tools for Cardiovascular Disease Risk Stratification in South Africa', *Journal of Personalized Medicine*, 12(9), p. 1360. Available at: https://doi.org/10.3390/jpm12091360.

Khera, A.V. *et al.* (2018) 'Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations', *Nature Genetics*, 50(9), pp. 1219–1224. Available at: https://doi.org/10.1038/s41588-018-0183-z.

Kullo, I.J. *et al.* (2024) 'The PRIMED Consortium: Reducing disparities in polygenic risk assessment', *American Journal of Human Genetics*, 111(12), pp. 2594–2606. Available at: https://doi.org/10.1016/j.ajhg.2024.10.010.

Li, L. *et al.* (2024) 'Integration of a polygenic score into guideline-recommended prediction of cardiovascular disease.', *European heart journal*, 45(20), pp. 1843–1852. Available at: https://doi.org/10.1093/eurheartj/ehae048.

Manrai Arjun K. *et al.* (no date) 'Genetic Misdiagnoses and the Potential for Health Disparities', *New England Journal of Medicine*, 375(7), pp. 655–665. Available at: https://doi.org/10.1056/NEJMsa1507092.

Martin, R. *et al.* (2016) 'Genetic diagnosis of familial hypercholesterolaemia using a rapid biochip array assay for 40 common LDLR, APOB and PCSK9 mutations.', *Atherosclerosis*, 254, pp. 8–13. Available at: https://doi.org/10.1016/j.atherosclerosis.2016.09.061.

Miller, P.E. and Martin, S.S. (2016) 'Approach to Statin Use in 2016: an Update.', *Current atherosclerosis reports*, 18(5), p. 20. Available at: https://doi.org/10.1007/s11883-016-0578-1.

Mulder, N. *et al.* (2018) 'H3Africa: current perspectives', *Pharmacogenomics and Personalized Medicine*, 11, pp. 59–66. Available at: https://doi.org/10.2147/PGPM.S141546.

Nonterah, E.A. *et al.* (2025) 'Obesity phenotypes and dyslipidemia in adults from four African countries: An H3Africa AWI-Gen study.', *PloS one*, 20(1), p. e0316527. Available at: https://doi.org/10.1371/journal.pone.0316527.

Onywera, H. *et al.* (2024) 'Boosting pathogen genomics and bioinformatics workforce in Africa.', *The Lancet. Infectious diseases*, 24(2), pp. e106–e112. Available at: https://doi.org/10.1016/S1473-3099(23)00394-8.

Owolabi, P., Adam, Y. and Adebiyi, E. (2023) 'Personalizing medicine in Africa: current state, progress and challenges.', *Frontiers in genetics*, 14, p. 1233338. Available at: https://doi.org/10.3389/fgene.2023.1233338.

Pisciotta, L. *et al.* (2006) 'Additive effect of mutations in LDLR and PCSK9 genes on the phenotype of familial hypercholesterolemia.', *Atherosclerosis*, 186(2), pp. 433–440. Available at: https://doi.org/10.1016/j.atherosclerosis.2005.08.015.

Raal, F.J. *et al.* (2020) 'Cascade Screening for Familial Hypercholesterolemia in South Africa', *Arteriosclerosis, Thrombosis, and Vascular Biology*, 40(11), pp. 2747–2755. Available at: https://doi.org/10.1161/ATVBAHA.120.315040.

Rosamilia Michael B. *et al.* (2024) 'Underrepresentation of Diverse Ancestries Drives Uncertainty in Genetic Variants Found in Cardiomyopathy-Associated Genes', *JACC: Advances*, 3(2), p. 100767. Available at: https://doi.org/10.1016/j.jacadv.2023.100767.

Rosenson, R.S. (2021) 'Existing and emerging therapies for the treatment of familial hypercholesterolemia.', *Journal of lipid research*, 62, p. 100060. Available at: https://doi.org/10.1016/j.jlr.2021.100060.

Sabatine, M.S. *et al.* (2017) 'Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease', *The New England Journal of Medicine*, 376(18), pp. 1713–1722. Available at: https://doi.org/10.1056/NEJMoa1615664.

Sirisena, N.D. and Dissanayake, V.H.W. (2019) 'Strategies for Genomic Medicine Education in Low- and Middle-Income Countries.', *Frontiers in genetics*, 10, p. 944. Available at: https://doi.org/10.3389/fgene.2019.00944.

Sliwa, K. *et al.* (2012) 'Different lipid profiles according to ethnicity in the Heart of Soweto study cohort of de novo presentations of heart disease.', *Cardiovascular journal of Africa*, 23(7), pp. 389–395. Available at: https://doi.org/10.5830/CVJA-2012-036.

Smyth, N., Ramsay, M. and Raal, F.J. (2018) 'Population specific genetic heterogeneity of familial hypercholesterolemia in South Africa.', *Current opinion in lipidology*, 29(2), pp. 72–79. Available at: https://doi.org/10.1097/MOL.0000000000000488.

Stroes, E.S. *et al.* (2015) 'Statin-associated muscle symptoms: impact on statin therapy-European Atherosclerosis Society Consensus Panel Statement on Assessment, Aetiology and Management.', *European heart journal*, 36(17), pp. 1012–1022. Available at: https://doi.org/10.1093/eurheartj/ehv043.

Talmud, P.J. *et al.* (2013) 'Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: a case-control study.', *Lancet (London, England)*, 381(9874), pp. 1293–1301. Available at: https://doi.org/10.1016/S0140-6736(12)62127-8.

The H3Africa Consortium *et al.* (2014) 'Enabling the genomic revolution in Africa', *Science*, 344(6190), pp. 1346–1348. Available at: https://doi.org/10.1126/science.1251546.

Tindana, P. *et al.* (2019) 'Engaging research ethics committees to develop an ethics and governance framework for best practices in genomic research and biobanking in Africa: the H3Africa model', *BMC Medical Ethics*, 20(1), p. 69. Available at: https://doi.org/10.1186/s12910-019-0398-2.

Toth, P.P. *et al.* (2018) 'Management of Statin Intolerance in 2018: Still More Questions Than Answers.', *American journal of cardiovascular drugs : drugs, devices, and other interventions*, 18(3), pp. 157–173. Available at: https://doi.org/10.1007/s40256-017-0259-7.

Totoń-Żurańska, J. *et al.* (2023) 'Targeted sequencing of a gene panel in patients with familial hypercholesterolemia from Southern Poland.', *Polish archives of internal medicine*, 133(6), p. 16417. Available at: https://doi.org/10.20452/pamw.16417.

Vilhjálmsson, B.J. *et al.* (2015) 'Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores', *The American Journal of Human Genetics*, 97(4), pp. 576–592. Available at: https://doi.org/10.1016/j.ajhg.2015.09.001.

Vrablik, M. *et al.* (2014) 'Statin-associated myopathy: from genetic predisposition to clinical management.', *Physiological research*, 63(Suppl 3), pp. S327-334. Available at: https://doi.org/10.33549/physiolres.932865.

Wang, J. *et al.* (2016) 'Polygenic Versus Monogenic Causes of Hypercholesterolemia Ascertained Clinically.', *Arteriosclerosis, thrombosis, and vascular biology*, 36(12), pp. 2439–2445. Available at: https://doi.org/10.1161/ATVBAHA.116.308027.

WHO (2017) *Cardiovascular diseases (CVDs)*, *Cardiovascular diseases (CVDs)*.

Yang, X.H. *et al.* (2022) 'Statin use and the risk of CVD events, stroke, and all-cause mortality in patients with diabetes: A systematic review and meta-analysis.', *Nutrition, metabolism, and cardiovascular diseases : NMCD*, 32(11), pp. 2470–2482. Available at: https://doi.org/10.1016/j.numecd.2022.07.018.

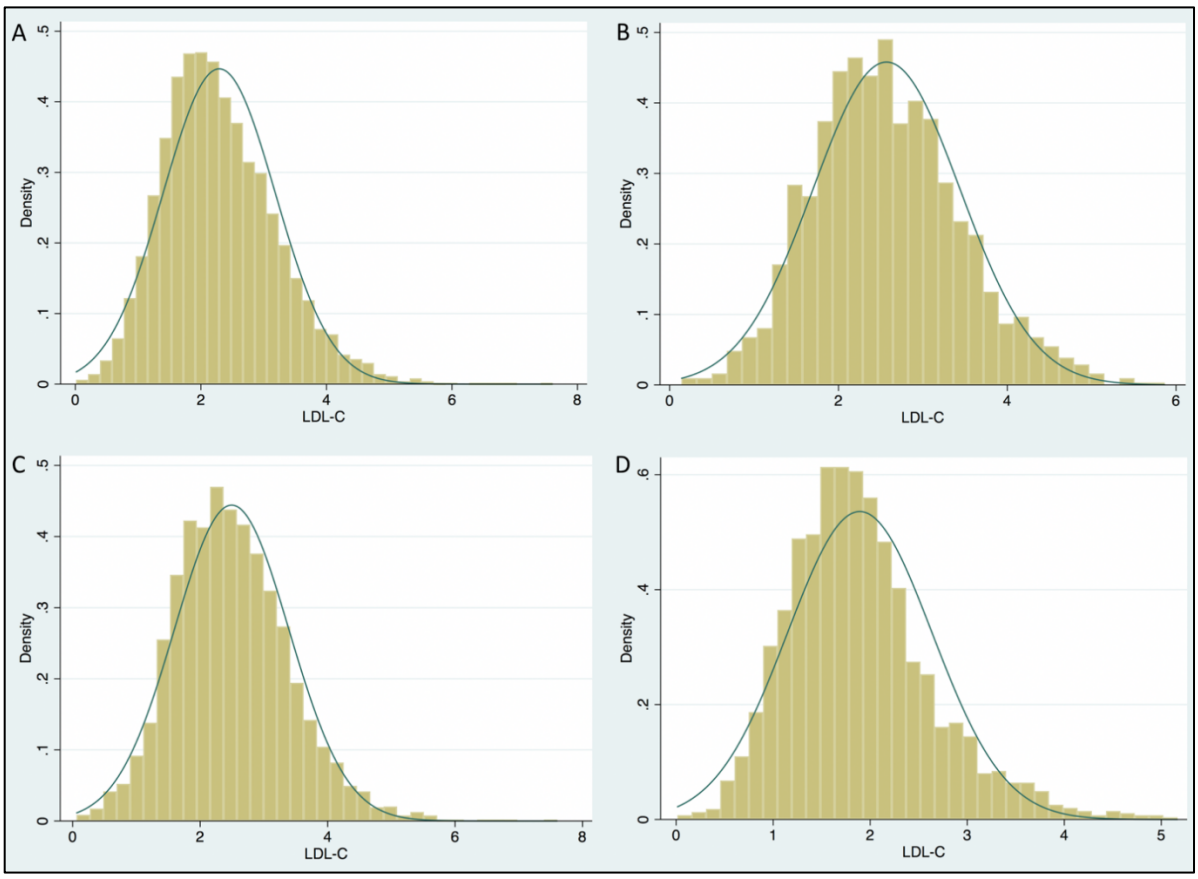
Yuyun, M.F. *et al.* (2020) 'Cardiovascular Diseases in Sub-Saharan Africa Compared to High-Income Countries: An Epidemiological Perspective.', *Global heart*, 15(1), p. 15. Available at: https://doi.org/10.5334/gh.403.

Supplementary Material

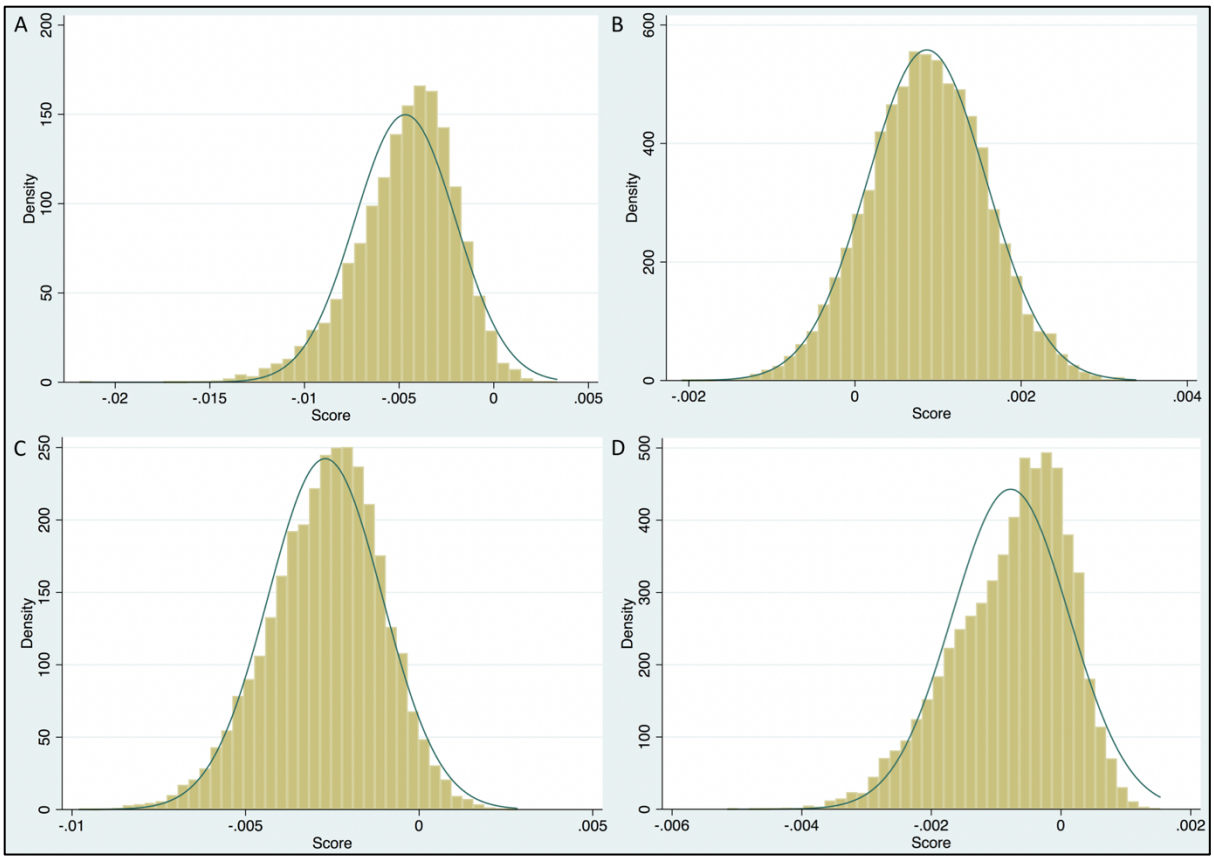
Chapter 2

Supplementary Table 1: List of genes in targeted sequencing panel

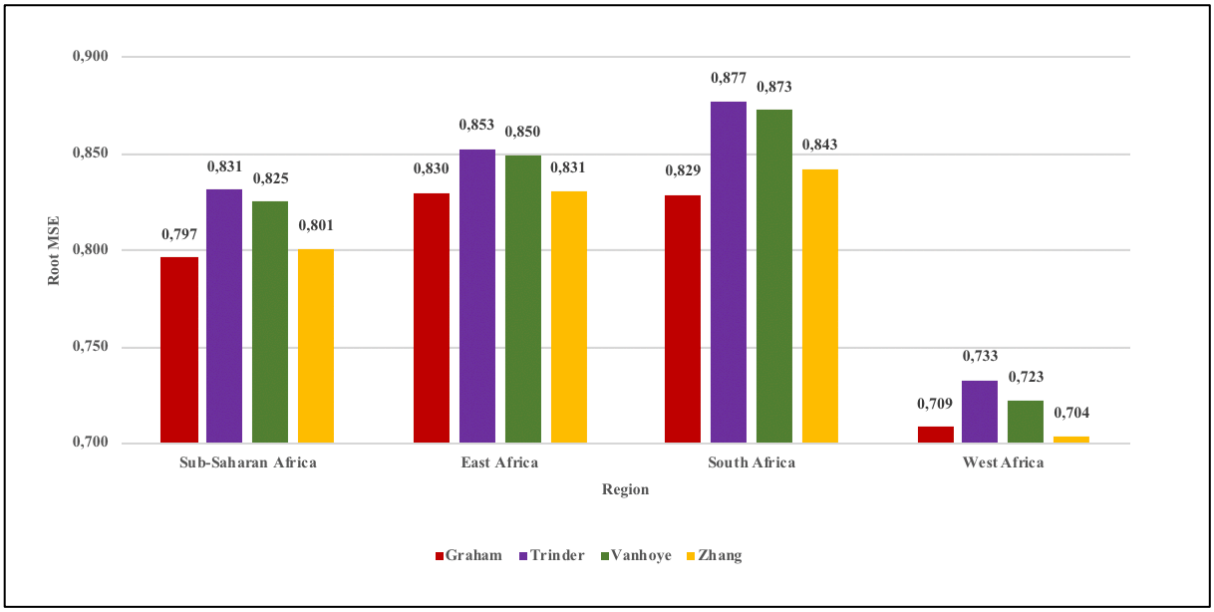
Gene	Protein
LDLR	Low Density Lipoprotein Receptor
APOB	Apolipoprotein A
PCSK9	Proprotein Convertase Subtilisin/Kexin Type 9
STAP1	Signal-transducing adaptor protein 1
APOE	Apolipoprotein E
LDLRAP1	Low density lipoprotein receptor adaptor protein 1
LIPA	Lysosomal acid lipase
ABCG5	ATP-binding cassette sub-family G member 5
ABCG8	ATP-binding cassette sub-family G member 8
SORT1	Sortilin
MYLIP	Myosin regulatory light chain interacting protein
NPC1L1	NPC1 like intracellular cholesterol transporter
MTTP	Microsomal triglyceride transfer protein
SAR1B	Secretion Associated ras related GTPase 1B
ANGPTL3	Angiopoietin-related protein 3
GATB	Glutamyl- TRNA Amidotransferase Subunit B



Supplementary Figure 1: The distribution of LDL-C (mmol/L) for A: All Regions in AWI-Gen, B: East Africa, C: South Africa and D: West Africa



Supplementary Figure 2: The distribution of A: Graham Score, B: Trinder Score, C: Vanhoye Score and D: Zhang Score across the AWI-Gen cohort



Supplementary Figure 3: Comparison of the root mean squared error for each PGS in each region in Africa

Appendices

Ethics Certificates

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

Office of the Deputy Vice-Chancellor (Research & Post Graduate Affairs)

TO: Ms N Smyth
School of Pathology
Department of Human Genetics
Sydney Brenner Institute for Molecular Bioscience
Medical School
University

E-mail: Natalie.Smyth@wits.ac.za

CC: Supervisor: Profs M Ramsay & FJ Raal; Dr D Sengupta
<Michele.Ramsay@wits.ac.za>
and <HREC-Medical.ResearchOffice@wits.ac.za>

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

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

DATE: 2020/01/02

REF: *R14/49*

PROTOCOL NO: **M1911192** *(This is your ethics application study reference number. Please quote this reference number in all correspondence relating to this study)*

PROJECT TITLE: *Understanding the genetic basis of extreme levels of high and low LDL-cholesterol in 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 the Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Ms N Smyth

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M1911192

NAME: (Principal Investigator)	Ms N Smyth
DEPARTMENT:	School of Pathology Department of Human Genetics Sydney Brenner Institute for Molecular Bioscience Medical School University
PROJECT TITLE:	Understanding the genetic basis of extreme levels of high and low LDL-cholesterol in African populations
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M170880
SUPERVISOR:	Profs M Ramsay & FJ Raal; Dr D Sengupta
APPROVED BY:	Dr N Naran, Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	2020/01/02

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary 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**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **November** and will therefore reports and re-certification will be due early in the month of **November** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

10 January 2025
Date

PLEASE QUOTE THE CLEARANCE CERTIFICATE NUMBER IN ALL ENQUIRIES

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

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

TO: Ms N Smyth

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

E-mail: Natalie.Smyth@wits.ac.za

CC: Supervisor: Professors M Ramsay & FJ Raal; Dr D Sengupta <Michele.Ramsay@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: 2025/03/28

REF: R14/49

PROTOCOL NO: **M2502100** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *Understanding the genetic basis of extreme levels of high and low LDL-cholesterol in 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 N Smyth

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M2502100

NAME: (Principal Investigator)	Ms N Smyth	**DEGREE:** PhD
SCHOOL:	School of Pathology Division of Human Genetics Faculty of Health Sciences	
PROJECT TITLE:	*Understanding the genetic basis of extreme levels of high and low LDL-cholesterol in African populations*	
DATE CONSIDERED:	Ad hoc	
DECISION:	Approved unconditionally	
CONDITIONS:	Renewal of M19/11/192 - expired on 2025/01/01	
NOTE:	For permission to use or contact student study participants, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>	
SUPERVISOR:	Professors M Ramsay & FJ Raal, Dr D Sengupta	
APPROVED BY:	Professor P Ruff, Chairperson, HREC (Medical)	
DATE OF APPROVAL:	2025/03/28	**EXPIRY DATE:** 2030/03/27

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

1 April 2025

Date

Plagiarism Declaration

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 Natalie Anne Smyth (Student number: 1331508) am a student registered for the degree of PhD in the academic year 2019.

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: 6 November 2025