

VARIANTS IN FOUR GENES ASSOCIATED WITH LIPID LEVELS: A STUDY IN AFRICAN POPULATIONS

Mahtaab Hayat

Student number: 603401

Supervisor: Professor Michèle Ramsay

Co-supervisor: Dr. Robyn Kerr



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

A Dissertation submitted to the Faculty of Health Sciences, University of the
Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of
Masters (Med) in Human Genetics.

Johannesburg, 2018

Declaration

I, Mahtaab Hayat, declare that this Dissertation is my own, unaided work. It is being submitted for the Degree of Masters in Human Genetics at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

A handwritten signature in purple ink, reading 'M Hayat', is positioned above a horizontal line.

Mahtaab Hayat

11th day of June 2018 in Johannesburg

Dedication

for everyone who believed in me

Presentations from this research report

1. P5 Africa Congress, 23-24 March 2016, Cape Town, South Africa
Poster title: Genetic associations on lipid levels in an African population – implications for familial hypercholesterolemia
Authors: Mahtaab Hayat and Michèle Ramsay
2. Wits Health Science Research Day, 1 September 2016, Johannesburg, South Africa
Poster title: Genetic variants in four genes associated with lipid levels: a study in African populations
Authors: Mahtaab Hayat, Michèle Ramsay and Robyn Kerr
3. Southern African Society for Human Genetics Biennial Congress, 13-16 August 2017, Durban, South Africa
Poster title: Genetic Variants in Four Genes Associated With Lipid Levels: A Study in African Populations
Authors: Mahtaab Hayat, Michèle Ramsay and Robyn Kerr

Abstract

Non-communicable diseases, including cardiovascular disease (CVD), are on the rise in African populations. High serum LDL cholesterol (LDL-C) levels is a risk factor for CVD, but the contribution of high LDL-C levels to CVD in African populations remains poorly understood. Genetic variation in the *LDLR*, *APOB*, *PCSK9* and *LDLRAP1* genes is known to be associated with alteration in LDL-C levels in many populations. This study aims to examine whether genetic variants in these four genes are associated with differing LDL-C levels in African populations, considering LDL-C as a polygenic trait. Publicly available African whole genome sequence data were interrogated, and variants were selected for genotyping using functional predictive bioinformatics tools. Participants (n=1000) from the AWI-Gen study were selected using a case-control study design based on clinical cut-offs of LDL-C levels (500 with LDL-C>3.5 mmol/l, 500 with LDL-C<1.1 mmol/l). Genotyping was carried out on 19 selected SNPs chosen from across the four genes. Logistic regression analysis revealed that, after adjusting for sex, fasting glucose levels, BMI and geographic region, the minor alleles at two SNPs remained significantly associated ($p<0.05$) with low LDL-C levels - *LDLRAP1* rs12071264, c.533-22A>G (OR 0.5866, $p<0.01$) and *APOB* rs6752026, c.433G>A (OR 0.6898, $p=0.04$). The minor alleles G and A, were associated with lower LDL-C levels, suggesting gain of function effects. The variant alleles at both loci are extremely rare in European populations (MAF<0.001) and this may explain why they have not previously been reported in LDL-C association studies. Since African populations, in general, have reduced LDL-C levels these variants could be African-specific LDL-C associated variants and may suggest a unique gene-environment interaction. Using a limited number of potentially functional variants in African populations and after extensive adjustment for potential covariates, significant associations were detected with variants in two of the four genes studied.

Acknowledgements

First and foremost, thank you to my two supervisors: Professor Michèle Ramsay and Dr Robyn Kerr, without whom I would have never managed to get through this MSc. Your expert guidance and advice were much appreciated. Thank you!

Professor F Raal, thank you for your advice and expertise that aided in guiding my study.

My K – you dealt with my absence and mood swings splendidly. Thank you.

My family – my Ammi, Apa, Bhai Jaan, Haleema Bibi and Barai Abba – for their constant support and encouragement.

Jorge, buddy. Thanks for your unwavering support and encouragement. You always had answers to my (mostly silly) questions concerning coding and Ensembl!

Micke, Heather and Michaela – you guys have been the best group of friends I could ever ask for.

Thank you to Phelelani and Shaun for all the bioinformatics help, for the early morning and late afternoon chats and for all the encouragement and advice.

Carl, your guidance and crash tutorials in things that would take hours to understand was so appreciated.

Dhriti and Ananyo – thank you for allowing me to constantly ask questions and providing sound answers. You guys are awesome!

Romy and JT, thanks for your answers to my stats and R questions!

Thank you to the rest of my colleagues at SBIMB for making it feel like home and for always graciously extending a helping hand whenever it was needed.

Andrew May, thank you for your valuable insight and logical advice provided for the progress and write up of my dissertation.

The rest of my friends and family: thanks for not giving up on me even when I went MIA for 2 years.

Many thanks to the examiners of my dissertation. Your input was so valuable.

Financial acknowledgements:

DST/NRF South African Research Chair Initiative (grant holder Ramsay),
University of the Witwatersrand, Faculty of Health Sciences Faculty Research
Committee and the National Health Laboratory Service.

Table of Contents

Declaration	i
Dedication	ii
Presentations from this research report	iii
Abstract	iv
Acknowledgements	v
Table of Contents	vi
List of figures	ix
List of tables	x
List of Abbreviations	xi
1. Introduction	1
1.1. Dyslipidaemia.....	3
1.2. Clinical relevance of dyslipidaemia.....	5
1.3. Treatment of dyslipidaemia	7
1.4. Genes associated with dyslipidaemia.....	8
1.4.1. Low Density Lipoprotein Receptor (<i>LDLR</i>)	12
1.4.2. Apolipoprotein B (ABOP)	12
1.4.3. Proprotein Convertase Subtilisin Kexin Type 9 (<i>PCSK9</i>).....	12
1.4.4. Low Density Lipoprotein Adaptor Protein 1 (<i>LDLRAP1</i>)	13
1.5. Variant frequencies in the common dyslipidaemia genes.....	13
1.6. Dyslipidaemia is the South African Context.....	14
1.7. Study Rationale	15
1.8. Aim and objectives	16
1.8.1. Aims.....	16
1.8.2. Objectives.....	16

1.9.	Study approach	17
2.	Methods and Materials	18
2.1.	Participants	20
2.2.	Candidate gene selection	22
2.3.	Whole genome sequences available for study	22
2.4.	Variant Selection and Functional Annotation.....	24
2.5.	Genotyping approach	25
2.6.	Data Analysis	28
2.6.1.	Multiple testing.....	28
2.7.	Case-Control Association Analysis.....	29
2.7.1.	Logistic regression.....	29
2.7.2.	Polygenic risk score	30
2.7.3.	Visualisation of results	30
3.	Results	33
3.1.	Participants	33
3.2.	Selection of SNPs based on function and frequency	38
3.3.	Genotyping	38
3.4.	Quality control	41
3.4.1.	Sample Quality Control	41
3.4.2.	SNP Quality Control.....	41
3.5.	Population allele frequencies	42
3.6.	Case-Control Association Analysis.....	45
3.6.1.	Logistic regression.....	45
3.6.2.	Polygenic Risk Score	50
4.	Discussion	52
4.1.	Characterisation of participant phenotype	53

4.2. Genetic Associations with LDL-C	54
4.2.1. Genetic association analysis	55
4.2.2. Gene-Environment interactions	59
4.2.3. Polygenic Risk Score	61
4.2.4. Visualisation of genotype association with LDL-C levels	62
4.3. LDL-C Levels in African Populations	63
4.4. Limitations of this study	63
4.5. Future research to understand the genetic contribution to LDL-C levels among Africans.....	66
5. Conclusion	68
6. References.....	70
7. Appendices	82
7.1. Appendix A: Plagiarism report and document	82
7.2. Appendix B: Ethics clearance certificate	84
7.3. Appendix C: Various tables	85

List of figures

Chapter 1

Figure 1.1: Visual representations of hypercholesterolaemic symptoms	6
Figure 1.2: Components in LDL-C endocytosis	11
Figure 1.3: Flow diagram of study approach	17

Chapter 2

Figure 2.1: Map displaying sites that AWI-Gen participants were selected from	19
Figure 2.2: Hypothetical data following normal distribution	21
Figure 2.3: Map of Africa showing countries from which WGS data was publicly available for analysis	23
Figure 2.4: Process of MassARRAY genotyping	27
Figure 2.5: Flow diagram summarising the filtering and association analysis carried out for this study.	32

Chapter 3

Figure 3.1: Graphs showing the distribution of LDL-C levels across AWI-Gen Participants.....	37
Figure 3.2: Flow diagram of variant filtering.....	38
Figure 3.3: Histograms showing frequencies of 14 variants that were genotyped and passed QC in each of the four genes	44
Figure 3.4: Flow diagram of study outline with results.....	47
Figure 3.5: Forest plot for the 14 SNPs that were genotyped and passed QC	48
Figure 3.6: Box and whisker plots of the 2 SNPs significantly associated with LDL-C levels after logistic regression	49
Figure 3.7: Correlation of the polygenic risk score (PRS) with LDL-C levels in 993 individuals.....	51

List of tables

Chapter 1

Table 1.1: List of genes found to be associated with total cholesterol levels	9
Table 1.2: Estimated frequencies of variants in genes that cause FH in populations of European origin adapted from Benn et al., 2016	11
Table 1.3: Prevalence of FH in South African populations and variant profiles.....	14

Chapter 2

Table 2.1: Numbers of samples with available low coverage WGS data from each black African population from KGP and from AGVP	22
---	----

Chapter 3

Table 3.1: Distribution of 1000 participants across AWI-Gen sites stratified by sex	33
Table 3.2: Phenotype characterisation of 1000 AWI-Gen individuals: sex, age, BMI, glucose levels and LDL-C cholesterol	34
Table 3.3: Phenotype characterisation of 1000 AWI-Gen participants separated by sex: LDL-C levels, age, BMI and glucose levels	35
Table 3.4: HIV status and treatment of 1000 AWI-Gen participants	36
Table 3.5: List of the 19 SNPs that were genotyped in AWI-Gen participants using the MassARRAY platform	39
Table 3.6: Results of genotyping for the 19 SNPs selected for genotyping	40
Table 3.7: Samples removed due to SNP genotype failure QC criteria	41
Table 3.8: SNPs removed due to failing quality control measures.....	41
Table 3.9: Allelic association of 14 SNPs with and without adjustment for multiple testing using FDR in 993 individuals	45
Table 3.10: Logistic regression with 14 SNPs, adjusted for covariates sex, BMI, glucose levels and region in 993 individuals.....	46

List of Abbreviations

AGVP	African Genome Variation Project
ARV	Antiretroviral
<i>APOB</i>	Apolipoprotein B gene
AWI-Gen	Africa Wits INDEPTH partnership for genomic studies
BMI	Body mass index
CADD	Combined Annotation Dependent Depletion
CI	Confidence interval
CVD	Cardiovascular disease
DLCN	Dutch Lipid Clinic Network
ENCODE	Encyclopedia of DNA Elements
ESN	Esan population of Nigeria
FDR	False discovery rate
FH	Familial hypercholesterolaemia
GoF	Gain of function
GRCh37	Genome Reference Consortium Human Build 37
GWD	Gambian in Western Division, The Gambia - Mandinka
H3Africa	Human Hereditary and Health in Africa
HDL-C	High density lipoprotein cholesterol
HMG-CoA	3-hydroxy-3-methylglutaryl-CoA
INDEPTH	International Network for the Demographic Evaluation of Populations and their Health in low- and middle-income countries
KGP	1000 Genomes Project
LD	Linkage disequilibrium
LDL-C	Low density lipoprotein cholesterol
<i>LDLR</i>	Low density lipoprotein receptor gene
LDLR	Low density lipoprotein receptor protein
<i>LDLRAP1</i>	Low density lipoprotein receptor adaptor protein 1 gene
LoF	Loss of function
LWK	Luhya in Webuye, Kenya

MAF	Minor allele frequency
MEDPED	Make Early Diagnosis to Prevent Early Death
MSL	Mende population of Sierra Leone
NCDs	Non-communicable diseases
OR	Odds ratio
<i>PCSK9</i>	Proprotein casein subtilisin kexin type 9 gene
PolyPhen2	Polymorphism Phenotyping: variant effect predictor programme v. 2
PRS	Polygenic risk score
QC	Quality control
SAP	Shrimp alkaline phosphatase
SBE	Single base extension
SBRG	Simon Broome Registry Group
SIFT	Sorting Intolerant From Tolerant: variant effect predictor
SNP	Single nucleotide polymorphism
SSA	Sub-Saharan Africa
TC	Total cholesterol
UTR	Untranslated region
VCF	Variant call format
VEP	Variant Effect Predictor
WGS	Whole genome sequences
YRI	Yoruba population of Nigeria

1. Introduction

The incidence of non-communicable diseases (NCDs) is on the rise in Africa as well as Southern Africa (Mayosi *et al.*, 2009), with an estimated increase of 10% by 2030 (World Health Organisation, 2015). One reason is that the marginalisation of NCDs has resulted in the lack of treatment and prevention for NCDs. This has contributed to an increase in NCDs (Mayosi *et al.*, 2009). Another contributory factor is the change of lifestyle experienced by many urbanising African communities: an increase in the intake of high calorie, Westernised food, accompanied by reduced physical activity (Mayosi *et al.*, 2009; Agyepong *et al.*, 2017). In 2014, NCDs accounted for 43% of total deaths in South Africa (World Health Organisation, 2014). In addition, it is now becoming apparent that underlying genetic factors also play a role in predisposing to NCDs.

The most prevalent NCD worldwide is cardiovascular disease (CVD) (Mayosi *et al.*, 2009). CVDs are disorders of the heart and blood vessels which include, but are not exclusive to, coronary heart disease, congenital heart disease and rheumatic heart disease. These diseases predispose individuals to heart attacks and strokes and can often lead to premature death if not treated early enough. CVDs are the leading cause of death worldwide (Sidney *et al.*, 2016; World Health Organisation, 2017), with an estimate of 17.7 million deaths in 2015 due to CVD globally. This is approximately 33% of all deaths worldwide (World Health Organisation, 2017).

South Africa, and Africa, face a unique situation in which there is a rapid rise in the health burden (due to communicable and non-communicable diseases), along with an increasing population size. The population of people under age 25 years in sub-Saharan Africa (SSA) is estimated to double to 450 million by 2050, placing a foreseeable large health burden on the continent attributed to the predicted rise in NCDs (Agyepong *et al.*, 2017). Studies have demonstrated that genetic variants play a role in conferring risk to NCDs (Wooster *et al.*, 1995; Williams *et al.*, 2008; Walley, Asher and Froguel, 2009). Therefore, research is imperative in characterising genetic diseases and genetic profiles in African populations to aid the pathway to preventative, and hopefully to personalised, medicine. The emphasis on genetic research forms part of the Lancet Commission on the future of health in SSA (Agyepong *et al.*, 2017), specifically to reduce the rise of NCDs and to provide treatment for priority CVDs.

Over 75% of deaths caused by CVDs in 2012 occurred in low- and middle-income countries (LMICs) (World Health Organisation, 2017). In 2015, 1.6 million deaths were attributed to CVD in Africa, and this number is expected to increase to 2.6 million by 2030 (Barr *et al.*, 2016). Death due to CVD is common in South Africa (Steyn and Fourie, 2007) and 38% of deaths related to NCDs were due to CVDs in 2013 (Keates *et al.*, 2017). Two main factors contributing to the steady rise of CVDs in Southern Africa are increased urbanisation and prevalent CVD risk factors, like obesity and hypertension. Socioeconomic status is changing for many, and as a result, so do eating habits and physical activity. Intake of foods that are higher in fats and sugars are increased and physical activity decreases, causing a rise in the risk of CVD (Mayosi *et al.*, 2009; Soko, Masimirembwa and Dandara, 2016). Studies have found that infectious diseases such as HIV/AIDS and TB can also increase risk of CVD. Given the high burden of HIV in African populations, this increased risk of CVDs is significant (Huaman *et al.*, 2015; Soko, Masimirembwa and Dandara, 2016).

In European populations, common risk factors for CVD include hypertension, dyslipidaemia (Wilson *et al.*, 1998), diabetes (Bonora and Muggeo, 2001) and obesity (Bastien *et al.*, 2014). However, the contribution of these factors to CVD in Africa is unknown. As a result, the importance of research in this area is crucial to enable informed health care decisions going forward. In this study we focus on one of the known risk factors for CVD: high levels of cholesterol, particularly low density lipoprotein cholesterol (LDL-C). As data is limited on both CVD and LDL-C levels in African populations, it is important to investigate LDL-C levels and determine whether they are a major contributing factor to CVD in these populations.

Cholesterol is a type of lipid molecule, synthesised naturally by all animal cells, but primarily by liver cells, and transported in the blood to sites in the body where it is needed. It plays an essential role in the animal cell membrane by controlling fluidity and maintaining the barrier between the environment and the cell. The body also uses cholesterol to create steroid hormones and bile acids, and it is essential in making up the myelin sheath of axons (Brown and Goldstein, 1986). Most importantly, elevated cholesterol in the blood can lead to CVD.

When physicians measure cholesterol/lipid levels, they usually look at triglycerides, LDL-C and high-density lipoprotein cholesterol (HDL-C). LDL and HDL are lipoproteins that carry cholesterol around the body. High levels of HDL-C are protective and are caused by positive changes in lifestyle: exercise, losing weight and reducing cigarette smoking. It is protective because HDL helps to carry cholesterol found in blood vessel walls back to the liver (Toth, 2005). In contrast, high levels of LDL-C have a deleterious effect by contributing to causing CVD.

High LDL-C levels are often caused by consumption of fatty, processed foods and physical inactivity, but can also be caused by predisposing genetic factors. A study done by Snieder, van Doornen and Boomsma, (1999) found that the heritability of LDL-C is more than 50%. A study conducted by Weiss *et al.*, (2006) reports similar heritability estimates in the range of 40-60%. Another study by Beekman *et al.*, (2002) shows that there was 83% heritability of LDL-C in a pair of monozygotic Dutch twins. Despite the high heritability, there is still variance that is not accounted for. In the general population, LDL-C levels have multifactorial origins (genetics plus the environment).

Mendelian or monogenic disorders involving high LDL-C are most often highly penetrant. This makes the identification of a causal or associated variant less complex than multifactorial disorders (Antonarakis and Beckmann, 2006). A common Mendelian disorder caused by high LDL-C levels is familial hypercholesterolaemia (FH) (Stecher and Hersh, 1949; Arnett and Shah, 2014).

1.1. Dyslipidaemia

Dyslipidaemia refers to abnormal lipid profiles, where an individual has lipid levels (LDL-C, HDL-C, triglycerides or total cholesterol (TC)) higher (LDL-C > 5 mmol/l) or lower (LDL-C < 3.5 mmol/l) than the normal range (Durrington, 2003). In this study the focus is on high and low levels of LDL-C, specifically, in the blood plasma. Dyslipidaemia and lipids will refer to high or low LDL-C levels in this study. Cholesterol is transported in the blood by lipoproteins. It moves into and out of cells using cellular mechanisms and specific proteins. Too much cholesterol in the bloodstream results in a build-up of plaque in artery walls which leads to CVD, therefore, the level of cholesterol in the bloodstream needs to be kept within an optimal range.

Several lipoproteins transport cholesterol in the blood, with the two most common being LDL and HDL. LDL-C binds to its receptor, low density lipoprotein receptor (LDLR), and this complex is internalised into the cell. Once inside the cell, LDL-C regulates cholesterol levels by inhibiting the production of cholesterol in cells by suppressing the cholesterol producing enzyme: 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) reductase (Brown, Faust and Goldstein, 1975). Various variants in the genes involved in cholesterol metabolism, namely *LDLR*, low density lipoprotein receptor adaptor protein (*LDLRAP1*), apolipoprotein B (*APOB*) and proprotein convertase subtilisin kexin type 9 (*PCSK9*), result in elevated levels of LDL-C in the plasma.

High levels of LDL-C, referred to as hypercholesterolaemia, over prolonged periods, have negative effects. For example, LDL-C can build up on artery walls to form plaques, restricting blood flow to and from the heart. In monogenic cases, e.g. FH, the build-up of LDL-C starts from birth and although asymptomatic, can present with end-stage consequences such as premature heart attacks or strokes (Wiegman *et al.*, 2015).

FH, an autosomal dominant trait, is the most common single gene disorder in the world (Genest, 2017), with between 14 and 34 million affected people worldwide (Nordestgaard *et al.*, 2013). It is characterised by high levels of LDL-C in the plasma (fasting LDL-C level of >5 mmol/l) (Catapano *et al.*, 2016). FH can be caused by variants in several different genes that affect the uptake of cholesterol by cells, or the metabolism of cholesterol once taken up by cells. Amongst the many types of familial hyperlipidaemias, FH is classified as type IIa, where variants affect levels of only LDL-C (Hegele, 2009). Other types of hyperlipidaemias affect different lipids, such as HDL-C, triglycerides and very dense lipoproteins.

The phenomenon of low levels of LDL-C in the blood is known as hypolipidaemia. While this form of dyslipidaemia is significantly more rare than hyperlipidaemia, it can be protective against CVD and can promote longevity (Hooper, van Bockxmeer and Burnett, 2005; Langsted *et al.*, 2016). Familial or genetic forms of hypolipidaemia are not well characterised due to many different genes having been associated with the disorder (Hooper, van Bockxmeer and Burnett, 2005). Specific types of variants, i.e. gain of function or loss of function variants in *PCSK9* specifically, are now known to cause low or high levels of LDL-C, respectively. Two of the most well described genes associated

with hypolipidaemia are *PCSK9* (Abifadel *et al.*, 2003) and *APOB* (Hooper, van Bockxmeer and Burnett, 2005; Schonfeld, Lin and Yue, 2005).

1.2. Clinical relevance of dyslipidaemia

There have been reports of loss of function variants in *PCSK9* causing low levels of LDL-C in different populations (Marais *et al.*, 2015). One loss-of-function variant (R46L) was identified in a Danish population (Langsted *et al.*, 2016). Two nonsense variants (Y142X and C679X) were identified in African Americans, with an observed 40% reduction in LDL-C levels in carriers (Cohen *et al.*, 2005). In a group of 653 Zimbabwean women, only the Y142X variant was identified that lowered levels of LDL-C (Hooper *et al.*, 2007). Individuals with low levels of LDL-C are at an advantage, as there is an 88% reduction in risk of CVD (Cohen *et al.*, 2006) and a 30% lower incidence of ischemic heart disease (Benn *et al.*, 2010; Langsted *et al.*, 2016). There are no particular symptoms associated with low levels of LDL-C.

On the other hand, high levels of LDL-C have deleterious effects in humans, having a direct correlation to CVD and death at early ages. The most dangerous aspect of high LDL-C is probably that there are no obvious symptoms. Often, individuals are only made aware of their high LDL-C status only once they, or close family members, have experienced an early CVD event such as cardiac arrest before the age of 55 (men) or 60 (women) (Nordestgaard *et al.*, 2013; Farnier *et al.*, 2017). Thus, it is imperative to diagnose hypercholesterolaemia as soon as possible, as cholesterol build-up begins in the foetus and increases over time, putting emphasis on preventative treatment and lifestyle changes (Wiegman *et al.*, 2015).

Familial hypercholesterolaemia (FH), an autosomal dominant trait, is diagnosed according to defined clinical criteria. The three most commonly used guidelines are defined by the Make Early Diagnosis to Prevent Early Death (MEDPED) programme, the Dutch Lipid Clinic Network (DLCN) and the Simon Broome Registry Group (SBRG). A patient is diagnosed with hypercholesterolaemia if they have:

- a. LDL-C > 4 mmol/l
- b. Family history of CVD
- c. Phenotypic presentations of hypercholesterolaemia

The phenotypic presentation of hypercholesterolaemia includes xanthomas, xanthelasmas, corneal arcus (figure 1.1) and premature CVD. However, these symptoms generally only present on patients with extremely high LDL-C levels, and in older patients (Brown and Goldstein, 1974). Heterozygous FH patients often experience CVD before the age of 55, and homozygous patients before the age of 20 (Nordestgaard *et al.*, 2013).

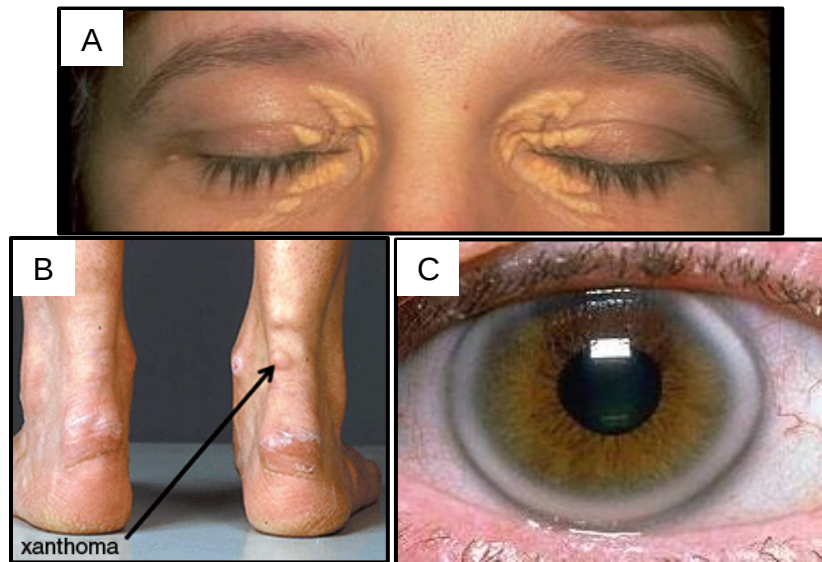


Figure 1.1: Visual representations of hypercholesterolaemic symptoms. A: Xanthelasma: fat deposits around the eye. B: Xanthoma: fat deposits around tendons. C: Corneal arcus: circular grey ring formed by fat deposits around iris. Images taken from: A: <https://en.wikipedia.org/wiki/Xanthoma>, B: <http://www.texasfootdoctor.org/xanthomas-of-the-achilles-tendon.html>, C): <http://www.iridology-swanssea.co.uk/corneal-arcus/>

There are several potential confounders that affect lipid levels, the most important being HIV infection, particularly relevant in SSA. HIV infection and its treatment are reported to have various effects on lipid levels. In a South African cohort from Limpopo, Vos *et al.* (2017) report on HIV infection being associated with low LDL-C levels. Anastos *et al.* (2007) found that women treated for HIV had higher levels of LDL-C. Ritonavir, indinavir/ritonavir and nelfinavir were specifically associated with high LDL-C, indicating that the type of treatment affects lipid levels. This in turn affects the chances of a patient experiencing CVD (Sandler *et al.*, 2014; Zhou *et al.*, 2015; Zidar *et al.*, 2015). These studies suggest that HIV infection and HIV treatment affect LDL-C levels independently. In addition, there is evidence to suggest that HIV also affects levels of insulin, where HIV

infected men more often have diabetes than women (Palmer *et al.*, 2014; Koethe *et al.*, 2016).

Diet, as an environmental factor, also affects lipid levels. A study carried out on a Japanese population showed that with conforming to a Western way of life, LDL-C levels increased (Egusa *et al.*, 1993). A group of Japanese between the ages of 30 and 69 years from Los Angeles and Hawaii were used as Westernised participants. Japanese participants of the same ages were recruited from Hiroshima, where a more traditional diet was followed. They found that the Japanese population that was Westernised had increased intake of animal protein, and decreased consumption of vegetable/plant protein. As a result, they had increased levels of LDL-C levels. Diabetes was also increased in the population that adopted the Westernised diet.

Obesity is also found to be a factor that affects LDL-C levels. A Canadian study of 16 000 individuals, aged 18 to 74 years, found an association between obesity and lipid levels, amongst other phenotypes. The study found that obesity was correlated with raised LDL-C levels (Reeder *et al.*, 1997). A study conducted in Thailand reports three times higher LDL-C levels in obese individuals compared to non-obese individuals (Kulanuwat *et al.*, 2015) and Magkos, Mohammed and Mittendorfer (2008) found a 50% increase in LDL-C levels in obese individuals compared to lean individuals.

It is understood that HIV infection, diet and obesity can act as confounders when studying LDL-C levels.

1.3. Treatment of dyslipidaemia

The revised world frequency for heterozygous FH is closer to 1 in 250 (Akioyamen *et al.*, 2017) individuals than the previously reported 1 in 500 individuals (Austin *et al.*, 2004) and this can be attributed to FH being largely underdiagnosed in earlier studies, or the definition having been revised in the more recent studies (Nordestgaard *et al.*, 2013; Watts *et al.*, 2014). It is recommended the LDL-C levels higher than 5 mmol/l (160 mg/dl) in adults be treated with medication and a change in lifestyle needs to be introduced (Jialal and Barton Duell, 2016).

Homozygous FH cases are rare with frequencies of 1 in 160 000 to 300 000 (Cuchel *et al.*, 2014). These patients present with very high levels of LDL-C, i.e. >10 mmol/l

(400mg/dl) and are especially difficult to treat (Jialal and Barton Duell, 2016).

Homozygous patients require immediate treatment and care as lack of treatment can lead to early CVD and death before the age of 20, making early diagnosis and intervention very necessary (Versmissen *et al.*, 2008).

Treating hypercholesterolaemia is done by prescribing medication and encouraging the incorporation of lifestyle changes in patients. Treatment has mostly been done by administering statins. Statins are a group of drugs that inhibit the enzyme HMG-Co A, thereby preventing the production of cholesterol in the liver and allowing the plasma levels of LDL-C to decrease. Lifestyle changes include following a healthy diet and regular exercise (Klug *et al.*, 2015).

There has been a recent breakthrough in the treatment of high LDL-C, including homozygous FH, by inhibiting the enzyme PCSK9 (Stein, 2012; Raal *et al.*, 2015). When LDL-C is taken into the cell, the receptors are not targeted for degradation by PCSK9 and return to the cell surface, hence increasing the intake of LDL-C into cells and reducing plasma levels of LDL-C.

Ference *et al.*, 2012 found that starting treatment earlier, i.e. in childhood, decreases the chances of the patient experiencing CVD, highlighting the need of detecting and diagnosing hyperlipidaemia as early as possible. Hypolipidaemia cases require no treatment as low LDL-C levels are protective against CVD and other heart conditions.

1.4. Genes associated with dyslipidaemia

In monogenic forms of dyslipidaemia, one variant in any of the known FH genes can alter lipid levels, enough to cause disease. However, as LDL-C levels are a continuous variable in a population, there must be variants in other unknown genes that influence an alteration in LDL-C levels. In individuals where variants are not detected in the monogenic genes, many variants in many different genes can cumulatively contribute to altering lipid levels.

There are many genes described in the literature that are associated with dyslipidaemia. Over 90 loci have been associated with the quantitative trait of TC levels as identified by meta-analysis genome wide association studies (GWAS) carried out on different populations all over the world (Kathiresan *et al.*, 2009; Teslovich *et al.*, 2010).

Associated variants for TC as a quantitative trait were identified in and close to the following genes (table 1.1) (Talmud, Futema and Humphries, 2014; García-Giustiniani and Stein, 2016):

Table 1.1: List of genes found to be associated with total cholesterol levels

Genes associated with TC levels				
<i>ABCA1</i>	<i>CAPN3</i>	<i>GPD1</i>	<i>MAFB</i>	<i>PTRF</i>
<i>ABCA1</i>	<i>CAV1</i>	<i>GPIHBP1</i>	<i>MAP3K1</i>	<i>PYGM</i>
<i>ABCA8</i>	<i>CETP</i>	<i>HFE</i>	<i>MC4R</i>	<i>RAB3GAP1</i>
<i>ABCG1</i>	<i>CIDEA</i>	<i>HMGCR</i>	<i>MLXIPL</i>	<i>RAF1</i>
<i>ABCG5</i>	<i>CILP2</i>	<i>HNF1A</i>	<i>MOSC1</i>	<i>SAR1B</i>
<i>ABCG8</i>	<i>CITED2</i>	<i>HNF4A</i>	<i>MSL2L1</i>	<i>SBNO1</i>
<i>ABO</i>	<i>CMIP</i>	<i>HPR</i>	<i>MTTP</i>	<i>SCARB1</i>
<i>AGPAT2</i>	<i>COBLL1</i>	<i>IRF2BP2</i>	<i>MVK</i>	<i>SLC22A8</i>
<i>AMPD1</i>	<i>COQ2</i>	<i>IRS1</i>	<i>MYLIP</i>	<i>SLC25A40</i>
<i>AMPD3</i>	<i>CPT2</i>	<i>JMJD1C</i>	<i>NAT2</i>	<i>SLC39A8</i>
<i>ANGPTL3</i>	<i>CTF1</i>	<i>KLF14</i>	<i>NPC1L1</i>	<i>SLCO1B</i>
<i>ANGPTL4</i>	<i>CYP26A1</i>	<i>KLHL8</i>	<i>NYNRIN</i>	<i>SORT1</i>
<i>APOA1</i>	<i>CYP2D6</i>	<i>LACTB</i>	<i>OSBPL7</i>	<i>SPTY2D1</i>
<i>APOA1 –C3 –A4 –A5</i>	<i>CYP7A1</i>	<i>LCAT</i>	<i>PABPC4</i>	<i>ST3GAL4</i>
<i>APOA5</i>	<i>DNAH11</i>	<i>LDLR</i>	<i>PCSK9</i>	<i>STARD3</i>
<i>APOB</i>	<i>ERGIC3</i>	<i>LDLRAP1</i>	<i>PDE3A</i>	<i>TOP1</i>
<i>APOC2</i>	<i>EVI5</i>	<i>LILRA3</i>	<i>PGS1</i>	<i>TRIB1</i>
<i>APOC3</i>	<i>FADS1 – 2 – 3</i>	<i>LIPC</i>	<i>PINX1</i>	<i>TRPS1</i>
<i>APOE</i>	<i>FLJ36070</i>	<i>LIPG</i>	<i>PLA2G6</i>	<i>TTC39B</i>
<i>APOE–C1–C2</i>	<i>FRK</i>	<i>LMF1</i>	<i>PLEC1</i>	<i>TYW1B</i>
<i>ARL15</i>	<i>FRMD5</i>	<i>LMNA</i>	<i>PLIN1</i>	<i>UBASH3B</i>
<i>BRAP</i>	<i>GALNT2</i>	<i>LOC55908</i>	<i>PLTP</i>	<i>UBE2L3</i>
<i>BSCL2</i>	<i>GCKR</i>	<i>LPA</i>	<i>PPARA</i>	<i>ZMPSTE24</i>
<i>C6orf106</i>	<i>GPAM</i>	<i>LPL</i>	<i>PPARG</i>	<i>ZNF648</i>
		<i>LRP1</i>	<i>PPP1R3B</i>	<i>ZNF664</i>

TC = total cholesterol

This list is not by any means a complete comprehensive set of genes with allelic variants that affect lipid levels; however, it does serve as an indication of how many genes can influence LDL-C levels.

In patients with raised LDL-C levels (LDL-C > 5mmol/l), pathogenic variants are identified in 60 to 80% of cases (Talmud, Futema and Humphries, 2014). In patients where a single variant could not be identified for FH, it is expected that other loci cumulatively contribute to the polygenic phenotype (Talmud *et al.*, 2013). With the number of genes listed in table 1.1, it is suggested that dyslipidaemia is not only monogenic, and can be caused by variants in unknown genes (Benn *et al.*, 2016).

Although there are a small number of 'core' genes that have a big effect on the phenotype (LDL-C levels), there are other genes, with small effects, that contribute to the phenotype (Boyle, Li and Pritchard, 2017). With this in mind, a genetic risk score could be generated to assess the combined impact of "small-effect—genes" on the phenotype. This is known as a polygenic risk score (PRS). The variants are weighted according to the odds ratio generated in logistic regression (see chapters 2 and 3) and may be population specific, with most research being done in populations of European origin.

Monogenic FH is primarily caused by pathogenic variants in four genes: *LDLR*, *APOB*, *PCSK9* and *LDLRAP1*. *PCSK9* and *APOB* have also been associated with hypolipidaemia. Variants in *LDLR* account for most cases of FH that have been documented thus far (table 1.2). The four genes that cause monogenic forms of dyslipidaemia will be discussed in further detail below.

All four genes code for proteins that aid in the intake and metabolism of LDL-C and the LDL-C receptor. Due to the monogenic nature, and therefore high impact on dyslipidaemia and LDL-C metabolism of the above mentioned four genes, they have been chosen for further investigation to examine LDL-C levels as a quantitative trait in Africans.

Table 1.2: Estimated frequencies of variants in genes that cause FH in populations of European origin adapted from Benn *et al.*, 2016

Gene	Estimated frequency of variants that cause FH due to high LDL-C levels
<i>LDLR</i>	95%
<i>APOB</i>	2-11%
<i>PCSK9</i>	~1%
<i>LDLRAP1</i>	Rare
Other	Unknown

These genes play a vital role in the metabolism of LDL-C in liver cells (figure 1.2). LDL-C transports cholesterol which circulates in the arteries. There is a protein called apolipoprotein B on LDL-C that binds to the LDL-C receptor on cells (Innerarity *et al.*, 1987), which is internalised with the help of an adaptor protein called LDLRAP1 (Eden *et al.*, 2002). Once internalised, the LDL-C:LDLR protein complex separates and the receptor is targeted for degradation by the enzyme PCSK9 (Maxwell and Breslow, 2004).

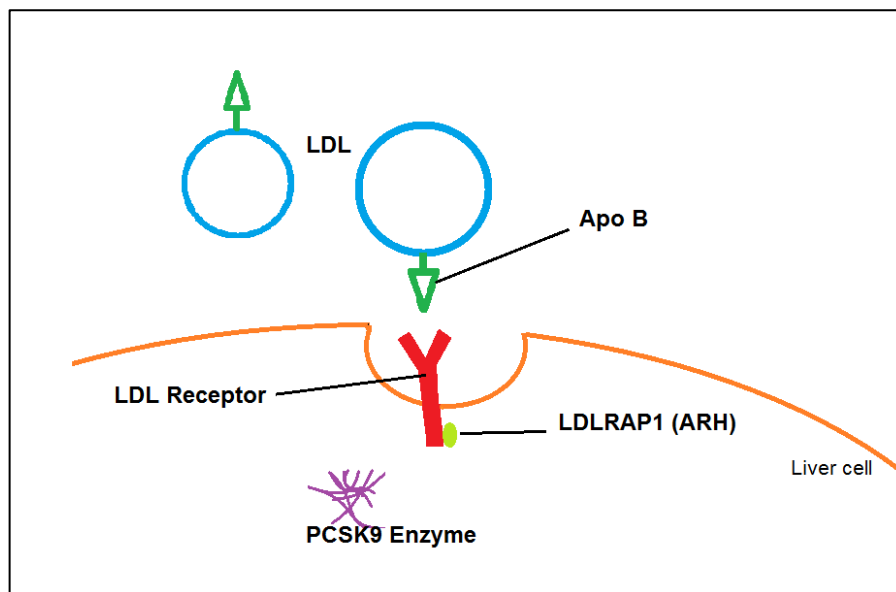


Figure 1.2: Components in LDL-C endocytosis. APOB, a component on LDL, binds to LDLR. An adaptor protein, LDLRAP1, mediates the intake of the APOB-LDLR complex into the cell. Once inside the cell, LDL-C dissociates from the receptor and LDLR is metabolised by an enzyme called PCSK9.

1.4.1. Low Density Lipoprotein Receptor (*LDLR*)

The *LDLR* gene codes for the low-density lipoprotein receptor protein. The gene is mapped to chromosome 19p13.2 and comprises 18 exons (NCBI, 2017). The receptor functions to maintain the plasma levels of LDL-C by binding to LDL-C and internalising it through receptor-mediated endocytosis. Once internalised, the receptor dissociates from its ligand (LDL-C) and LDLR returns to the cell surface to bind plasma LDL-C again.

Loss of function (LoF) variants in this gene render the receptor ineffective, thus leaving increased levels of LDL-C in the blood stream. Heterozygous variants at this locus result in haploinsufficiency of the receptor (Seidman and Seidman, 2002), therefore FH causing variants in *LDLR* are inherited in an autosomal dominant manner (Brown and Goldstein, 1974). LDL-C levels are high (5-15 mmol/l) and medical intervention and lifestyle changes are necessary.

1.4.2. Apolipoprotein B (*APOB*)

The *APOB* gene maps to chromosome 2p24.1 and has 29 exons (Deeb *et al.*, 1986; NCBI, 2017). It codes for the APOB protein which is a component on LDL-C and binds to LDLR and is internalised into liver cells for metabolism (Innerarity *et al.*, 1987; Hooper, van Bockxmeer and Burnett, 2005). Loss of function variants in this gene affect the binding affinity of APOB to LDLR, thus decreasing the internalisation of LDL-C (Innerarity *et al.*, 1987, 1990; Hooper, van Bockxmeer and Burnett, 2005). Variants in *APOB* are associated with both hyperlipidaemia and hypolipidaemia (Lee *et al.*, 2017). Variants that cause hyperlipidaemia are inherited in an autosomal dominant fashion (Innerarity *et al.*, 1987). For example, in a study conducted in a Caucasian European population, the variant R3500Q causes raised lipid levels (Benn *et al.*, 2016). Rare variants in *APOB* have also been identified in a Korean population (Lee *et al.*, 2017).

1.4.3. Proprotein Convertase Subtilisin Kexin Type 9 (*PCSK9*)

The *PCSK9* gene maps to chromosome 1p32.2 and has 12 exons (Varret *et al.*, 1999; NCBI, 2017). This gene codes for a protease that binds LDLR and targets the receptor for degradation (McNutt, Lagace and Horton, 2007). The mode of inheritance of variants in *PCSK9* is autosomal dominant (Abifadel *et al.*, 2003).

Gain of function (GoF) variants in *PCSK9* increase the degradation of LDLR in the cell, therefore restricting the number of receptors that return to the surface of the cell. This results in an increased level of LDL-C in the plasma (Abifadel *et al.*, 2003). Loss of function (LoF) variants in this gene cause low levels of LDL-C as the degradation of LDLR is reduced and, therefore, more receptors return to the surface of liver cells, allowing more plasma LDL-C to be internalised and metabolised (Cohen *et al.*, 2005). This finding has been exploited in treating hypercholesterolaemia very successfully (Robinson *et al.*, 2015).

1.4.4. Low Density Lipoprotein Adaptor Protein 1 (LDLRAP1)

The final gene examined in this study was *LDLRAP1*. The gene maps to chromosome 1p36.11 and has 9 exons (NCBI, 2017). The adaptor protein assists the receptor-mediated endocytosis mechanism of the LDLR:LDL-C complex (Soutar, Naoumova and Traub, 2003; Michaely *et al.*, 2004). Variants in this gene cause hypercholesterolaemia, but the mode of inheritance of hyperlipidaemia is autosomal recessive in this case as two deleterious mutations are necessary to result in high LDL-C levels (Soutar and Naoumova, 2004).

1.5. Variant frequencies in the common dyslipidaemia genes

There are many variants that have been identified in the four genes. For instance, a search on Ensembl gave 9923 associated variants listed for *LDLR*, 9758 for *APOB*, 4919 for *PCSK9* and 4427 for *LDLRAP1* (<https://www.ensembl.org>, accessed 21 Oct 2017). Most of these variants have no functional impact on the gene product and are neutral. The Leiden Open (source) Variation Database (LOVD) (Fokkema, den Dunnen and Taschner, 2005) is a database that contains FH causal variants ; however, the data is outdated as it was last updated in 2011. It can be accessed at <http://www.ucl.ac.uk/ldlr/Current/>

It has now been well established that the presence/absence and/or frequency of particular variants is variable in different populations. For example, the Y142X variant in *PCKS9* is observed in 2% of African Americans, but < 0.1% in European Americans (Cohen *et al.*, 2005,). The R46L variant in *PCSK9* is observed in 3% of Canadians (Saavedra *et al.*, 2014) and 1.7% in a European American population, but in only 0.15%

in the African American population (Hallman *et al.*, 2007). The former is accompanied by an 88% reduction in risk for CVD (Cohen *et al.*, 2006) and the latter is associated with a 86% reduction in CVD (Saavedra *et al.*, 2014). Accounting for population-specific genetic substructure is therefore important to consider in studies of multifactorial traits.

1.6. Dyslipidaemia is the South African Context

In some South African populations, the prevalence of FH is significantly higher than the worldwide prevalence due to founder effect. Variants, specifically in *LDLR*, have been well documented in the South African Afrikaner (Jenkins *et al.*, 1980), Indian (Rubinsztein *et al.*, 1992) and Jewish (Seftel *et al.*, 1989) populations. These variants are high impact and cause monogenic FH (table 1.3). While other studies may quote different prevalences (Ibe *et al.*, 2017), the prevalences are similar to those listed below, highlighting the fact that FH is a common disorder.

Table 1.3: Prevalence of FH in South African populations and variant profiles

Population	Estimated Frequency	Common deleterious variants in <i>LDLR</i>	Reference
South African Afrikaner	1/75 individuals	c.681 C>G	(Leitersdorf <i>et al.</i> , 1989)
		c.1285 G>A	(Leitersdorf <i>et al.</i> , 1989)
		c.523 G>A	(Kotze <i>et al.</i> , 1990)
South African Ashkenazi Jewish	1/67 individuals	654_656del	(Meiner <i>et al.</i> , 1991)
South African Indian	1/100 individuals	c.2054 C>T	(Soutar, Knight and Patel, 1989)

There is no equivalent data available on variants causing hyper- or hypolipidaemia for the black South African population. This may be merely due to the lack of genetic studies in this population, or it could be due to a general low level of lipids in the black African populations (Hooper *et al.*, 2007), with consequent low frequencies of diseases associated with hyperlipidaemia. Therefore, the purpose of my study was to identify variants that have a phenotypic effect on lipid levels as a quantitative trait in black African populations. A case-control study design based on clinical cut-offs (highest LDL-C and lowest LDL-C in the study participants) was used and the study approach was a

candidate SNP screen of markers in selected genes in a case-control (high LDL-C vs low LDL-C) study cohort. Four genes were chosen as good candidates as they are known to be associated with monogenic dyslipidaemia. Variants in these genes were identified and their frequency assessed in African populations.

Because hyperlipidaemia is underdiagnosed and undertreated globally, identifying variants which alter lipid levels could prove pivotal in increasing diagnoses and implementing treatment earlier to prevent premature CVD and death.

1.7. Study Rationale

To identify functional variants in candidate genes with the potential to affect LDL-C levels, this study used publicly available whole genome sequence data from African populations. Variant frequencies in different populations were considered. To establish whether variants in the candidate genes were predicted to contribute to high and low LDL-C levels, the variants were examined using bioinformatics tools to identify potentially functional variants.

Suitable variants were chosen, based on potential function and allele frequency, to be investigated in participants from the AWI-Gen study. AWI-Gen is an Human Heredity and Health in Africa Consortium (H3Africa) study referred to as AWI-Gen (Africa Wits INDEPTH partnership for genomic studies) (Ramsay *et al.*, 2016). The AWI-Gen study includes >10 000 participants from different sites across Africa (South Africa, Ghana, Burkina Faso and Kenya). One thousand participants between the ages of 35 and 85 years were chosen for the current study, primarily based on their LDL-C levels (500 participants with high LDL-C levels and 500 participants with low LDL-C levels).

The participants, although originating from different sites across Africa, were assigned into two groups: 500 with the lowest LDL-C levels (controls) and 500 with highest LDL-C levels (cases). We recognise that different African populations harbour genetic differences (Durbin *et al.*, 2010; Pickrell *et al.*, 2012) that may influence our ability to detect population-specific LDL-C associated variants. Population substructure is the distinct genetic difference between populations due to ancestral origin of the population and can lead to false positive associations. If there are differences in the relative proportions of different populations in cases, compared to controls, the associations

could reflect populations differences rather than differences due to disease (Tian, Gregersen and Seldin, 2008). To avoid this, various factors need to be accounted for, including ethnicity. Sometimes country of origin is used as a proxy for ethnicity, as objective measures are relatively poor when only a small number of genetic markers are included in a study. In this study, the participants were grouped together to increase the power to detect variants associated with LDL-C levels, but geographic region was included as a potential confounder. The genotype data was corrected for multiple testing, and geographic region was used as a covariate in logistic regression analysis. As data for African populations is very limited concerning this subject, this study should be considered a small pilot study that aims to serve as a starting point for future studies focusing on lipid levels.

This study aimed to identify variants in candidate genes that may play a role in LDL-C levels in Africans. It would have been ideal to examine the full DNA sequence of all four genes, however, due to cost constraints, a genotyping approach was used and only a small set of selected variants was chosen for investigation.

1.8. Aim and objectives

1.8.1. Aims

To use a computational approach to identify potentially functional genetic variants in African whole genome sequence data in established candidate genes known to alter LDL-C levels (*LDLR*, *APOB*, *PCSK9* and *LDLRAP1*), and to test their association with LDL-C levels in participants from the AWI-Gen study.

1.8.2. Objectives

1. To identify variants in the coding and flanking regions of *LDLR*, *APOB*, *PCSK9* and *LDLRAP1* from African whole genome sequences (WGS). The WGS data were obtained from the 1000 Genomes Project (KGP) and the African Genome Variation Project (AGVP).
2. To assess the potential functional impact of the variants by using relevant bioinformatics tools and to choose a small number for further investigation.

3. To test variants for association with high or low LDL-C levels in 1000 participants from the AWI-Gen study and to analyse the data in a “case-control” study design (high LDL-C considered as cases and low LDL-C considered as controls).

1.9. Study approach

Since it is known that the four genes *LDLR*, *APOB*, *PCSK9* and *LDLRAP1* are monogenetically associated with LDL-C levels in populations worldwide, they were chosen as candidates and used as a starting point for investigating the possible aetiology and landscape of LDL-C levels in black African populations. Variants were identified in these genes from African WGS data through a process of assessing the predicted functional impact, minor allele frequency and type of variant (missense, regulatory, synonymous and nonsense) for each variant. Selected variants were genotyped in a cohort of 1000 AWI-Gen participants: 500 with low LDL-C levels and 500 with high LDL-C levels. Regression analysis was done with the expectation of finding an association between LDL-C levels and alleles. Figure 1.3 summarises the study approach.

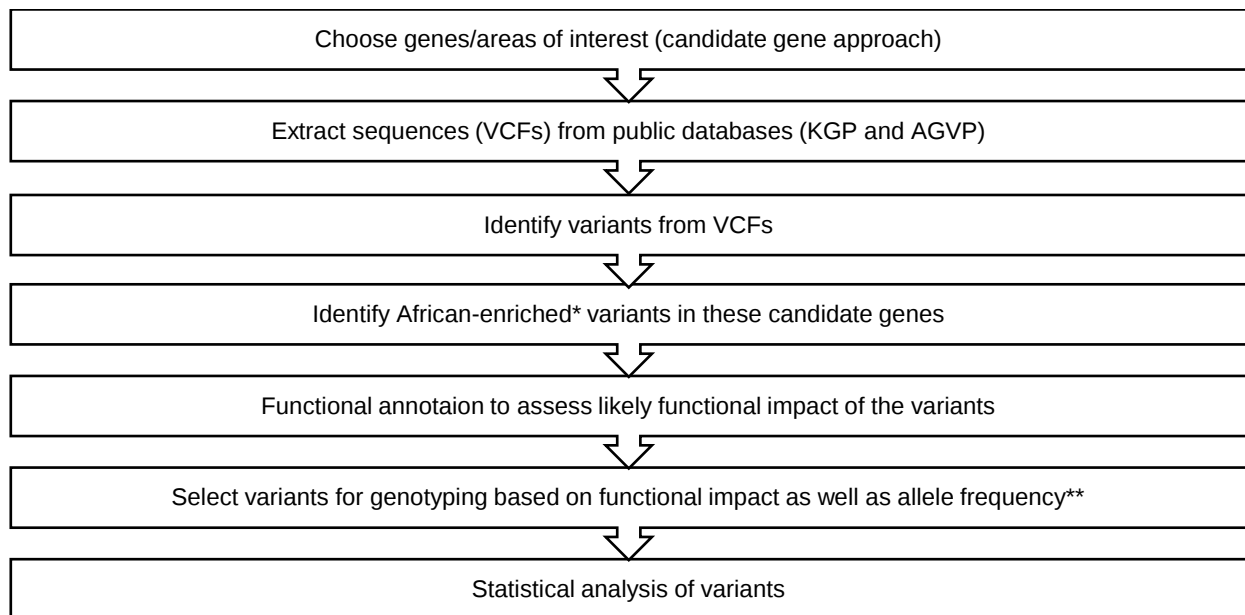


Figure 1.3: Flow diagram of study approach.

*Variants may be present in other populations, too, but are more common in African populations

**Variants were chosen if predicted to have an effect on the protein, or if the minor allele was present in at least 20% of the population. KGP frequencies were used.

2. Methods and Materials

The aim of this project was to identify variants in WGS from African populations, and to test their association with LDL-C levels in a cohort from the AWI-Gen study. This project can be divided up into 2 sections, namely, the initial bioinformatics section which aimed at identifying variants in appropriate genes, and the subsequent laboratory section which aimed to look at the frequency of these variants in a “case-control” (high vs low LDL-C) approach. The bioinformatics section makes use of online databases and tools that are freely available. Figure 2.5, at the end of this chapter, shows a flow diagram of the methods used for this study.

Study participants were chosen from the AWI-Gen (**A**frica, **W**its-INDEPTH Partnership for **GEN**omic studies) project, a Human Heredity and Health in Africa (H3Africa) Consortium study (Ramsay *et al.*, 2016). Together with the International Network for the Demographic Evaluation of Populations and their Health in low- and middle-income countries (INDEPTH), the study aims to understand the interaction between genetic, epigenetic and environmental risk factors for obesity and cardiometabolic diseases in sub-Saharan Africa. AWI-Gen has over 10 000 participants that have been recruited from six sites across East, West and South Africa in four countries: Kenya, South Africa, Ghana and Burkina Faso (figure 2.1). Extensive phenotypic data was collected for each participant.

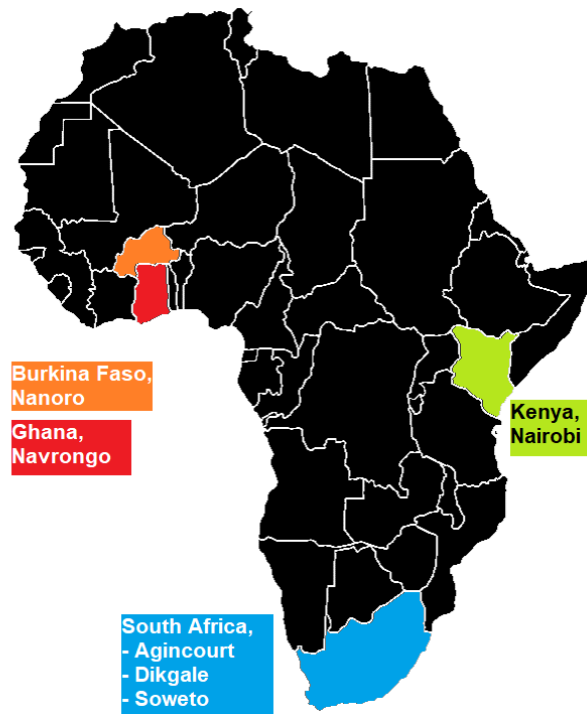


Figure 2.1: Map displaying sites that AWI-Gen participants were selected from. Participants were selected from six sites from four countries: Burkina Faso (Nanoro), Ghana (Navrongo), Kenya (Nairobi) and South Africa (Agincourt, Dikgale and Soweto).

WGS data were pulled from public databases. Public bioinformatics databases provide easily accessible genetic data, freely, to the public. The data can be stored in many different file formats, and some databases occasionally require a letter of request that informs the centre/group to which the data belongs what the data will be used for. The two public databases with WGS data that were used are the 1000 Genomes Project (KGP) and the African Genome Variation Project (AGVP) databases. These databases were chosen because they both contain genome sequences of African populations. The files that were extracted were VCF files (variant call format) – which stores all the variation identified in the individuals' WGS.

The variants were examined for deleteriousness by means of online tools that are freely available to the public, e.g. Variant Effect Predictor (VEP) (McLaren *et al.*, 2010) and Combined Annotation Dependent Depletion (CADD) (Kircher *et al.*, 2014). These tools are provided through a genome annotation portal such as Ensembl (ensembl.org) and by universities such as the University of Washington (cadd.gs.washington.edu), respectively. The tools assess and then predict the effect variants have on the function

of a protein, enabling one to determine whether the variants are predicted to have a functional effect. Suitable variants, based on predicted function or minor allele frequency (MAF>0.20) in the combined African data, were chosen to be genotyped in a cohort of AWI-Gen participants.

Thereafter, the chosen variants were genotyped in individuals from the AWI-Gen study, selected for high or low LDL-C levels. Association analysis was carried out to determine whether the variants were associated with the two extremes of lipid levels in African populations.

2.1. Participants

Participants from the AWI-Gen study were used in this study. The AWI-Gen study collected fasting LDL-C levels for all participants and 1000 participants were selected for this study. The study also collected additional phenotype data from participants: fasting glucose levels, medication use, body mass index (BMI) and alcohol use, among many others.

The AWI-Gen study consists of over 10 000 participants with various phenotype data. Only 1000 individuals were chosen in this study due to budget constraints. For this study, participants were excluded based on the following criteria:

- Diabetes
- BMI > 35
- Problematic alcohol use
- Individuals on medication for hyperlipidaemia
- DNA with a concentration less than 4 ng/μl

The age of the participants ranged between 35 and 80 years old. Looking at the bell curve of normal LDL-C levels (figure 2.2), 1000 participants on the two extremes were chosen: 500 with the highest LDL-C levels (> 3.5 mmol/l) and 500 participants with the lowest LDL-C levels (< 1.1 mmol/l). This approach is based on extreme phenotyping (Li *et al.*, 2011; Gurwitz and McLeod, 2013).

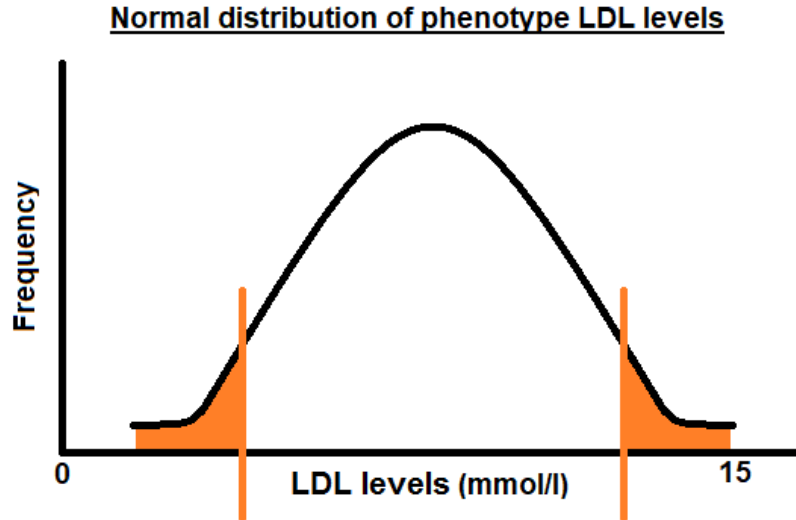


Figure 2.2: Hypothetical data following normal distribution. Coloured areas show the two groups of participants chosen, based on LDL-C levels, the group with low LDL-C levels were used as a 'control' group against which the group with high LDL-C ('cases') were compared.

The approach of extreme phenotyping is used to separate samples into “cases” and “controls” when studying a quantitative trait. This study includes 500 cases (high LDL-C levels) and 500 controls (low LDL-C levels). The power of this study is 97.91%, to detect a variant with a MAF of 0.1 and odds ratio (OR) of 0.5 using a sample size of 1000 samples. The power would be 83.43% to detect a variant with the same parameters but with an OR of 1.5. Quanto was used to determine the power of the study (Gauderman, 2002).

A post-hoc power analysis was done on two SNPs that were significantly associated with LDL-C levels. The two SNPs were rs12071264 and rs6752026. A study of 500 cases and 500 controls would be 85.94% powered to detect association with the first SNP (MAF = 0.090, OR = 0.5866). The power would be 75.09% to detect association with the second SNP (MAF = 0.135, OR = 0.6898) in a study of the same size.

A chi square test was done to determine whether sex was equally represented between the high and low LDL-C level groups. Two-sample Wilcoxon rank-sum tests were carried out to determine whether LDL-C, BMI and age were significantly different between the two LDL-C groups.

Ethics was obtained from the Wits Human Research Ethics Committee (Medical): M160833 (Appendix A).

2.2. Candidate gene selection

Four genes were selected to be investigated: *LDLR*, *APOB*, *PCSK9* and *LDLRAP1*. As described in the introduction, these genes are known to cause monogenic FH when they contain pathogenic mutations. In this study, however, they were studied to identify variants that would be associated with LDL-C levels as a polygenic trait.

2.3. Whole genome sequences available for study

WGS data were used in this study. They were extracted from two public databases namely, KGP and AGVP, in VCF file format with 1000 bp flanking region on either side. AGVP and KGP (Durbin *et al.*, 2010) called their variants by aligning the sequences to the human reference genome (GRCh37) (Gurdasani *et al.*, 2015). For this study, a total of 975 individuals from eight different populations were investigated from low coverage (2 to 4 X) WGS data. Coverage is the number of times a nucleotide has been read or sequenced (Sims *et al.*, 2014). Low coverage data can reduce the chances of accurately calling a variant, especially when a variant is rare, as the data is only supported by 2 to 4 reads. There were 655 individuals from KGP and 320 individuals from AGVP. The breakdown of population distribution is presented in Table 2.1.

Table 2.1: Numbers of samples with available low coverage WGS data from each black African population from KGP and from AGVP

Population	Number of samples
KGP	
Esan, Nigeria (ESN)	111
Gambian, West Gambia (GWD)	220
Luhya, Kenya (LWK)	115
Mende, Sierra Leone (MSL)	98
Yoruba, Nigeria (YRI)	111
AGVP	
Baganda, Uganda	100
Ethiopian, Ethiopia	120
Zulu, South Africa	100

KGP= The Thousand Genome Project

AGVP = African Genome Variation Project

The countries from which the populations originate are highlighted in figure 2.3.

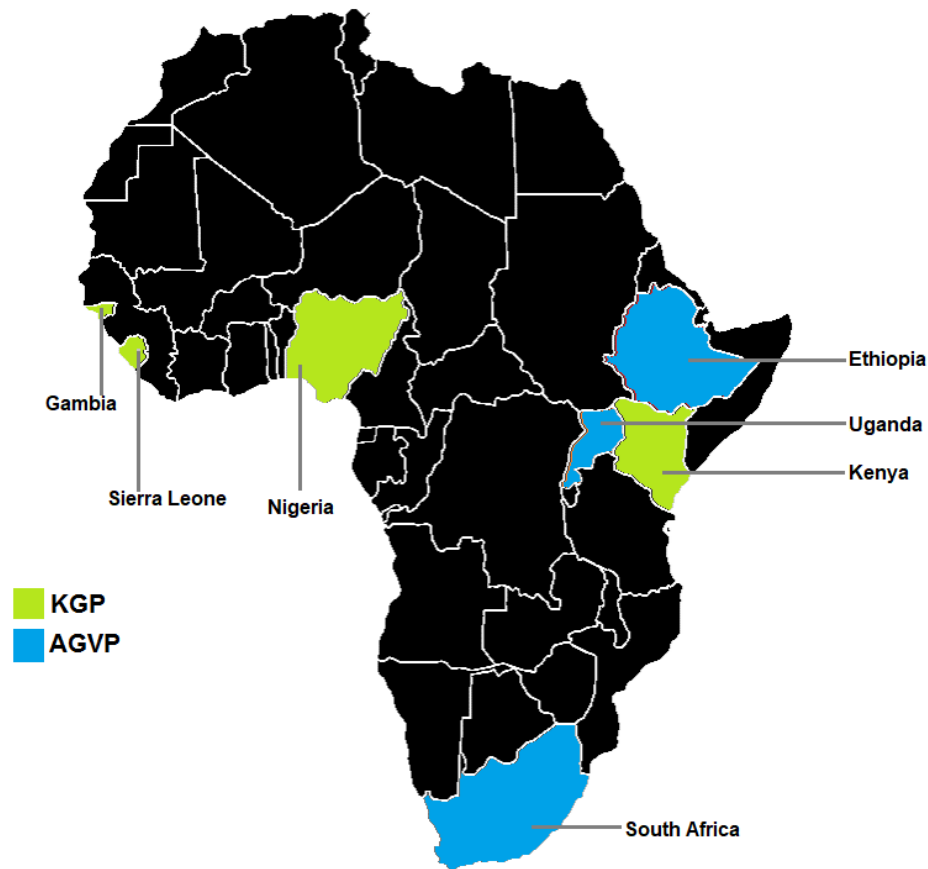


Figure 2.3: Map of Africa showing countries from which WGS data was publicly available for analysis: Ethiopia, Gambia, Kenya, Nigeria, Sierra Leone and South Africa

The KGP aimed to create a catalogue of human genome data and find variants with frequencies of at least 1% in the population. In this project a total of 2504 whole genomes from different populations were sequenced. In total, 26 populations were sampled from Africa, East Asia, South Asia, Europe, and North and South America. The sequences from KGP were constructed using low coverage genome sequencing (2-4 X), deep exome sequencing as well as microarray genotyping. It is interesting to note that 28% of all novel variants found were from African populations alone (The 1000 Genomes Project Consortium *et al.*, 2015).

The AGVP (Gurdasani *et al.*, 2015) identified that while African populations are the most genetically diverse, the characterisation of variation in African genomes is limited. The

paper describes 1481 individuals from 18 different populations from sub-Saharan Africa, but whole genome sequences (4X coverage) are only available for 320 individuals from 3 populations: Baganda, Ethiopian and Zulu. There is also genotype data available for all 1481 individuals from the HumanOmni2.5M genotyping array.

2.4. Variant Selection and Functional Annotation

From these African whole genome sequences, data from the four genes of interest (*LDLR*, *APOB*, *PCSK9* and *LDLRAP1*) were extracted, with a 1000 bp flanking region on either side. A total of 3541 variants were identified. The SNPs from these regions were functionally annotated using CADD and Ensembl's VEP to identify deleterious variants. VEP includes the tools SIFT and PolyPhen2 (see descriptions below). SNPs were selected in two stages: firstly, based on predicted deleteriousness, MAF (4% to 20%) and type of variant, and secondly based on a minimum MAF of 20%, and maximum of 45%, in African populations. Sequences from KGP and AGVP were mapped to GRCh37, therefore VEP was used on Ensembl's archive site for GRCh37.

CADD (Combined Annotation Dependent Depletion) measures deleteriousness (variants that reduce the fitness of an organism), and not just molecular and functional pathogenicity of a variant. CADD combines annotation from sources such as VEP and data from the ENCODE Project which includes information on conservation metrics, regulatory information and transcription factor binding regions to name a few. These variants are then contrasted to simulated variants and deleteriousness scores are generated (Kircher *et al.*, 2014). Variants that had a score greater than 10 were selected as potentially deleterious.

SIFT (Sorting Intolerant From Tolerant) (Kumar, Henikoff and Ng, 2009) is primarily used in predicting whether nonsynonymous variants are deleterious or not by assessing whether an amino acid substitution is likely to affect the function of a protein. The algorithm compiles a dataset of aligned protein sequences for a sequence of interest. Each position is scanned, and a probability of deleteriousness is calculated, and a matrix generated for each of the 20 amino acids occurring at that position. A substitution is then predicted, and a score is assigned to that position. Positions that are highly conserved are generally more intolerant than less conserved positions. The scores that are given to each variant are interpreted as follows: <0.05 = damaging, >0.05 = tolerated

(Kumar, Henikoff and Ng, 2009). In this study, variants that were scored <0.05 were selected as potentially deleterious.

PolyPhen2 (Polymorphism Phenotyping v.2) (Adzhubei *et al.*, 2010) also focuses on predicting whether nonsynonymous variants are likely to be harmful or not. The tool does this by calculating the impact of an amino acid change on the structure and function of a human protein. It calculates a Naïve Bayes probability score, reporting variants as benign (0), possibly damaging (0.5) or probably damaging (1). In this study, variants that were scored as possibly damaging and higher (> 0.5) were selected as potentially deleterious.

After generating scores for all the variants in all four genes, only those variants that occurred in six or more of the eight populations being studied were chosen. After this, variant selection happened in two phases: the first phase filtered variants based on deleteriousness, the type of variant (missense, synonymous, nonsense, regulatory) and MAF (between 0.04 and 0.2). Variants had to have at least one moderately deleterious score (PolyPhen2 > 0.5 , SIFT <0.05 , CADD >10). The second phase filtered variants based on only MAF. As the sample size of the total cohort was 1000, the variants that were chosen for a MAF between 0.2 and 0.45 to increase the power of finding an association. The MAF for African populations was used from dbSNP, which is generated from KGP. Linkage disequilibrium (LD) was assessed for the selected variants. LD analysis was carried out using Haploview ($r^2>0.4$) (Barrett *et al.*, 2005).

2.5. Genotyping approach

There are many methods to carry out genotyping including, TaqMan Real-Time PCR, ARMS-PCR and array genotyping. The most economical option available for this project proved to be the use of the MassARRAY System by Agena Bioscience. The MassARRAY System is provided as a commercial service by Inqaba Biotech in Pretoria, South Africa.

The DNA used for the genotyping was obtained from the Biobank based at the Sydney Brenner Institute for Molecular Bioscience (SBIMB) after receiving appropriate approval from the steering group. The DNA had been extracted using either the salting out method (Miller, Dykes and Polesky, 1988) or the automated Qiasymphony (Qiagen,

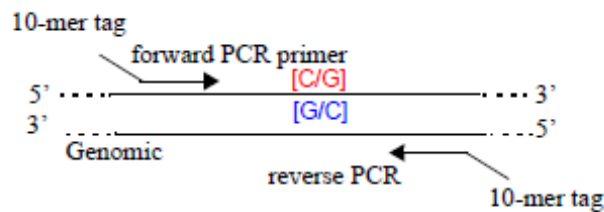
Hilgen, Germany) method. The Biobank stores working aliquots at 4°C, and a storage aliquot at -80°C. DNA for this project was taken from the working aliquots. The concentration for each of the 1000 samples was normalised to ~30 ng/μl and ~10 μl DNA was sent for genotyping.

The MassARRAY platform measures the mass of molecules and assigns a nucleotide base to a position based on the difference in mass. When designing an assay for the MassARRAY system, specific software is used to test whether SNPs of interest can be genotyped (failed SNPs are indicated in tables 1 and 2 in appendix B and were excluded from further downstream steps).

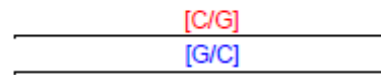
After confirming SNPs that could be genotyped using the MassARRAY through an *in silico* test, the area of interest was amplified and PCR products of 80-120 bp in length were generated. The PCR product was cleaned with SAP (shrimp alkaline phosphatase) to neutralise dNTPs that haven't been incorporated. A single base extension (SBE) reaction followed, where only one base (which is the SNP of interest) was added at the end of the primer. The SBE primers were specifically designed so that they were 15 to 30 bp long, or 4500 to 9000 Da in mass.

The mass of each SBE primer, as well as each individual nucleotide was known prior to the SBE reaction. In the MassARRAY system, the SBE product was excited, the DNA was ionised and it moved through a vacuum to a detector. The time that it took for it to get to the detector is determined by the mass and the system calculated the difference of the mass before and after extension, hence determining the genotype of the SNP of interest. Figure 2.1 shows the stepwise process of the MassARRAY system.

Amplification



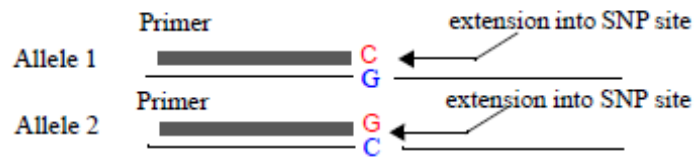
PCR Product



SAP Treat-

SAP treatment to neutralize unincorporated dNTPs

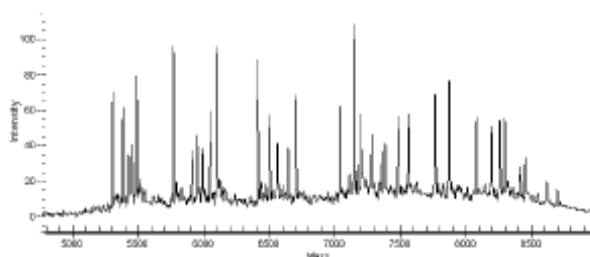
iPLEX Gold Reaction



iPLEX Gold cocktail containing primer, enzyme, buffer, and terminator nucleotides

Sample conditioning, dispensing, and MALDI-TOF MS

Spectrum



MALDI-TOF mass spectrometry analysis

24-plex spectrum

Figure 2.4: Process of MassARRAY genotyping. The area of interest was amplified, and PCR products were cleaned to remove unincorporated nucleotides. Single base extension followed which included a specific primer that extends with only one base. This base is at the position of the SNP of interest. The PCR products were put into a mass spectrometry machine where the masses of the products were measured. Genotypes were then derived from the differences in mass. (Image taken from MassARRAY manual supplied by Ingaba biotec).

2.6. Data Analysis

PLINK v.1.9 is an open source tool that is used primarily for genome wide association studies (GWAS) analysis (Purcell *et al.*, 2007; Chang *et al.*, 2015). PLINK v.1.9 (www.cog-genomics.org/plink/1.9) was used to analyse genotype data from the MassARRAY system. Files that are compatible with PLINK v.1.9 (.map and .ped) were generated from the raw data received from the MassARRAY system. These were then converted into .bim, .bam and .fam files in PLINK v.1.9 for further association analysis. The data was separated into high and low LDL-C participant categories so that a case-control type of analysis could be carried out. High LDL-C level participants were tagged as cases, and low LDL-C participants were tagged as controls.

Quality control measures for samples and SNPs after genotyping were followed as per Anderson *et al.* (2011). There were some genotyping failures, so not all samples have data for each SNP. Therefore, samples that had more than 30% missing SNP data were excluded from further analysis. Similarly, some SNPs fail genotyping in many samples, therefore, SNPs that failed in more than 40% of samples were excluded. In addition, SNPs incompatible with Hardy-Weinberg equilibrium (HWE) at $p < 0.005$, differential missingness < 0.00001 and MAF < 0.01 were excluded from further analysis. The dataset was not assessed for extreme heterozygosity and duplications because the number of SNPs genotyped were too few to accurately calculate estimates (Anderson *et al.*, 2010).

SNP frequencies were calculated from the genotyping data for each region (West, East and South Africa). A Fisher's Exact Test was carried out to see if the SNP counts for 14 SNPs differed across the three regions.

2.6.1. Multiple testing

When carrying out association tests, more than one comparison is being carried out. As the number of inferences (tests) increases, the probability of an erroneous inference occurring increases, i.e. the probability of detecting a false positive association. The chance of false positives occurring and rejecting the true null hypothesis, therefore needs to be corrected for. In this study, 14 SNPs were tested. Therefore, 14 hypothesis tests were carried out resulting in an increased probability of a false association

occurring. A statistical p value of, for example, 5% will tell you that there is a 5% chance of false positives occurring. When one does many tests, it is important to correct, or adjust p values, for multiple testing.

The Benjamini-Hochberg method, or the false discovery rate (FDR) adjusts p values so that the number of false positives that will be reported are reduced (Benjamini and Hochberg, 1995). The method considers the proportion of the rejected null hypotheses which are rejected in error. This ensures that the significantly associated SNPs are less likely to be false positives.

P values were only corrected for in allelic association. Correction for multiple testing was not done in logistic regression after adjusting for covariates, as the study is a hypothesis based study rather than an exploratory based study. We, therefore, think that correction for multiple testing does not need to be done after adjusting for covariates.

2.7. Case-Control Association Analysis

2.7.1. Logistic regression

Allelic association was carried out using PLINK v.1.9 to determine which alleles of the 14 SNPs were significantly associated with LDL-C levels. The p values were corrected for multiple testing using the Benjamini-Hochberg method. All variants with a significance level of $p < 0.05$, corrected for multiple testing, were considered significant. The odds ratios (OR) and 95% confidence intervals (CI) were calculated using the major allele (A2) in this study as a reference. Therefore, the OR explains the effect the minor allele (A1) has on the phenotype. The OR is a numerical value that represents the association between a SNP and LDL-C levels (Szumilas, 2010).

Logistic regression is predictive regression analysis carried out on a dichotomous dependent variable (high or low LDL-C in this case). It is a predictive analysis that describes the relationship between a dependent variable and an independent variable (SNP variant) (Walker and Duncan, 1967). Logistic regression gives an OR that is the estimated measure of association between an exposure (variant) and outcome (LDL-C levels). The OR gives the odds of an outcome occurring if an exposure is present, compared to odds of the outcome occurring without the exposure variable. It gives a log odds increase of the outcome. An OR > 1 signifies that the minor allele (A1) is

associated with increased levels of LDL-C (>3.5 mmol/l). In contrast, an OR < 1 means that the minor allele is associated with decreased levels of LDL-C (<1.1 mmol/l).

Logistic regression was carried out for the 14 SNPs that passed quality control. The analysis was adjusted for potential covariates, namely: sex, BMI, glucose levels and region of origin of participants (East (Nairobi), West (Ghana and Burkina Faso) and South Africa (Dikgale, Soweto and Agincourt)).

2.7.2. Polygenic risk score

A polygenic risk score was calculated using six significantly associated SNPs with $p < 0.05$ in the allelic association, before adjusting for potential confounders. For each SNP, if the OR was >1 (i.e. the minor allele was associated with a high LDL-C level), the minor allele homozygous genotype was given a score of 0, the heterozygote genotype was given a score of 1, and the major allele homozygous genotype was given a score of 2. On the other hand, if the OR was <1 (i.e. the minor allele was associated with a low LDL-C level), the minor allele homozygote genotype was given a score of 2, the heterozygote genotype a score of 1 and the major allele homozygous genotype was given a score of 0.

R was used to create a frequency plot with the polygenic risk score for cases and controls. A Kruskal-Wallis rank test was done using STATA v.14.2 (StataCorp, 2015) to determine whether the polygenic risk score for cases and controls are significantly different to each other.

In addition, to show the linear correlation of the polygenic risk score against the mean of LDL-C level per risk score, a plot was generated using R v.3.4 (R Core Team, 3.4).

2.7.3. Visualisation of results

Box and whisker plots were generated for two SNPs that were significantly associated with LDL-C levels after logistic regression and correction using the false discovery rate adjustment. The effect of the addition of each allele is shown, plotted against LDL-C levels in the 1000 participants. The plots were generated in R.

A Shapiro-Wilk test was carried out to assess the normality of LDL-C levels across the genotypes. This was followed by a Kruskal-Wallis test to determine whether the LDL-C levels were statistically different ($p < 0.05$) across the three genotypes, followed by a post-hoc contrast analysis. This was done for the two SNPs that were statistically significant after adjusting for covariates in logistic regression analysis (rs6752026 and rs12071264). This was done using Real-Statistics software (Charles Zaiontz, 2018).

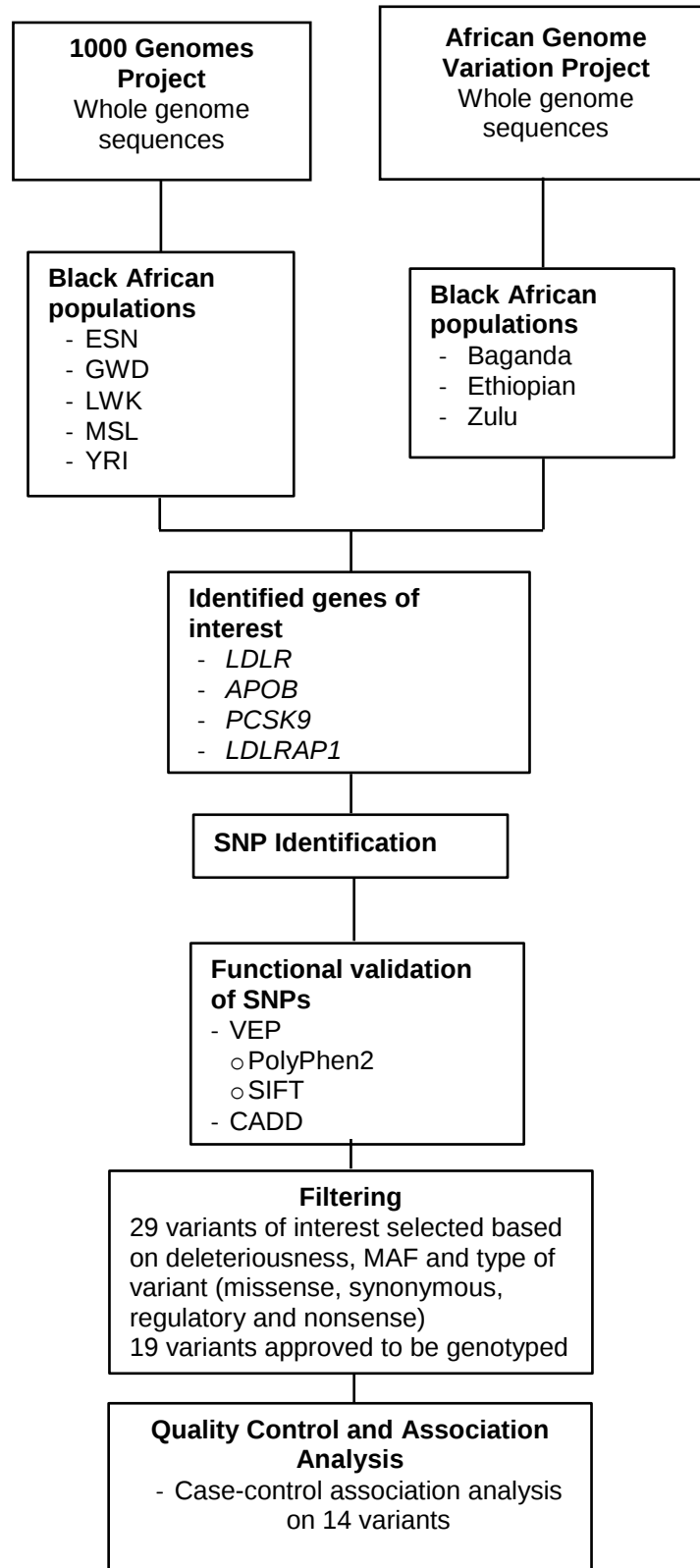


Figure 2.5: Flow diagram summarising the filtering and association analysis carried out for this study.

3. Results

3.1. Participants

The 1000 participants from the AWI-Gen study were chosen from among six sites: three sites from South Africa (Agincourt, Dikgale and Soweto), and one each in Burkina Faso (Nanoro), Kenya (Nairobi) and Ghana (Navrongo). They were selected from among those at the extremes of the LDL-C level distribution and numbers of participants from each site are shown in table 3.1. Despite the participants originating from different sites, they were pooled into one sample set for regression analysis. Due to the diversity of African populations, population substructure was corrected for in the logistic regression analysis using geographic location as a proxy. The distribution of participants from each region (South, West and East Africa) is distributed evenly. However, the number of cases and controls from each region is greatly skewed; East Africa, there are 30 controls and 276 cases, West Africa: 302 controls, 23 cases, and South Africa: 168 controls, 201 cases. This was adjusted for in the logistic regression analysis.

Table 3.1: Distribution of 1000 participants across AWI-Gen sites stratified by sex

	High LDL-C (n=500)		Low LDL-C (n=500)				
AWI-Gen Site	M	F	M	F	Total (site)	Region	Total (Region)
Agincourt (SA)	52	75	51	49	227	South Africa	369
Dikgale (SA)	11	18	16	35	80		
Soweto (SA)	45	0	17	0	62		
Navrongo (Ghana)	0	6	84	148	238	West Africa	325
Nanoro (Burkina Faso)	12	5	12	58	87		
Nairobi (Kenya)	133	143	16	14	306	East Africa	306
Total (sex)	253	247	196	304	1000		

M = Male, F = Female, SA = South Africa

The phenotype data (outlined in tables 3.2, 3.3 and 3.4) used for this study on the AWI-Gen participants were: age, BMI, fasting glucose levels, sex, HIV status and antiretroviral treatment (ARV), and LDL-C levels. LDL-C was significantly different since they were divided into two groups based on the LDL-C levels. Sex, BMI and glucose levels show a significant difference ($p < 0.05$) between the two groups of participants. The mean age of the high LDL-C level group and low LDL-C level group was 51.07 and 50.32 years, respectively. Age was the only phenotype to show no significant difference ($p = 0.1938$) between the two groups.

Table 3.2: Phenotype characterisation of 1000 AWI-Gen individuals: sex, age, BMI, glucose levels and LDL-C cholesterol

Phenotype Characterisation					
Phenotype	High LDL-C (n=500)		Low LDL-C (n=500)		p value
Sex (%F)	49.40%		60.80%		$<1.00 \times 10^{-3}$
	Mean	Standard deviation	Mean	Standard deviation	
Age	51.07	± 7.10	50.32	± 6.31	0.19
BMI	26.27	± 4.32	21.52	± 3.62	$<1.00 \times 10^{-3}$
Glucose*	5.10	± 0.69	4.62	± 0.68	$<1.00 \times 10^{-3}$
LDL-C	4.37	± 0.70	0.85	± 0.19	$<1.00 \times 10^{-3}$

* Fasting glucose levels, %F = % females

There were more women in the low LDL-C group, and the high LDL-C group was characterised by a higher BMI, and higher fasting glucose levels. The phenotypes were also evaluated by sex in table 3.3. The only phenotype that was significantly different between the sexes was LDL-C levels ($p = 0.005$). The remainder of the phenotypes: age, BMI and glucose levels showed no significant difference between males and females.

Table 3.3: Phenotype characterisation of 1000 AWI-Gen participants separated by sex: LDL-C levels, age, BMI and glucose levels

Phenotype	Female (n=551)		Male (n=449)		P value
	Mean	Standard deviation	Mean	Standard deviation	
LDL-C	2.41	±1.78	2.85	±1.87	5.00x10 ⁻⁴
Age	50.71	±6.62	50.68	±6.86	0.74
BMI	23.96	±4.68	23.82	±4.60	0.74
Glucose*	4.85	±0.69	4.87	±0.76	0.23

* Fasting glucose levels

Data on HIV infection (table 3.4) was missing for 57 individuals (11.4%) of cases and 297 (59.4%) of controls, with little information on treatment. In the high LDL-C group, most participants were recorded as HIV negative (383/500), with the rest of the participants either being HIV positive, or no data was recorded. In the low LDL-C group, there was no data recorded for most participants (297/500).

Although HIV infection was not used as an exclusion criterion when participants were selected, it has been identified as a potential confounder due to the effects HIV infection and ARV treatment have on LDL-C levels. However, due to the amount of missing data on HIV for the AWI-Gen individuals, HIV status and treatment could not be included in logistic regression analysis as a covariate.

In the high LDL-C group, 13.54% are HIV positive, and from the low LDL-C group, 37.93% are HIV positive. From these individuals that are HIV positive, 63.33% and 42.86% have received ARV treatment from the high and low LDL-C groups, respectively. It seems that more individuals in the low LDL-C group are HIV positive, however, more individuals from the high LDL-C group receive treatment.

Table 3.4: HIV status and treatment of 1000 AWI-Gen participants

Treatment status	High LDL-C (n=500)				Low LDL-C (n=500)			
HIV status	Positive: 60	Negative: 383	Unknown: 57	Total 500	Positive: 77	Negative: 126	Unknown: 297	Total 500
On Treatment	38	1*	0	39	33	0	0	33
Not on treatment	1	1	0	2	12	1	0	13
No data	21	381	57	459	32	125	297	454

*This participant has a negative status but is documented as having received treatment. This could be a potential data capture error in AWI-Gen.

The distribution of the LDL-C levels across 5940 AWI-Gen individuals (figure 3.1 A) of the parent study indicates that low LDL-C levels are common, with 66% of the individuals having LDL-C levels below 4 mmol/l. LDL-C levels in the control group with low LDL-C for the present study ranged from 0.4 mmol/l - 1.2 mmol/l. The case group with high LDL-C group had LDL-C levels ranging from 3.7-14.2 mmol/l. Two individuals were excluded from further analysis in my study due to very high LDL-C levels of 14.2 mmol/l and 8.23 mmol/l (sample ID TNG0S and TEJ0C), as they could potentially skew the analysis as outliers. It is possible that these individuals have FH, rather than a polygenic form of hypercholesterolaemia, which also prompted their exclusion. Figure 3.1 B and C represents the LDL-C data for both the high (cases) and low (controls) groups, excluding the outliers.

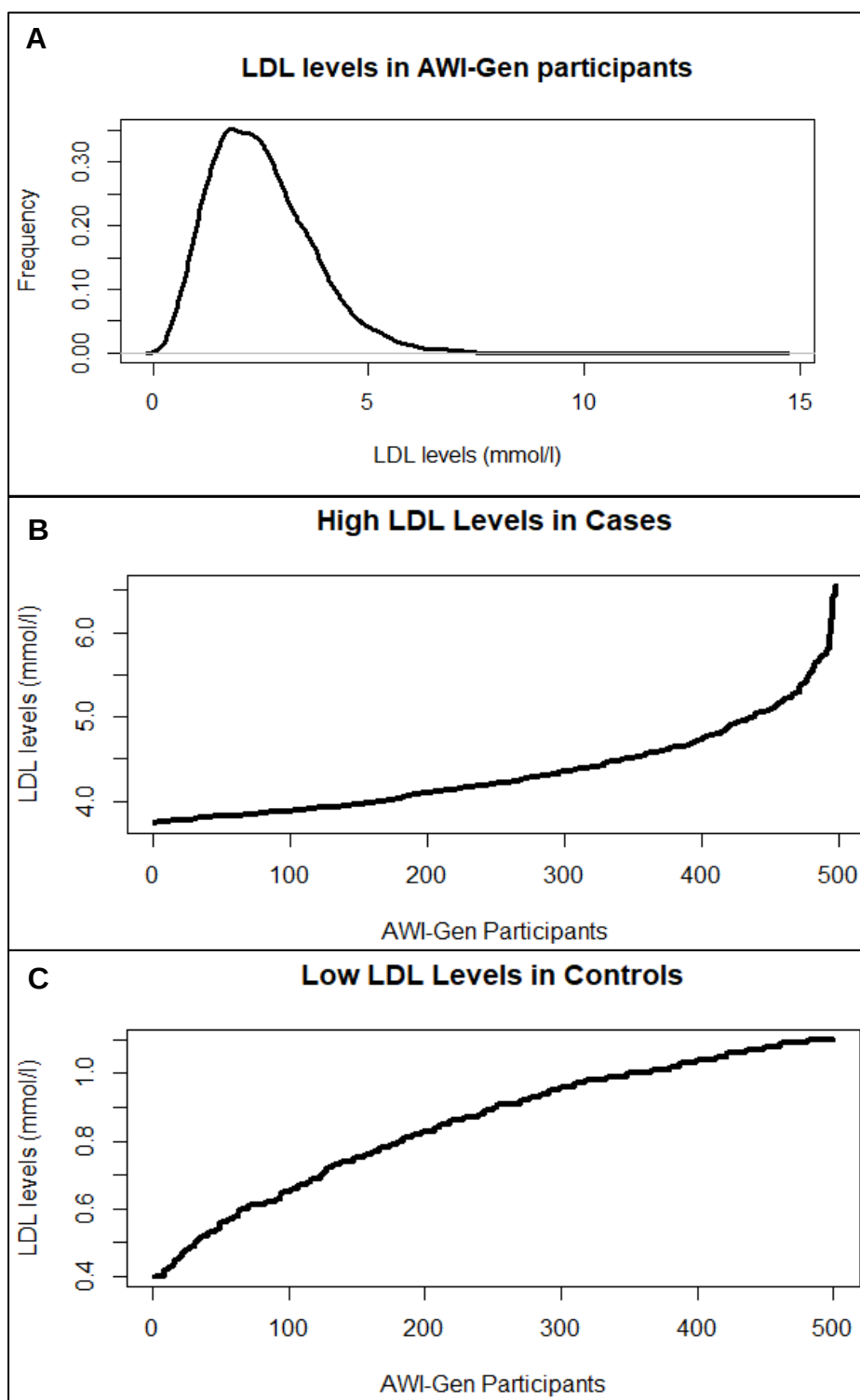


Figure 3.1: Graphs showing the distribution of LDL-C levels across AWI-Gen Participants. **A:** Distribution of LDL-C levels in 5940 AWI-Gen participants. **B:** High LDL-C levels in 498 AWI-Gen participants (cases). Two individuals were excluded due to extreme outlying high LDL-C levels. **C:** Low LDL-C levels in 500 AWI-Gen participants (controls).

3.2. Selection of SNPs based on function and frequency

There were 3541 variants identified in all eight populations in all four genes. Seventy-six variants that occurred in six or more of the eight populations were selected. The variants were filtered based on deleteriousness, the type of variant (missense, start/stop lost, regulatory) and MAF. Thirteen variants were chosen based on moderate deleteriousness as well as MAF (4%-20%) and an additional 16 were chosen solely on MAF (20%-45%). There were 29 variants that were finally selected to be tested. Figure 3.2 is a summary of the selection of SNPs.

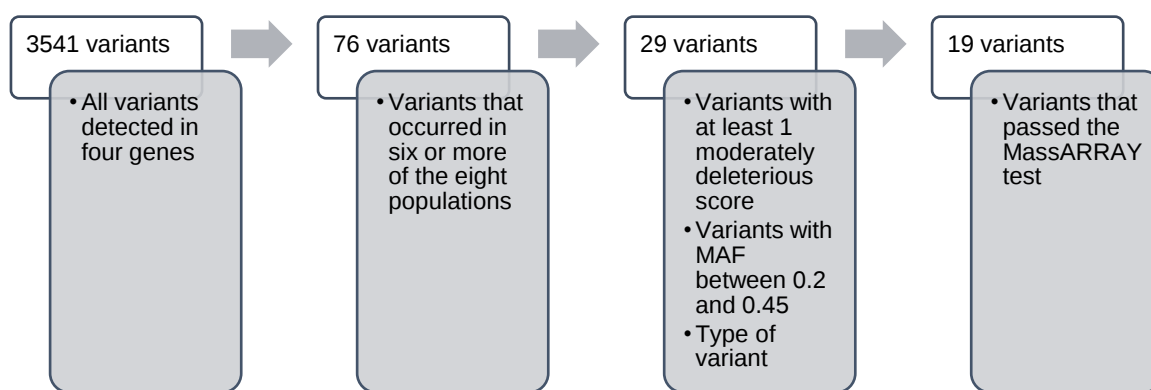


Figure 3.2: Flow diagram of variant filtering. Filtering and quality control measures were done on over 3000 variants, leaving 19 variants that were finally selected for genotyping.

3.3. Genotyping

An *in silico* MassARRAY assessment test was done on the 29 variants selected. Only 19 variants of the initial 29 variants remained to be finally genotyped. The 19 SNPs listed in table 3.5. Linkage disequilibrium (LD) was assessed for the 19 variants. None of the selected 19 variants were in LD with each other. A comprehensive table of all 29 SNPs that had initially been chosen can be found in Appendix C table 1 and 2. Table 3.6 is a summary of the genotype data from the 19 SNPs.

Table 3.5: List of the 19 SNPs that were genotyped in AWI-Gen participants using the MassARRAY platform

Gene	Variant	Minor allele (A1)	Type of variant	Amino acid change
<i>LDLR</i>	rs72658855	T	Synonymous	N to N
	rs11669576	A	Missense	A to T
	rs5929	T	Synonymous	P to P
	rs3826810	A	Regulatory region	N/A
	rs5925	C	Synonymous	V to V
	rs2569540	C	Intronic	N/A
	rs17242635	A	Intronic	N/A
<i>APOB</i>	rs12720855	G	Missense	S to P
	rs679899	A	Missense	A to V
	rs6752026	A	Missense	P to S
	rs12720820	C	Intronic	N/A
	rs12714102	C	Intronic	N/A
	rs3791981	G	Intronic	N/A
<i>PCSK9</i>	rs7552471	T	Synonymous	S to S
	rs4927193	C	Intronic	N/A
	rs45613943	C	Intronic	N/A
<i>LDLRAP1</i>	rs12071264	G	Intronic	N/A
	rs35910270	-	Deletion in 3'UTR	N/A
	rs13373894	A	Intronic	N/A

Genomic locations are reported using NCBI Build 37 (hg19)

Table 3.6: Results of genotyping for the 19 SNPs selected for genotyping

Gene	rs number	SNP Missingness	No. Individuals with data (n)	No. individuals with missing data	Minor Allele (A1)	HWE	MAF	High LDL-C					Low LDL-C				
								A1/A1 n(f)	A1/A2 n(f)	A2/A2 n(f)	MAF	HWE	A1/A1 n(f)	A1/A2 n(f)	A2/A2 n(f)	MAF	HWE
LDLRAP1	rs12071264	0.005	995	5	G	0.697	0.090	3	60	435	0.066	0.467	6	101	390	0.114	1.000
	rs13373894	0.006	994	6	A	0.000	0.465	33	364	101	0.432	<0.001	71	353	72	0.499	<0.001
	rs35910270	0.005	995	5	del	0.843	0.400	65	242	191	0.373	0.444	96	232	169	0.427	0.314
PCSK9	rs4927193	0.010	990	10	C	0.159	0.237	40	157	297	0.240	0.006	24	185	287	0.235	0.455
	rs7552471	0.009	991	9	T	0.691	0.086	2	84	408	0.089	0.409	4	75	418	0.084	0.766
	rs45613943	0.006	994	6	C	0.579	0.277	31	174	292	0.237	0.458	49	217	231	0.317	0.918
APOB	rs12720855	0.005	995	5	G	0.505	0.076	4	63	431	0.071	0.301	3	75	419	0.081	1.000
	rs12720820	0.399	601	399	C	0.003	0.305	9	133	154	0.255	0.002	31	153	121	0.352	0.103
	rs3791981	0.006	994	6	G	0.657	0.475	103	253	141	0.462	0.652	125	236	136	0.489	0.282
	rs12714102	0.198	802	198	C	0.434	0.460	77	199	117	0.449	0.684	98	188	123	0.469	0.136
	rs679899	0.005	995	5	A	0.026	0.123	18	110	370	0.147	0.012	5	89	403	0.100	1.000
	rs6752026	0.005	995	5	A	0.340	0.135	4	93	401	0.101	0.805	10	147	340	0.168	0.260
LDLR	rs72658855	0.005	995	5	T	0.234	0.031	0	20	478	0.020	1.000	2	37	458	0.041	0.197
	rs11669576	0.105	895	105	A	0.000	0.198	25	139	290	0.208	0.153	28	109	304	0.187	<0.001
	rs5929	0.006	994	6	T	0.878	0.117	2	98	398	0.102	0.145	12	106	378	0.131	0.168
	rs5925	0.004	996	4	C	0.141	0.152	15	134	350	0.164	0.625	14	111	372	0.140	0.132
	rs2569540	0.006	994	6	G	0.561	0.432	89	249	159	0.430	0.648	101	229	167	0.434	0.171
	rs17242635	0.271	729	271	G	0.000	0.484	0	363	13	0.483	<0.001	0	343	10	0.486	<0.001
	rs3826810	0.005	995	5	A	0.314	0.106	5	80	413	0.090	0.583	9	102	386	0.121	0.403

A1 = minor allele, A2 = major allele, MAF = minor allele frequency, HWE = Hardy-Weinberg equilibrium, Genomic locations are reported using NCBI Build 37 (hg19)

3.4. Quality control

3.4.1. Sample Quality Control

Five individuals (AB0484, AB0675, NRE0G, AB0459 and NBI0D) were removed from the analysis due to too many SNPs having failed genotyping (> 0.3 failure rate). Of the five samples that failed, three had no results, hence achieving a missingness score of 1 as no data was generated for these individuals. Two had failed due to bad DNA quality, as reported by Inqaba. Two more individuals were removed as they were outliers for LDL-C levels, as mentioned earlier. Table 3.7 shows the number of SNPs with missing data for each individual.

In total, seven individuals were excluded, leaving 993 individuals in the study.

Table 3.7: Samples removed due to SNP genotype failure QC criteria

Sample ID	Number of missing SNPs	Missingness	260/280
AB0484	19	1.00	1.93
AB0675	19	1.00	2.07
NRE0G	19	1.00	2.06
AB0459	18	0.95	2.00
NBI0D	18	0.95	1.83

260/280 ratio = purity of DNA. A ratio of ~ 1.8 is considered good quality

3.4.2. SNP Quality Control

Table 3.8 shows five SNPs that were removed due to having failed quality control measures. Four SNPs (*LDLR* rs11669576, *LDLR* rs17242635, *APOB* rs12720820 and *APOB* rs12714102) were removed due to high missingness ($>10\%$). A fifth SNP was removed because it was not in HWE (*LDLRAP1* rs1333894) with a p value <0.005 .

Table 3.8: SNPs removed due to failing quality control measures

Gene	rs number	HWE p value	Missingness
<i>LDLRAP1</i>	rs13373894	<0.005	0.006
<i>LDLR</i>	rs11669576	<0.001	0.105
	rs17242635	<0.001	0.271
<i>APOB</i>	rs12720820	0.003	0.399
	rs12714102	0.434	0.198

HWE = Hardy-Weinberg equilibrium

All SNPs showed no differential missingness, which checks whether controls and cases differ in missingness, i.e. that there is no stark difference in missingness between cases and controls. No SNPs exceeded the cut off of $p = 0.00001$. Standard MAF cut off was 1%. The lowest MAF was 3%, therefore no SNPs were flagged for removal. Data from a total of 14 SNPs were used for further analysis.

The 14 SNPs were analysed using a case-control approach. Case-control analysis was done because participants were chosen using a method based on extreme phenotyping. Participants with high LDL-C levels were classified as cases and those with low LDL-C levels as controls.

3.5. Population allele frequencies

Figure 3.3 depicts the minor allele frequencies of 14 SNPs that were evaluated in logistic regression (section 3.6). The African (green bar) and European (red bar) frequencies were taken from KGP, and the Eastern ($n=306$), Western ($n=323$) and Southern ($n=364$) frequencies were calculated from the genotype data of the AWI-Gen participants (bars are shades of blue). The frequencies of the alleles at some loci were very similar and different in others between the AWI-Gen participant regions, indicating some variation between the Eastern, Western and Southern African populations (figure 3.3) However, there is a stark difference in allele frequency between most of the African population frequencies and the European frequencies, supporting the notion that African populations often have different allele frequencies compared to non-African populations.

A p value cut-off of <0.05 was used to indicate that the frequencies between East, West and South African populations differ significantly. Five SNPs are similar in frequency across the three regions. Allele frequencies for nine SNPs do differ across all three African regions, namely: *LDLR* rs72658855, rs5925, *APOB* rs679899, rs3791981, *PCSK9* rs7552471, rs4927193, rs45613943, *LDLRAP1* rs12071264 and rs35910270. It is interesting to note that the differences usually are because of the allele frequencies in West Africans being more different to East and South African frequencies.

The allele frequencies were compared across the three African regions (East, West and South Africa) and are shown relative to the combined African and European population in KGP (Figure 3.3). Nine of the 14 loci showed significant frequency

differences within Africa. Generally, however, the frequencies in East and South Africa are more similar to one another, when compared to West Africa, in line with their demographic histories. A table comparing pairwise frequencies can be found in Appendix C (table 4).



Figure 3.3: Histograms showing frequencies of 14 variants that were genotyped and passed QC in each of the four genes. The African (green) and European (red) frequencies were obtained from KGP. East, West and Southern African (shades of blue) frequencies were calculated using genotyping data. The p values indicate frequency differences across East, West and South African populations. A significant difference across East, West and South African population frequencies ($p < 0.05$) is indicated with an asterisk (*).

3.6. Case-Control Association Analysis

3.6.1. Logistic regression

Logistic regression was carried out to determine whether there was an association between the variants and LDL-C levels, where cases were participants with high LDL-C levels, and controls were those with low LDL-C levels. Allelic association was carried out using PLINK v.1.9. Six SNPs (bold and highlighted in table 3.9) were significantly associated with LDL-C, tested under multiple correction using Benjamini-Hochberg (FDR). FDR reduces the chances of a SNP falsely being selected as associated by statistical means. Variants were considered as significant with an adjusted $p < 0.05$ which is indicated in bold and shaded in the table below.

Table 3.9: Allelic association of 14 SNPs with and without adjustment for multiple testing using FDR in 993 individuals

Gene	SNP	Unadjusted p value	Adjusted p value	Minor allele (A1)	Frequency in cases	Frequency in controls	OR	L95	U95
APOB	rs6752026	<0.0001	0.0002	A	0.1018	0.1680	0.5613	0.4309	0.7314
PCSK9	rs45613943	0.0001	0.0005	C	0.2374	0.3169	0.6709	0.5503	0.8180
LDLRAP1	rs12071264	0.0002	0.0011	G	0.0665	0.1137	0.5557	0.4046	0.7632
APOB	rs679899	0.0013	0.0047	A	0.1472	0.0998	1.5570	1.1860	2.0430
LDLR	rs72658855	0.0065	0.0181	T	0.0202	0.0413	0.4783	0.2782	0.8223
LDLRAP1	rs35910270	0.0148	0.0346	del	0.3730	0.4266	0.7997	0.6680	0.9573
LDLR	rs3826810	0.0297	0.0595	A	0.0907	0.1207	0.7267	0.5445	0.9699
LDLR	rs5929	0.0424	0.0741	T	0.1018	0.1310	0.7516	0.5702	0.9909
LDLR	rs5925	0.1448	0.2252	C	0.1633	0.1398	1.2010	0.9388	1.5350
APOB	rs3791981	0.2407	0.3370	G	0.4626	0.4889	0.8999	0.7544	1.0730
APOB	rs12720855	0.4059	0.5166	G	0.0716	0.0815	0.8689	0.6237	1.2100
PCSK9	rs7552471	0.6389	0.7454	T	0.0894	0.0835	1.0780	0.7877	1.4750
PCSK9	rs4927193	0.8367	0.8464	C	0.2388	0.2349	1.0220	0.8306	1.2580
LDLR	rs2569540	0.8464	0.8464	G	0.4293	0.4336	0.9826	0.8226	1.1740

OR = odds ratio, L95 = lower 95% confidence interval, U95 = upper 95% confidence interval, Genomic locations are reported using NCBI Build 37 (hg19)

Logistic regression was carried out using PLINK v.1.9. Sex, BMI, glucose levels and the region the participant originated from were identified as potential confounders.

Sex, BMI and glucose levels were significantly different between the cases and

controls. The region (East, West and South Africa) the participant originated from was also identified as a confounder because we acknowledge that different African populations show population stratification. These factors were, therefore, corrected for. Age was not a confounder as it was not significantly different between cases and controls. Two SNPs, (bold and shaded in table 3.10) *LDLRAP1* rs12071264 and *APOB* rs6752026, remain significant ($p < 0.05$) after adjusting for above mentioned covariates. Both SNPs have OR values below one (meaning that the minor allele (A1) is associated with low LDL-C levels; with the major allele (A2) is associated with high LDL-C levels).

Table 3.10: Logistic regression with 14 SNPs, adjusted for covariates sex, BMI, glucose levels and region in 993 individuals

GENE	SNP	Minor Allele (A1)	N	OR	L95	U95	P value
<i>LDLRAP1</i>	rs12071264	G	993	0.5866	0.3914	0.8791	0.0097
<i>APOB</i>	rs6752026	A	993	0.6898	0.4847	0.9816	0.0391
<i>LDLR</i>	rs72658855	T	993	0.4983	0.2463	1.0080	0.0527
<i>LDLRAP1</i>	rs35910270	G	993	0.8004	0.6341	1.0100	0.0608
<i>APOB</i>	rs679899	A	992	1.3620	0.9691	1.9130	0.0753
<i>LDLR</i>	rs5929	T	992	0.7452	0.5258	1.0560	0.0984
<i>PCSK9</i>	rs7552471	T	989	1.3560	0.8964	2.0510	0.1492
<i>PCSK9</i>	rs45613943	C	992	0.8280	0.6401	1.0710	0.1508
<i>LDLR</i>	rs3826810	A	993	0.8147	0.5688	1.1670	0.2637
<i>PCSK9</i>	rs4927193	C	988	1.1460	0.8811	1.4910	0.3093
<i>APOB</i>	rs3791981	G	992	0.9115	0.7272	1.1420	0.4211
<i>LDLR</i>	rs5925	C	993	1.0400	0.7653	1.4130	0.8022
<i>APOB</i>	rs12720855	G	993	1.0470	0.6879	1.5950	0.8288
<i>LDLR</i>	rs2569540	G	992	0.9810	0.7810	1.2320	0.8693

N = no. of individuals genotyped, OR = odds ratio, L95 = lower 95% confidence interval, U95 = upper 95% confidence interval, Genomic locations are reported using NCBI Build 37 (hg19)

Figure 3.4 is a recap of the study design. The flow diagram in figure 1.3 was reproduced and additional information was added to indicate the results generated for each step in the study.

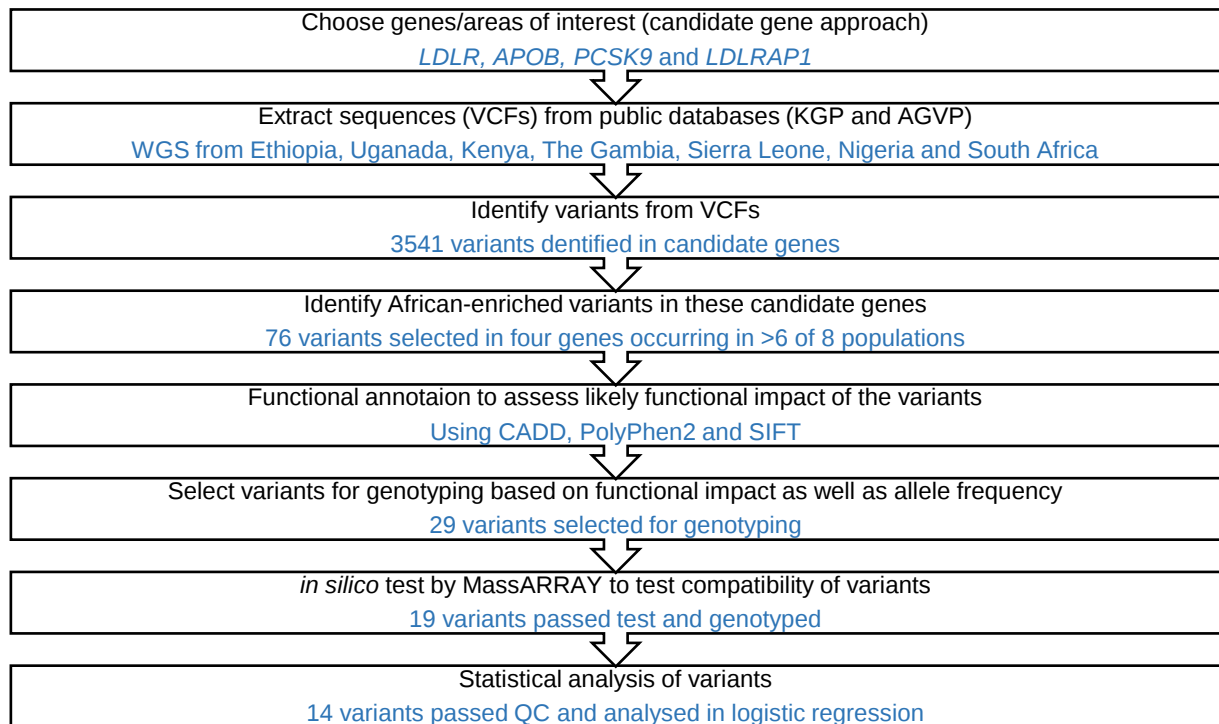


Figure 3.4: Flow diagram of study outline with results. This flow diagram is a replica of figure 1.3, with the results of each step given in blue.

Using the data for the 14 SNPs from table 3.10, a forest plot was generated using the OR and 95% confidence interval values (figure 3.5). The forest plot distinctly shows how the minor allele for two significant SNPs is associated with low LDL-C levels, and that the remaining 12 SNPs are not significantly associated with LDL-C. The top two SNPs are the significant variants because the CI lines do not cross the $x=1$ (red line) intercept. The CI lines of the remaining 12 SNPs do cross the intercept, indicating that they are not significantly associated with LDL-C levels.

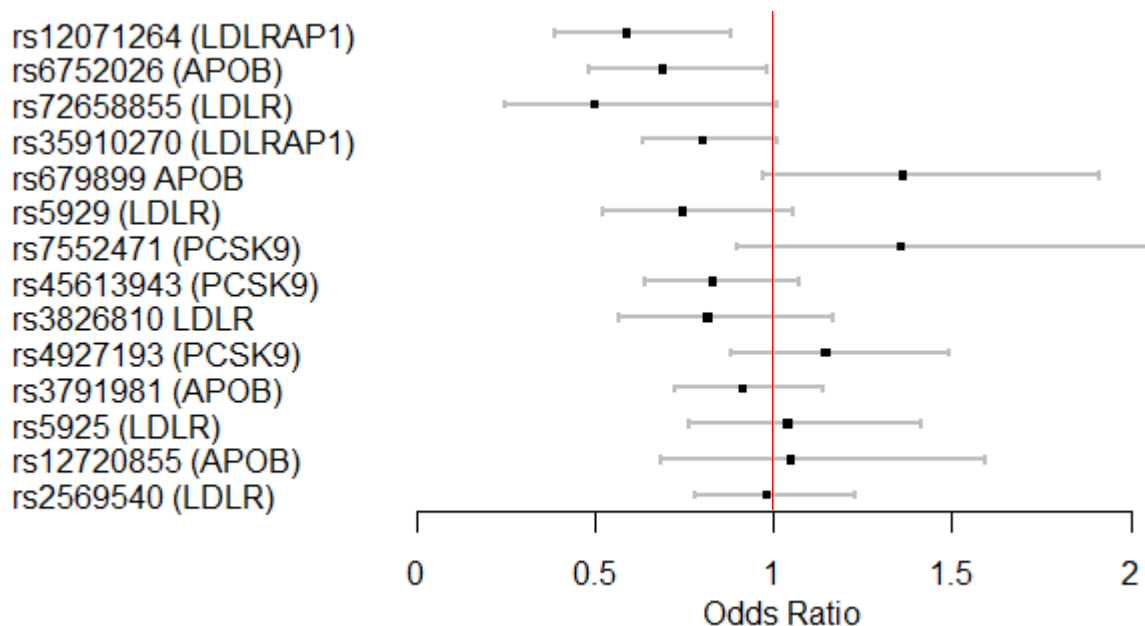


Figure 3.5: Forest plot for the 14 SNPs that were genotyped and passed QC. The plot shows data for SNPs after adjusting for covariates: sex, BMI, fasting glucose levels and region (logistic regression). Bars represent 95% confidence interval (CI). The first two SNPs are significantly associated with LDL-C levels as the CI bars do not cross $x=1$ (red line). Since the OR is <1 for the first two SNPs, the minor alleles of the two SNPs are significantly associated with low LDL-C levels: rs12071264 and rs6752026. The remaining 12 SNPs are not significantly associated with LDL-C levels because the CI bars cross the $x=1$ line.

Figure 3.6 shows the association of the genotypes with LDL-C levels. The genotype AA for rs12071264 *LDLRAP1* (OR 0.5866) correlates with higher LDL-C levels (figure 3.6 A), supporting the allelic association of the common A allele with increased LDL-C. With more G alleles (heterozygous and minor allele homozygous genotypes) the trend was associated with lower LDL-C levels, suggesting that the presence of the minor allele G contributes to lower LDL-C levels in this populations.

The other associated SNP, rs6752026 *APOB* (OR 0.6898), shows that the genotype GG correlates with high LDL-C levels (figure 3.6 B) supporting the allelic association of common allele G with higher LDL-C levels. The heterozygous and homozygous minor allele A genotypes are associated with lower LDL-C levels.

There is a decrease in LDL-C only when the minor allele is present (in both the heterozygous and homozygous genotype) for both SNPs. This suggests that these alleles have a dominant mode of action.

The LDL-C levels were not normal for each genotype group, therefore a Kruskal-Wallis test was carried out to assess whether LDL-C levels were statistically different ($p < 0.05$) across the genotype groups for both SNPs. Thereafter, a post-hoc contrast test was done to determine which 2 groups were statistically different ($p < 0.05$) from each other. For rs12071264 and rs6752026, the Kruskal-Wallis p values were 0.0007 and 0.0008, indicating that the LDL-C levels are different across the three genotypes. The post-hoc contract test showed a statistical difference between rs12071264 AA and GA ($p = 0.0011$) and rs6752026 GG and AG ($p = 0.0148$).

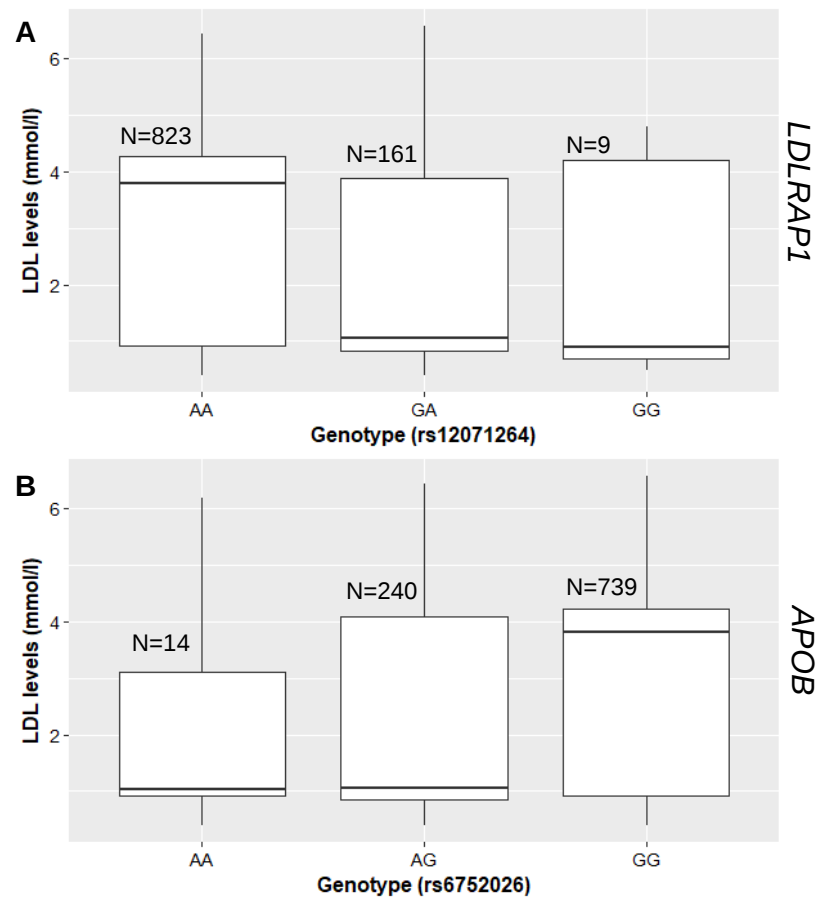


Figure 3.6: Box and whisker plots of the 2 SNPs significantly associated with LDL-C levels after logistic regression. The 993 AWI-Gen individuals with low and high LDL-C are included in the plots. **A:** rs12071264 *LDLRAP1* (OR 0.5866), shows how LDL-C levels decrease with presence of minor allele G. **B:** rs6752026 *APOB* (OR 0.6898) shows how LDL-C levels decrease with minor allele A. For each locus, having at least one minor allele is associated with lower LDL-C.

3.6.2. Polygenic Risk Score

A polygenic risk score (PRS) was calculated as per the methods section (2.7.2) using the six SNPs significantly associated with LDL-C with $p < 0.05$ after correction for multiple testing, but before including the potential confounders. Significantly associated SNPs before adjusting for confounders were used so more SNPs could be included in generating the PRS. In the plot, the number of high risk alleles is counted for each individual and can range from 0 to a maximum of 12 alleles. The plot (figure 3.7 A) shows the frequency (proportional number of individuals) of cases and controls for each score. The curve of the controls (low LDL-C) is shifted to the right (higher score), as expected. This indicates that a higher score is correlated with lower LDL-C levels (controls) and therefore is suggested to correlate with a protective effect to CVD. The two groups are significantly different from each other ($p = 0.001$).

Figure 3.7 B clearly shows the correlation of the generated risk score and LDL-C levels. It is apparent that individuals with a greater number of high risk alleles have lower LDL-C levels. Alleles individually have a small effect on the phenotype, but when considering alleles across several associated genetic loci, the additive effect is clearly seen to influence the phenotype.

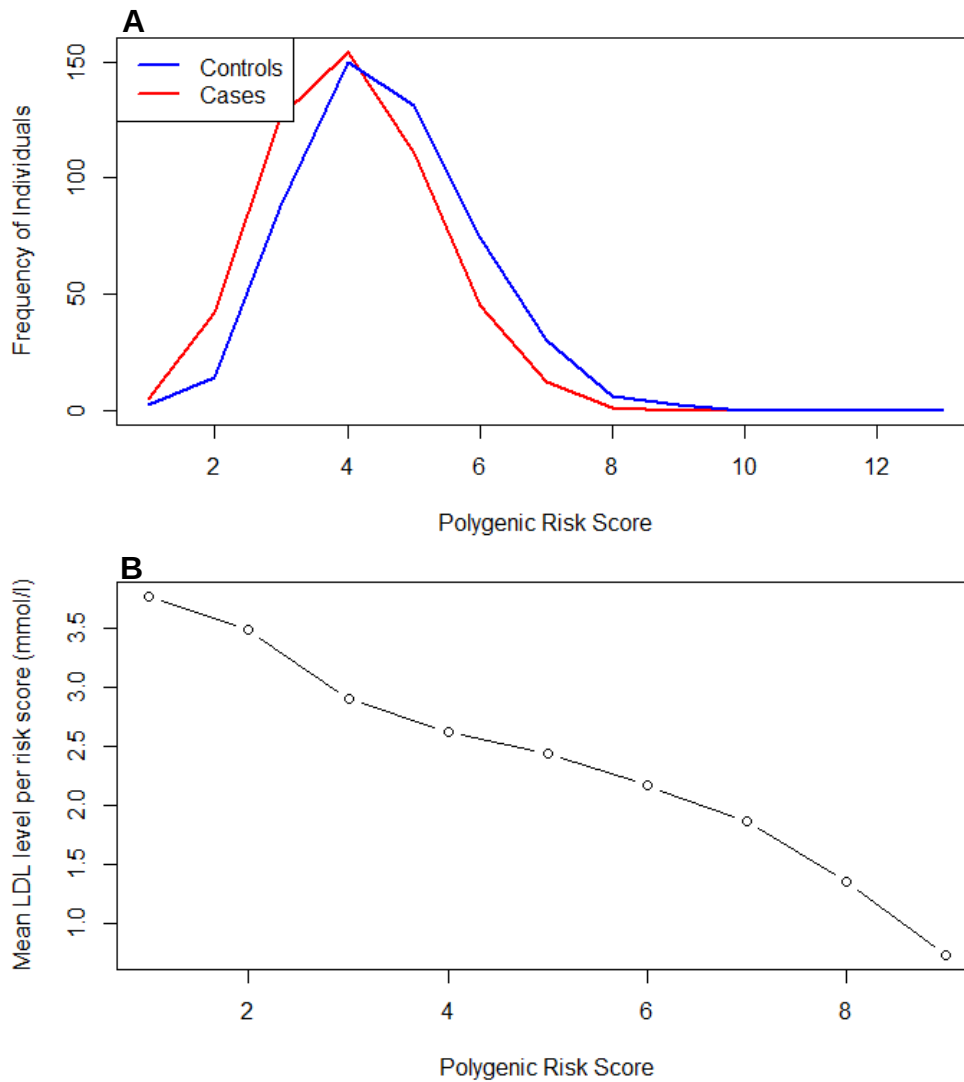


Figure 3.7: Correlation of the polygenic risk score (PRS) with LDL-C levels in 993 individuals. Six LDL-C associated SNPs $p < 0.05$ after adjustment for multiple testing but before adjusting for covariates were included in the calculation of the PRS. **A:** PRS calculated using six SNPs. Plot shows the frequency of cases and controls for each score. The curve of the controls is shifted to the right, indicating that in controls the LDL-C levels decrease with the addition of alleles associated with lower LDL-C levels (either common or minor allele). **B:** Plot of risk score against mean LDL-C level per risk score. It was found that with the addition of each allele associated with lower LDL-C levels (common or minor allele), the mean LDL-C level of the participants decreased.

4. Discussion

The aim of this study was to identify variants that may be associated with variation in LDL-C levels in black African populations. Four genes were investigated: *LDLR*, *APOB*, *PCSK9* and *LDLRAP1*. Variants in these genes have been found to be associated with LDL-C levels as a quantitative trait in various populations worldwide. In addition, pathogenic variants in these genes are known to be associated with monogenic forms of dyslipidaemia, more specifically familial hypercholesterolaemia (FH). Using a computational approach, variants were identified in the four genes in WGS data in African individuals sourced from public databases (AGVP and KGP) where WGS data are freely available.

Over 3500 variants were identified across the four genes. These variants were then filtered, using a computational approach which considered base-line frequency of the minor allele across all eight African populations studied, as well as predicted deleteriousness of the variant. Following the filtering process, 29 variants were selected for laboratory genotyping, but only 19 were genotyped because 10 variants failed the MassARRAY *in silico* test. The 1000 participants were chosen by applying a case-control study design, where participants were specifically chosen because of either high or low LDL-C levels. The 19 variants were genotyped in a group of 998 AWI-Gen participants: 500 with low LDL-C levels (controls) and 498 with high LDL-C levels (cases). Two individuals were removed from the high LDL-C group because they were identified as outliers. Selecting participants using an adapted extreme phenotyping method enriches the presence of variants that contribute to a quantitative trait and increases the power for association as compared to random sampling (Barnett, Lee and Lin, 2013; Xu *et al.*, 2018). This selection method also reduces the number of individuals needed for genotyping to detect a specific effect size. Emond *et al.*, (2012) used an approach of extreme phenotyping to successfully identify a gene that modifies chronic infection in cystic fibrosis in 91 individuals. Yang *et al.*, (2015) found that extreme phenotyping was more effective for identifying variants associated with maize kernel phenotypes than conventional GWAS.

Logistic regression analysis was subsequently carried out using a case-control study design to determine whether the selected SNPs were associated with LDL-C levels.

4.1. Characterisation of participant phenotype

The AWI-Gen participants (approximately 10 000 in total) were sampled from six geographic sites (Agincourt, Dikgale, Navrongo, Nairobi, Nanoro, Soweto) representing populations across East, West and South Africa (Ramsay *et al.*, 2016). For this particular study, we selected 1000 participants from the 10 000 for further investigation. The individuals were chosen based on their LDL-C levels: 500 with low LDL-C levels and 500 with high LDL-C levels. As the selection criterion was LDL-C, irrespective of site of collection, the individuals included in this study were not evenly distributed across the six sites. Different African populations harbour different levels of genetic diversity (Durbin *et al.*, 2010; Pickrell *et al.*, 2012), predicting that the participants from the different sites may be genetically different to each other. Therefore, the pooling of six different groups that potentially exhibit population stratification into one sample group could be a confounder. The confounder of population substructure could possibly influence the detection of associated variants with LDL-C.

However, when participants were sorted into geographic regional groups (East, West and South Africa) instead of site (Agincourt, Dikgale, Navrongo, Nairobi, Nanoro, Soweto), the participants were evenly distributed across the sample of 1000 participants in the study. Though, to control for population sub-structure between regions, participants' regions were used as a covariate in logistic regression.

In addition to LDL-C levels, the following phenotype data was used in this study: age, BMI, sex, glucose levels, HIV status and HIV treatment. Age was the only phenotype variable to show no significant difference between cases and controls and was therefore excluded as a covariate.

HIV-related information from AWI-Gen was limited and many data points were missing (table 3.4). Consequently, HIV-related variables could not be factored into the combined analysis model. From the HIV data available to us, there are more participants who are HIV positive in the low LDL-C level (control) group. It seems that this correlates with a study by Vos *et al.* (2017) where HIV positive individuals were found to have lower LDL-C levels when compared to a control group. In addition, more participants in the high LDL-C level group (cases) were treated for HIV, suggesting that perhaps HIV treatment could be related to higher LDL-C levels (Anastos *et al.*, 2007). To reiterate, HIV infection status and treatment could not be

used as a covariate in logistic regression analysis due to many missing data points for the 1000 participants of this study.

The other three phenotype variables, namely, sex, BMI and fasting glucose levels, showed a significant difference between cases and controls (table 3.2) with p-values < 0.0001. They could be possible confounders; therefore, they were used as covariates in subsequent analyses.

Using the MassARRAY, 19 SNPs were genotyped. Following quality control of the genotype data, seven individuals were removed from the dataset. Two individuals were removed as outliers with extremely high LDL-C levels and a further five individuals were removed due to high SNP genotyping missingness. Five SNPs were removed due to high missingness and low HWE. This left 993 individuals and 14 SNPs to be further analysed.

The small number of exclusions of participants (seven) and SNPs (five) reflects the high quality of DNA.

4.2. Genetic Associations with LDL-C

This study investigated AWI-Gen participants in a case-control study design based on clinical cut-offs of LDL-C levels, with the aim of identifying variants associated with LDL-C in African populations. This study aimed to determine genetic contributions to a multifactorial trait, using individuals with extreme LDL-C levels.

There are studies that have shown that allelic variants in four genes, *LDLR*, *APOB*, *PCSK9* and *LDLRAP1*, have been associated with both polygenic and monogenic forms of dyslipidaemia. Kathiresan *et al.*, (2009) found that in a GWAS of 20 000 individuals, allelic variants of *LDLR*, *APOB* and *PCSK9* were associated with LDL-C levels. A study that was done on 841 Amish individuals found that a variant in *APOB* was associated with an increase in LDL-C (Shen *et al.*, 2010). Loss of function variants in *LDLRAP1* were found to be associated with raised lipid levels in six Spanish patients (Sánchez-Hernández *et al.*, 2018). Therefore, allelic variants associated with LDL-C levels have been identified in all four genes when examining LDL-C as a multifactorial trait.

In African populations, or populations with African ancestry, few studies have been undertaken looking at association with alleles at the four genes being investigated and LDL-C levels. A GWAS carried out on 3263 individuals of African descent from

different populations identified alleles in *LDLR* associated with LDL-C levels (Willer *et al.*, 2013). Two nonsense variants in *PCSK9* were found to be associated with low LDL-C levels in a sample of 128, of which 64 individuals were African American (Cohen *et al.*, 2005), and a nonsense variant identified in Zimbabwean women associated with low LDL-C levels (Hooper *et al.*, 2007).

4.2.1. Genetic association analysis

Following quality control, 496 cases with high LDL-C levels and 497 controls with low LDL-C levels were included in the association analysis. These participants were genotyped for 19 SNPs that were selected from data on African populations with a view to choosing the SNPs based on potential functional impact of the variant allele, and with substantive frequencies of the variant allele in African populations. After quality control, only 14 SNPs were included for analyses going forward. The allelic association analysis showed six SNPs (*APOB* rs6752026, rs679899, *LDLR* rs72658855, *LDLRAP1* rs12071264, rs35910270 and *PCSK9* rs45613943) were significantly associated with LDL-C levels ($p < 0.05$) after correcting for multiple testing using the Benjamini-Hochberg method (table 3.9) and before adjusting for potential confounders. Three of the associated SNPs (*APOB* rs6752026, rs679899, *LDLR* rs72658855) are missense variants, *LDLRAP1* rs12071264 is a variant in intron 5, close to the intron:exon splice junction, *LDLRAP1* rs35910270 is a 1 bp deletion variant in the 3' untranslated region and *PCSK9* rs45613943 is found towards the end of intron 5. At least one variant in each of the four genes showed allelic association with LDL-C, suggesting that genetic variation at all four genes likely contribute toward LDL-C level variation in East, West and South African populations.

4.2.1.1. Significant association with two variants

Logistic regression was carried out with adjustment for four potential covariates (sex, BMI, fasting glucose levels and participant region). After adjustment, four SNPs that showed allelic association were no longer statistically associated with LDL-C levels in the logistic regression (table 3.10). Only two variants remained significantly associated after adjusting for the four covariates: *LDLRAP1* rs12071264 ($p < 0.01$) and *APOB* rs6752026 ($p < 0.04$).

The first variant is in *LDLRAP1*: rs12071264, c.533-22A>G (OR: 0.5866, 95%CI: 0.40-0.88). The variant is found in the middle of the gene in intron 5. The minor G allele was found to be associated with low LDL-C levels and occurs at a frequency of

0.1415 in the KGP combined African sample. This variant is extremely rare in non-African populations (absent in the KGP European samples) and is not represented in frequently used GWAS arrays. ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) was used to check if this variant has any previous associations with LDL-C. However, no clinically relevant information for rs12071264 was found. It must be noted that ClinVar records studies with variants primarily associated with monogenic traits and that have mainly been carried out in populations other than African. It will, therefore, not always be predictive of the effect for polygenic associations or for variants in African populations. There is no information regarding this SNP on the GWAS catalogue (<https://www.ebi.ac.uk/gwas/home>).

Using SplicePort (Dogan *et al.*, 2007), this SNP (c.533-22A>G) was found to be close to a splice site. Variants that are found a few base pairs away from a splice site can affect the way the gene is spliced. Consequently, the altered transcript of the *LDLRAP1* gene could contribute toward lowered LDL-C levels. The major allele in all populations is associated with higher LDL in Africans. The fact that the minor allele is associated with lower levels of LDL-C may suggest that this variant acts as a gain of function variant associated with lower LDL-C, or causes allelic insufficiency of *LDLRAP1* due to alternate splicing and that this contributes to lower LDL levels in African populations.

The true effect of this variant remains unknown. Variants in *LDLRAP1* have not previously been shown to associate with low LDL-C levels, but variants causing high LDL-C have been identified (Sánchez-Hernández *et al.*, 2018). This SNP is absent in European populations and was likely not included in GWAS arrays used to date. This may therefore be a unique African LDL-C associated variant. There are no LDL-C association with *LDLRAP1* in African populations.

The second variant significantly associated with LDL-C levels was a missense variant in exon 5 of the *APOB* gene: rs6752026, c.433C>T (OR: 0.6898, 95%CI: 0.69-0.48). *APOB* is transcribed from the reverse strand. In the literature the variant is listed as a C to T variant, even though the actual variant is G to A. This SNP is a proline to serine missense variant and the A allele (encoding the serine) is associated with lower LDL-C levels. When a variant changes the amino acid property/charge at a specific position, the structure and/or function of the protein may be affected (Sauna and Kimchi-Sarfaty, 2011). The amino acid change here is from a polar amino acid to a non-polar amino acid. It is reported as benign by ClinVar, although it had a

“probably damaging” PolyPhen2 score (0.919) and a “deleterious” SIFT score (0.03) suggesting that this SNP could be deleterious as a consequence of altering the protein structure. The consequence of the variant could be a gain of function variant. This would result in the APOB protein binding with a higher affinity to LDLR, decreasing levels of LDL-C in the blood plasma.

ClinVar does not usually report on genetic associations with complex traits, and this may be why it is reported as benign, even though other prediction tools suggest a high probability of being damaging. This SNP shows association with the trait, and due to the predicted functional impact it may contribute to altering the phenotype. As most available published studies are based on European populations, it is highly likely that this variant has not been previously reported because it occurs at very low frequencies in European KGP populations (0.1%) and is not represented on GWAS arrays.

Even though there are no studies reporting on this *APOB* variant, previous studies have identified other *APOB* variants to be associated with LDL-C in non-African populations. A missense variant causing FH was found in *APOB* in a French family (Elbitar *et al.*, 2018). In contrast, two studies in a Spanish and French population found deleterious variants that are thought to cause low LDL-C and APOB protein levels (Martín-Morales *et al.*, 2013; Rimbart *et al.*, 2016). Studies on variants in *APOB* affecting lipid levels in African populations were not found in the literature.

In association studies, very often the associated variant is not the casual/functional variant itself. So, it is highly likely that these two SNPs, if not causal, may be in LD with the actual functional variants with functional impact on the phenotype (Clarke *et al.*, 2011).

4.2.1.2. Associations of *LDLR*, *APOB*, *PCSK9* and *LDLRAP1* with LDL-C in the literature

In this study, allelic association with variants in these four genes have shown suggestive or significant association with LDL-C levels. For two of the genes investigated, *LDLR* and *PCSK9*, no SNPs remained significantly associated ($p < 0.05$) with LDL-C levels after logistic regression and adjusting for covariates.

The effect of the PCSK9 protein on LDL-C levels is a relatively new discovery. However, it is now well documented the functional *PCSK9* variants are associated with both high and low levels of LDL-C in many populations (Abifadel *et al.*, 2003;

Cohen *et al.*, 2005). The fact that none of the *LDLR* and *PCSK9* variants tested in this study were found to be significantly associated in logistic regression does not indicate that these genes are not involved in variation in LDL-C levels in Africans.

There is relatively little literature available for comparative studies. A Korean study (Lee *et al.*, 2017) aimed to identify rare variants in 22 patients with extremely low LDL-C. The study sequenced both genes and used PolyPhen2 and SIFT as well as Mutation Taster to functionally annotate the variants. They found eight rare variants in two candidate genes (*APOB* and *PCSK9*) that were associated with low LDL-C.

Another study (Lange *et al.*, 2014) also used extreme phenotyping to select their participants, approximately half of which were European American, and performed exome sequencing. They used ANNOVAR and GENCODE to annotate the variants as nonsense, splice, missense, synonymous, UTR and noncoding. In addition to finding known and novel variants in three of the genes we investigated (*LDLR*, *APOB* and *PCSK9*), Lange *et al.*, (2014) discovered a locus (*PNPLA5*) previously not associated with LDL-C. The *PNPLA5* protein appears to have a role in lipogenesis in the liver (Wilson *et al.*, 2006), however, the exact function of the protein remains unclear. Variants in *APOB* and *PCSK9* were found to be associated with low LDL-C levels, and variants in *LDLR* with high LDL-C.

As already mentioned, Sánchez-Hernández *et al.*, (2018) and Pirillo *et al.*, (2017) identified several duplications, deletions and SNPs that cause high LDL-C levels in Spanish and Italian populations, respectively.

To identify rare variants with a large effect on LDL-C levels in an African setting, an approach that would include sequencing of a gene panel with genes known to be involved in lipid levels, in a sample of participants chosen by clinical cut-offs of LDL-C levels could lead to the identification of alleles with a large contribution to the phenotype. Few studies have been carried out on black African populations exploring the effect of variants in *LDLR*, *APOB*, *PCSK9* and *LDLRAP1*. Another approach to identify rare variants with large effects on LDL-C in African populations, would be to carry out an exploratory GWAS to search for novel high-effect loci contributing to the trait.

A study carried out on 1860 black South African individuals (Setswana speaking) aimed to identify variants associated with LDL-C levels. Van Zyl *et al.*, (2014) investigated whether variants in *LDLR* associate with LDL-C levels in these black

South Africans. The *LDLR* gene was sequenced and two novel LDL-C associated variants were identified in 30 individuals. The study showed that a C to T promoter variant (rs17249141) was associated with lower LDL-C levels. Four variants were associated with higher LDL-C levels; one intron variant (rs2738447) and three variants in the 3' UTR (rs14158, rs2738465 and rs3180023).

In addition to causing dyslipidaemia in other populations, a *PCSK9* nonsense variant, C679X (rs28362286) showed an association with low LDL-C levels in a cohort of Zimbabwean women (Hooper *et al.*, 2007). The frequency of this SNP in the cohort of Zimbabwean women was 3.7%. This variant was initially identified in my study while carrying out functional annotation; however, it was excluded for genotyping due to its low allele frequency in the African populations (0.008) and it would have had insufficient power in my study.

The MAF used to select the variants was taken from the combined African frequency reported in dbSNP (which come from KGP). This was the most suitable proxy population to use for the African frequency estimate during the SNP selection process.

In conclusion, my study has provided suggestive evidence of the role of genetic variation of all four genes in LDL-C levels in black Africans. However, the sample size was underpowered to detect modest to small effects. After adjusting for both multiple testing and potential confounders, only 2 out of the 14 SNPs tested in this study remained significantly associated with LDL-C levels. In both cases the rare allele was associated with lower LDL-C. This may suggest that the allelic variants have a gain of function impact, or are in close LD with functional variants, that contribute to decreasing LDL-C levels in Africans.

4.2.2. Gene-Environment interactions

LDL-C levels are influenced by genetic variants at many different loci; however, the phenotype is a complex trait also influenced by other factors. Lipid levels can be influenced by sex, BMI and age, as well as lifestyle choices such as diet. Lipid levels have been reported to have an estimated heritability ranging between 40 and 60% (Weiss *et al.*, 2006). GWAS studies of very large sample sizes have generally explained only 10-12% of the variability in LDL-C levels (Teslovich *et al.*, 2010). Some of the missing heritability could be explained by gene-environment interactions and gene-gene interactions (also referred to as epistasis) (De *et al.*, 2017).

Gene-environment interactions have been identified in some studies, where medication, like hormone replacement therapy, and lifestyle choices (exercise and diet) can affect lipid levels (Hagberg, Wilund and Ferrell, 2000). A significant correlation was found between dietary cholesterol and plasma cholesterol levels, where increasing dietary cholesterol increases lipid levels (Kim *et al.*, 2013). If an increase in dietary cholesterol results in an increase in plasma cholesterol levels, then it follows that one may expect a decrease in dietary cholesterol, in addition to an increase in daily exercise, to result in a decrease in plasma cholesterol. As an exception, this would not necessarily be the case for monogenic causes on high LDL-C, like FH, that are primarily the result of a highly penetrant pathogenic mutation.

It is interesting that African American populations generally have lower lipid levels than non-African populations (Bentley and Rotimi, 2012). Diet and exercise could contribute to explaining the generally low levels in African populations – as dietary cholesterol tends to be low, and daily activity is higher than in more urban populations – but a genetic predisposition to lower LDL-C cannot be excluded. Data on diet and physical exercise on the AWI-Gen participants was not available at the time of this study, and this information was therefore not used as potential covariates in logistic regression. In addition, the data is self-reported and may not be an accurate reflection of behaviour.

In the AWI-Gen study, among the 1000 selected participants, there are more individuals with high LDL-C levels in East Africa than there are in West Africa. This could be due to the type of diet patterns followed in East Africa. Although fat intake is lower in African populations than in the global west, East African populations seem to consume more fat compared to West Africans. Burlingame (2003) reports that East African populations get 30-32% of energy from fat, whereas populations in West Africa get 23% of their energy from fat. Diet, could therefore, be one explanation for this skewed distribution.

HIV data on the AWI-Gen participants was not complete. It is known that infection and ARV treatment affect LDL-C levels and therefore this is an important omission and could influence the outcome of this study. More HIV-related data should be included in future studies to properly assess the impact of HIV status and treatment on LDL-C levels in African populations.

Alternatively, there could be genetic predisposing factors to low LDL-C levels in different African populations. In our study, participants from East and South African differ in frequencies to West African frequencies for most of the SNPs (figure 3.2 and appendix C table 4). There could be genetic predisposing factors that contribute differently to the phenotype in each population.

However, the LDL-C level landscape in African populations is likely due to a combination of genetic predisposing factors, as well as environmental effects. For both significantly associated SNPs identified in this study, it was the major alleles that were associated with higher LDL-C levels. Although this may suggest that the normal distribution of LDL-C levels in African populations would be expected to be higher, the rare alleles may have some gain of function effect that associates them with lower LDL-C levels.

4.2.3. Polygenic Risk Score

A polygenic risk score (PRS) is used to determine the combined effect of multiple variants identified in an association study on a phenotype (Dudbridge, 2013). Variants are chosen for inclusion in the score based on significantly associated loci and knowledge of the risk allele. This score is used to predict the genetic risk for developing a disease and may use a simple additive or weighted model. In the latter, alleles at specific loci are weighted according to the extent of the effect on the phenotype (OR or beta value) and/or significant p value. A simple PRS was generated in this study based on the additive effect of alleles associated with a low LDL-C phenotype.

We included six SNPs with $p < 0.05$ (after correcting for multiple testing but before adjusting for potential confounders) from the allelic association in this PRS, rather than the two SNPs that were significantly associated ($p < 0.05$) after adjusting for covariates. This was done because a PRS can include markers with small effects and which show a trend toward the phenotype under investigation and because a PRS does not work well with a very small number of SNPs (Guo *et al.*, 2017; Paquette *et al.*, 2017).

Figure 3.7 A shows that the curve based on data from the controls shifted to the right in terms of a protective effect for LDL-C levels, compared to the PRS for the case group. The shift is modest, but significant ($p = 0.001$) which suggests that there is an association of the variants identified with LDL-C levels. This shift would be better

assessed with the inclusion of the larger sample size in the full AWI-Gen cohort, but unfortunately, we do not have data to test this.

A post hoc power analysis was done on the two significant SNPs. rs12071264 would be detected with 85.94% power (MAF 0.090), and rs6752026 (MAF 0.135) would be detected with 75.09% power. The SNPs are therefore well powered to detect an association in a case-control group of 500 (1000 individuals in total).

Figure 3.7 B shows clearly that with each addition of an allele associated with lower LDL-C, there is a trend to low LDL-C levels and the protective effect for LDL-C increases.

4.2.4. Visualisation of genotype association with LDL-C levels

Figure 3.6 A depicts a box and whisker plot for *LDLRAP1* rs12071264 (OR 0.5866). Genotypes with the minor allele (G) are associated with lower LDL-C levels and therefore in an individual who has the genotype AA, the LDL-C levels are higher, as expected. The minor allele of this SNP is associated with lower LDL-C levels in this African population. The SNP is an intronic variant and may affect transcription by affecting the affinity of transcription factors binding to the DNA, or it may cause alternate splicing amongst other consequences (Pagani and Baralle, 2004). This variant has not been detected in Asian and European populations according to KGP data, however, the frequency of the variant ranges from 0.07 to 0.20 in African American and African populations. Therefore, this variant may be associated with LDL-C levels only in African and African American populations.

The second SNP, *APOB* rs6752026 (OR 0.6898) also shows significant association of the minor allele (A) with low LDL-C levels. The homozygous AA and heterozygous AG are associated with low LDL-C levels (figure 3.6 B), and the homozygous GG genotype is associated with increased LDL-C levels. This SNP has a very low frequency in European populations but has frequencies of 0.05-0.2 in African and African American populations. This could also therefore also be an African-specific association.

Statistical tests showed a significant difference between the homozygous major allele genotype and the heterozygous genotype for both SNPs. This indicates that the presence of one minor allele reduces LDL-C levels, supporting the association of the minor allele with low LDL-C levels.

4.3. LDL-C Levels in African Populations

The LDL-C distribution in African populations is generally considered to be lower than in non-African populations, therefore, it is counter intuitive that the common alleles at the two associated SNPs would associate with higher LDL-C levels in Africans. As suggested above, the rare alleles could have a gain of function effect for low LDL-C, and/or gene-environment interactions could play a role, and low-fat diets and high physical activity could also contribute to lower LDL-C levels in African populations. However, as African populations become more urbanised, a more western lifestyle will follow, which could increase LDL-C, especially in those with a genetic predisposition for high LDL-C levels.

Detecting hyperlipidaemia early in individuals and administering treatment and lifestyle changes can reduce the number of CVD related events, and subsequently reduce the health burden on the African and South African health service (Kromberg, Sizer and Christianson, 2013). Precision public health is using data to implement intervention strategies that will most efficiently benefit the majority of individuals in a population (Dowell, Blazes and Desmond-Hellmann, 2016). Using population specific genetic variants to predict LDL-C levels will only be effective if they have good predictive potential and the assays are affordable. At present a serum cholesterol test remain a better and more cost-effective measure of LDL-C levels. Intervention strategies such as lifestyle changes and appropriate prescription of medication for high LDL-C that is effective for the population in question could be implemented for a better outcome.

4.4. Limitations of this study

The study has several strengths, but also some weaknesses. On the positive side, this study was a big African-based study in comparison to other African-based studies. It was designed to study groups with very high and very low LDL-C levels, thereby increasing the chances of finding genetic associations with LDL-C levels. The variants genotyped in this study were selected carefully for informativeness based on predicted function and MAF. Finally, a candidate gene approach testing genes previously shown to be associated with LDL-C was used, further enriching the potential for finding association signals with LDL-C levels.

The AWI-Gen participants were all African, however, they were multi-ethnic, with uneven distribution across ethnic groups and this could have caused bias due to

population sub-structure. False positives could have arisen due to the differences in allele frequencies across the three geographical regions. However, because lipid data on African populations are limited, this study serves as a starting point for subsequent research endeavours on understanding genetic associations with LDL-C levels in African populations. Due to this, it is likely that the two significant SNPs identified in logistic regression could be false positives as their frequencies differ across the three geographical regions. Replication studies are necessary to confirm or reject this finding.

A small number of SNPs were tested per gene and it would have been ideal to sequence the entire gene or have a more representative set of markers to capture all the haplotype blocks across each gene. Only a limited number of SNPs were genotyped due to budget constraints. There was also attrition of SNPs due to the elimination of 10 SNPs by the *in silico* MassARRAY assay validation platform due to the system predicting that primers would not bind effectively to the DNA sequence, preventing amplification of the region of interest, and therefore genotyping would fail. In the end, only 19 SNPs were genotyped in this study.

African populations have lower linkage disequilibrium (LD) among loci in comparison to European and Asian populations (Teo, Small and Kwiatkowski, 2010). Due to the low LD in African populations, more SNPs are needed to capture all the haplotype blocks, however, when applying fine mapping (a high density of SNPs in the region of interest), if an association were found, it would make it easier to find the causal variant because it is likely to be on a smaller haplotype block. In other populations, LD blocks are larger, therefore not many SNPs are required to be genotyped to cover all LD blocks. With the small selection of SNPs genotyped in this study, we reduce the chance of finding associations, and therefore also the chance of finding a causal variant in LD with the selected SNPs. It is, therefore, imperative that more SNPs are selected when carrying out a GWAS on African populations to cover the many LD blocks.

The sample size was relatively small. However according to Quanto, this study was powered enough to detect association with the variants in this cohort. The post hoc analysis on power resulted in 85.94% and 75.09% power to detect association with rs12071264 and rs6752026. To fully assess the potential predictive value of the polygenic risk score would require a much larger study. An increase in sample size

would also increase the detection of a number of significant SNPs with smaller effects that could be found using logistic regression analysis.

Most studies reporting on LDL-C in the literature are based on non-African participants, therefore we need to assess whether the effects of the variants are similar in African populations or whether different variants (sometimes in the same genes) may be involved. African populations are largely under studied and may have different genetic aetiologies with respect to variation in LDL-C levels. In addition, many studies have examined gene mutation involvement in monogenic forms of high LDL-C related disorders (e.g. FH) and have therefore failed to detect low penetrance alleles associated with LDL-C levels as a multifactorial trait. The GWAS catalogue, which is better suited to report on complex trait associations, as opposed to ClinVar, did not have information on associations with the two associated SNPs this study identified. In part this could be because they are absent or extremely rare in non-African populations.

Despite knowing that HIV status can affect LDL-C levels, HIV data could not be used in logistic regression due to many missing data points. HIV status should have been used as a criterion for exclusion if enough data had been available. In areas where HIV is common, the genetic association of LDL-C in the context of HIV should be studied specifically. This should be done with a good understanding of the effects of ARVs on lipid levels, which is a complex focus area.

Diet and physical exercise were not used as confounders in logistic regression as the data for the AWI-Gen participants was not available when this study was started.

As this study employed a candidate gene approach, we chose genes previously identified in the literature – so we chose genes known to be involved in LDL-C level variation in European populations, but for which African studies are lacking. Miller *et al.*, (2016) investigated seven genes in a South African Xhosa woman to find associations with low LDL-C levels. In the study *APOB*, *MTTP*, *PCSK9*, *ANGPTL3*, *SAR1B* and *APOC3* were sequenced and *APOE* was genotyped. They found two novel variants in *APOB* that showed no association with LDL-C. However, they did find a homozygous *APOE* $\epsilon 2$ association with low LDL-C levels. This indicates that in African populations, there may be other genes that contribute to LDL-C levels.

4.5. Future research to understand the genetic contribution to LDL-C levels among Africans

To more thoroughly evaluate the contribution of genetic variation from the four genes of interest, a comprehensive set of LD pruned SNPs for each gene could have been selected to ensure that all LD blocks were covered. If this were done, haplotype analysis could determine more accurately what the contribution of genetic variation from each gene is relative to the phenotype. Furthermore, to capture all genetic variation fully, it would be ideal to sequence the four genes explored in this study, along with other genes implicated with LDL-C as a complex trait, in African populations. Deep whole genome sequencing on African populations with attached relevant phenotype data would therefore be useful. Association analysis could identify genes implicated with LDL-C levels that may not show the same associations in other populations.

If the sample size is large enough, associations could be found with rare variants of low or modest effect on the phenotype that would give additional insight into the genetic landscape of African populations concerning the prediction of LDL-C levels.

In addition to genetic susceptibility, gene-gene and gene-environment effects contribute greatly to LDL-C levels, as only 10% of variability can be attributed to common variants. Genetic variation, together with DNA methylation epigenetic effects, could explain the phenotype variation more extensively (Soto-Ramírez *et al.*, 2013). Gain of function effects for the rarer alleles could help to explain why, despite the fact that the common alleles for the two associated SNPs from this study are associated with high LDL-C levels, the LDL-C levels in African populations tend to be generally low.

Extensive research on the functional impact of the SNPs highlighted in this study as well as other studies, particularly variants in *PCSK9*, could help understand the underlying mechanisms of the variants and the reason for the effects they have on LDL-C levels in African populations. This would aid in clarifying whether variants contributing to LDL-C levels in African populations are different to variants in non-African populations.

From our limited study, it seems likely that there are African-specific gene variants that affect LDL-C levels in African populations and that additional African-specific genes and variants are involved (Miller *et al.*, 2016). To fully assess this possibility, a

different study approach is necessary, and could include a full GWAS analysis with a SNP array enriched for common African genetic variation, whole exome sequencing or whole genome sequencing in a large cohort of African participants with the relevant phenotype and behavioural data.

5. Conclusion

Few studies have been done on genetic associations with LDL-C levels in black African populations. This study, albeit small, has shown that variants in four common LDL-C associated genes showed allelic association with LDL-C levels in 993 African participants from the AWI-Gen study. After logistic regression and adjusting for potential covariates, two significantly associated alleles (*LDLRAP1* rs12071264G and *APOB* rs67520260A) were found to associate with low LDL-C levels in East, West and South African populations, with suspected gain of function effects. The average LDL-C levels for individuals with different genotypes at these loci showed the expected correlation with LDL-C levels. One SNP was a missense SNP (rs6752026A) in *APOB* and the other was an intronic variant (rs12071264G) in *LDLRAP1*.

The *LDLRAP1* intronic variant is close to a splice site and if not directly be involved in alternate splicing, the SNP could possibly affect binding of the splicing machinery, thereby affecting the way or speed with which the transcript is spliced. This variant is absent in European populations. The effects of the *APOB* missense SNP remain unknown but SIFT and PolyPhen2 predicted it to be likely deleterious suggesting that it could alter protein functionality. This SNP specifically, could have a functional impact in LDL-C metabolism through its action on the APOB protein. It is observed in very low frequencies in European populations (0.1%). The absence and low frequency of the two SNPs, respectively in European populations, suggest that they may have an African-specific effect on LDL-C. The minor alleles of both variants are associated with low LDL-C levels, suggesting that the variants have a GoF effect.

The polygenic risk score using six LDL-C associated SNPs showed that with an increase in the number of alleles associated with high LDL-C levels, there was a significant difference in LDL-C levels between individuals in the high and low LDL-C groups. In addition, individuals with a higher risk score had, on average, higher LDL-C levels. Box and whisker plots also showed that with the presence of the minor allele (heterozygous and homozygous) in both variants, the LDL-C levels were lower.

Due to the population substructure, one might expect significant differences across African populations, but we observed that the minor allele frequencies for most variants were not significantly different across East, West and South African populations, with some exceptions. However, the allele frequencies for rs6752026

and rs12071264 differed significantly between African and European populations, with rs6752026 not observed in European and Asian populations. An improved future study design would be to consider sampling taking into consideration a population specific selection approach.

Genetic association studies rely on linkage disequilibrium and therefore associated variants are often not causal, but in LD with the LDL-C trait. Therefore, the variants that have been identified are not necessarily the functional variant itself, but may be in LD with the causal/functional variant. This is highly likely for rs12071264, as it was selected for high MAF rather than functional impact. This could explain why an intronic SNP, with no obvious functional impact (although it may affect splicing), was significantly associated with LDL-C levels in this study.

Africa is generally a genetic data-scarce region, therefore, extensive studies on the genetic aetiology of LDL-C level variation in African populations are few. The genes chosen for this study were based on data from previous associations in European populations. Whereas some SNPs may replicate in African populations it is important to use exploratory approaches that have the potential to identify novel African genetic contributions to the variation in lipid levels. A future exploratory approach should include use of the H3Africa SNP genotyping array on the full AWI-Gen study of over 10,000 individuals. This work is currently underway and could yield novel associations (possibly African-specific) that contribute to complex traits such as LDL-C levels, generating a wealth of information in data-scarce region.

6. References

- Abifadel, M., Varret, M., Rabès, J.-P., *et al.* (2003) 'Mutations in PCSK9 cause autosomal dominant hypercholesterolemia.', *Nature genetics*, 34(2), pp. 154–156. doi: 10.1038/ng1161.
- Adzhubei, I. A., Schmidt, S., Peshkin, L., *et al.* (2010) 'A method and server for predicting damaging missense mutations', *Nature Methods*, 7(4), pp. 248–249. doi: 10.1038/nmeth0410-248.
- Agyepong, I. A., Sewankambo, N., Binagwaho, A., *et al.* (2017) 'The path to longer and healthier lives for all Africans by 2030: the Lancet Commission on the future of health in sub-Saharan Africa', *The Lancet*, 17, pp. 1–57. doi: 10.1016/S0140-6736(17)31509-X.
- Akiyamen, L. E., Genest, J., Shan, S. D., *et al.* (2017) 'Estimating the prevalence of heterozygous familial hypercholesterolaemia : a systematic review and meta-analysis', *BMJ*, 7, pp. 1–14. doi: 10.1136/bmjopen-2017-016461.
- Anastos, K., Lu, D., Shi, Q., *et al.* (2007) 'Association of Serum Lipid Levels With HIV Serostatus, Specific Antiretroviral Agents, and Treatment Regimens', *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 45(1), pp. 34–42. doi: 10.1097/QAI.0b013e318042d5fe.
- Anderson, C. A., Pettersson, F. H., Clarke, G. M., *et al.* (2010) 'Data quality control in genetic case-control association studies.', *Nature protocols*, 5(9), pp. 1564–73. doi: 10.1038/nprot.2010.116.
- Antonarakis, S. E. and Beckmann, J. S. (2006) 'Mendelian disorders deserve more attention', *Nature Reviews Genetics*, 7(4), pp. 277–282. doi: 10.1038/nrg1826.
- Arnett, D. K. and Shah, S. J. (2014) *Cardiovascular Genetics and Genomics in Clinical Practice*. USA: Demos Medical Publishing. Available at: <https://books.google.co.za/books?hl=en&lr=&id=z34eBQAAQBAJ&oi=fnd&pg=PA12&dq=genetics+of+complex+traits&ots=kc5iWpLfJH&sig=3DYv1gNs06-TcMn98jHXj4ba-U#v=onepage&q=genetics+of+complex+traits&f=false> (Accessed: 25 September 2017).
- Austin, M. A., Hutter, C. M., Zimmern, R. L., *et al.* (2004) 'Genetic Causes of Monogenic Heterozygous Familial Hypercholesterolemia : A HuGE Prevalence Review', *American journal of epidemiology*, 160(5), pp. 407–420. doi: 10.1093/aje/kwh236.
- Barnett, I. J., Lee, S. and Lin, X. (2013) 'Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies', *Genetic Epidemiology*, 37(2), pp. 142–151. doi: 10.1002/gepi.21699.
- Barr, A. L., Young, E. H., Smeeth, L., *et al.* (2016) 'The need for an integrated approach for chronic disease research and care in Africa', *Global Health, Epidemiology and Genomics*, 1(19), p. doi.org/10.1017/gh.2016.16. doi: 10.1017/gh.2016.16.
- Bastien, M., Poirier, P., Lemieux, I., *et al.* (2014) 'Overview of Epidemiology and Contribution of Obesity to Cardiovascular Disease', *Progress in Cardiovascular Diseases*, 56(4), pp. 369–381. doi: 10.1016/j.pcad.2013.10.016.

- Beekman, M., Heijmans, B. T., Martin, N. G., *et al.* (2002) 'Heritabilities of Apolipoprotein and Lipid Levels in Three Countries', *Twin Research*, 5(02), pp. 87–97. doi: 10.1375/twin.5.2.87.
- Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of Royal Statistical Society. Series B (Methodological)*, 57(1), pp. 289–300.
- Benn, M., Nordestgaard, B. G., Grande, P., *et al.* (2010) 'PCSK9R46L, Low-Density Lipoprotein Cholesterol Levels, and Risk of Ischemic Heart Disease', *Journal of the American College of Cardiology*, 55(25), pp. 2833–2842. doi: 10.1016/j.jacc.2010.02.044.
- Benn, M., Watts, G. F., Tybjaerg-Hansen, A., *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, pp. 1384–1394. doi: 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. doi: 10.1586/eem.12.55.
- Bonora, E. and Muggeo, M. (2001) 'Postprandial blood glucose as a risk factor for cardiovascular disease in Type II diabetes: the epidemiological evidence', *Diabetologia*, 44(12), pp. 2107–2114. doi: 10.1007/s001250100020.
- Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) 'An Expanded View of Complex Traits: From Polygenic to Omnigenic', *Cell*, 169(7), pp. 1177–1186. doi: 10.1016/j.cell.2017.05.038.
- Brown, M. S., Faust, J. R. and Goldstein, J. L. (1975) 'Role of the low density lipoprotein receptor in regulating the content of free and esterified cholesterol in human fibroblasts', *Journal of Clinical Investigation*, 55, pp. 783–793. doi: 10.1172/JCI107989.
- Brown, M. S. and Goldstein, J. L. (1974) 'Familial hypercholesterolemia: defective binding of lipoproteins to cultured fibroblasts associated with impaired regulation of 3-hydroxy-3-methylglutaryl coenzyme A reductase activity.', *Proceedings of the National Academy of Sciences of the United States of America*, 71(3), pp. 788–792. doi: 10.1073/pnas.71.3.788.
- Brown, M. S. and Goldstein, J. L. (1986) 'A Receptor-Mediated Pathway for Cholesterol Homeostasis', *Science*, 232(4746), pp. 34–47. doi: 10.1126/science.3513311.
- Burlingame, B. (2003) 'The food of Near East, North West and Western African regions.', *Asia Pacific journal of clinical nutrition*, 12(3), pp. 309–12.
- Catapano, A. L., Graham, I., De Backer, G., *et al.* (2016) '2016 ESC/EAS Guidelines for the Management of Dyslipidaemias', *European Heart Journal*, 37(39), pp. 2999–3058. doi: 10.1093/eurheartj/ehw272.
- Chang, C. C., Chow, C. C., Tellier, L. C., *et al.* (2015) 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *GigaScience*, 4(1), p. 7. doi: 10.1186/s13742-015-0047-8.

Charles Zaiontz (2018) Real Statistics Resource Pack software (release 5.4), Copyright (2013-2018). Available at: <http://www.real-statistics.com> (Accessed: 9 June 2018).

Clarke, G. M., Anderson, C. A., Pettersson, F. H., *et al.* (2011) 'Basic statistical analysis in genetic case-control studies.', *Nature protocols*, 6(2), pp. 121–33. doi: 10.1038/nprot.2010.182.

Cohen, J., Pertsemlidis, A., Kotowski, I. K., *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. doi: 10.1038/ng1509.

Cohen, J. C., Boerwinkle, E., Mosley, T. H., *et al.* (2006) 'Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease', *New England Journal of Medicine*, 354(12), pp. 1264–1272. doi: 10.1056/NEJMoa054013.

Cuchel, M., Bruckert, E., Ginsberg, H. N., *et al.* (2014) 'Homozygous familial hypercholesterolaemia: new insights and guidance for clinicians to improve detection and clinical management. A position paper from the Consensus Panel on Familial Hypercholesterolaemia of the European Atherosclerosis Society.', *European heart journal*, 35(32), pp. 2146–57. doi: 10.1093/eurheartj/ehu274.

De, R., Verma, S. S., Holzinger, E., *et al.* (2017) 'Identifying gene–gene interactions that are highly associated with four quantitative lipid traits across multiple cohorts', *Human Genetics*, 136(2), pp. 165–178. doi: 10.1007/s00439-016-1738-7.

Deeb, S. S., Disteché, C., Motulsky, A. G., *et al.* (1986) 'Chromosomal localization of the human apolipoprotein B gene and detection of homologous RNA in monkey intestine.', *Proceedings of the National Academy of Sciences of the United States of America*, 83(2), pp. 419–22.

Dogan, R. I., Getoor, L., Wilbur, W. J., *et al.* (2007) 'SplicePort--an interactive splice-site analysis tool.', *Nucleic acids research*, 35, pp. W285-91. doi: 10.1093/nar/gkm407.

Dowell, S. F., Blazes, D. and Desmond-Hellmann, S. (2016) 'Four steps to precision public health', *Nature*, 540(7632), pp. 189–191. doi: 10.1038/540189a.

Dudbridge, F. (2013) 'Power and Predictive Accuracy of Polygenic Risk Scores', *PLoS Genetics*, 9(3), p. e1003348. doi: 10.1371/journal.pgen.1003348.

Durbin, R. M., Altshuler, D. L., Durbin, R. M., *et al.* (2010) 'A map of human genome variation from population-scale sequencing', *Nature*, 467(7319), pp. 1061–1073. doi: 10.1038/nature09534.

Durrington, P. (2003) 'Dyslipidaemia', *The Lancet*, 362(9385), pp. 717–731. doi: 10.1016/S0140-6736(03)14234-1.

Eden, E. R., Patel, D. D., Sun, X.-M., *et al.* (2002) 'Restoration of LDL receptor function in cells from patients with autosomal recessive hypercholesterolemia by retroviral expression of ARH1.', *The Journal of clinical investigation*, 110(11), pp. 1695–702. doi: 10.1172/JCI16445.

Egusa, G., Murakami, F., Ito, C., *et al.* (1993) 'Westernized food habits and concentrations of serum lipids in the Japanese', *Atherosclerosis*, 100(2), pp. 249–255. doi: 10.1016/0021-9150(93)90211-C.

Elbitar, S., Susan-Resiga, D., Ghaleb, Y., *et al.* (2018) 'New Sequencing

technologies help revealing unexpected mutations in Autosomal Dominant Hypercholesterolemia', *Scientific Reports*, 8(1), p. 1943. doi: 10.1038/s41598-018-20281-9.

Emond, M. J., Louie, T., Emerson, J., *et al.* (2012) 'Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis', *Nature Genetics*, 44(8), pp. 886–889. doi: 10.1038/ng.2344.

Farnier, M., Civeira, F., Descamps, O., *et al.* (2017) 'How to implement clinical guidelines to optimise familial hypercholesterolaemia diagnosis and treatment', *Atherosclerosis Supplements*, 26, pp. 25–35. doi: 10.1016/S1567-5688(17)30022-3.

Ference, B. A., Yoo, W., Alesh, I., *et al.* (2012) 'Effect of Long-Term Exposure to Lower Low-Density Lipoprotein Cholesterol Beginning Early in Life on the Risk of Coronary Heart Disease', *Journal of the American College of Cardiology*, 60(25), pp. 2631–2639. doi: 10.1016/j.jacc.2012.09.017.

Fokkema, I. F. A. C., den Dunnen, J. T. and Taschner, P. E. M. (2005) 'LOVD: Easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach', *Human Mutation*, 26(2), pp. 63–68. doi: 10.1002/humu.20201.

García-Giustiniani, D. and Stein, R. (2016) 'Genetics of Dyslipidemia.', *Arquivos brasileiros de cardiologia*, 106(5), pp. 434–438. doi: 10.5935/abc.20160074.

Gauderman, W. J. (2002) 'Sample size requirements for matched case-control studies of gene-environment interaction.', *Statistics in medicine*, 21(1), pp. 35–50.

Genest, J. (2017) 'Familial Hypercholesterolemia: Awareness, Appraisal, and Action', *Canadian Journal of Cardiology*, 33(3), pp. 298–299. doi: 10.1016/j.cjca.2016.09.011.

Guo, W., Samuels, J. F., Wang, Y., *et al.* (2017) 'Polygenic risk score and heritability estimates reveals a genetic relationship between ASD and OCD', *European Neuropsychopharmacology*, 27(7), pp. 657–666. doi: 10.1016/J.EURONEURO.2017.03.011.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., *et al.* (2015) 'The African Genome Variation Project shapes medical genetics in Africa', *Nature*, 517(7534), pp. 327–332. doi: 10.1038/nature13997.

Gurwitz, D. and McLeod, H. L. (2013) 'Genome-wide studies in pharmacogenomics: harnessing the power of extreme phenotypes.', *Pharmacogenomics*, 14(4), pp. 337–9. doi: 10.2217/pgs.13.35.

Hagberg, J. M., Wilund, K. R. and Ferrell, R. E. (2000) 'APO E gene and gene-environment effects on plasma lipoprotein-lipid levels', *Physiological Genomics*, 4(2), pp. 101–108. doi: 10.1152/physiolgenomics.2000.4.2.101.

Hallman, D. M., Srinivasan, S. R., Chen, W., *et al.* (2007) 'Relation of PCSK9 Mutations to Serum Low-Density Lipoprotein Cholesterol in Childhood and Adulthood (from the Bogalusa Heart Study)', *The American Journal of Cardiology*, 100(1), pp. 69–72. doi: 10.1016/J.AMJCARD.2007.02.057.

Hegele, R. A. (2009) 'Plasma lipoproteins: genetic influences and clinical implications', *Nature Reviews Genetics*, 10(2), pp. 109–121. doi: 10.1038/nrg2481.

Hooper, A. J., Marais, A. D., Tanyanyiwa, D. M., *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. doi: 10.1016/j.atherosclerosis.2006.08.039.

Hooper, A. J., van Bockxmeer, F. M. and Burnett, J. R. (2005) 'Monogenic hypocholesterolaemic lipid disorders and apolipoprotein B metabolism.', *Critical reviews in clinical laboratory sciences*, 42(5–6), pp. 515–545. doi: 10.1080/10408360500295113.

Huaman, M. A., Henson, D., Ticona, E., *et al.* (2015) 'Tuberculosis and Cardiovascular Disease: Linking the Epidemics.', *Tropical diseases, travel medicine and vaccines*, 1, pp. 1–15. doi: 10.1186/s40794-015-0014-5.

Ibe, U. K., Whittall, R., Humphries, S. E., *et al.* (2017) 'Analysis of mutations causing familial hypercholesterolaemia in black South African patients of different ancestry', *South African Medical Journal*, 107(2), pp. 145–148. doi: 10.7196/SAMJ.2017.v107i2.12022.

Innerarity, T. L., Weisgraber, K. H., Arnold, K. S., *et al.* (1987) 'Familial defective apolipoprotein B-100: low density lipoproteins with abnormal receptor binding.', *Proceedings of the National Academy of Sciences of the United States of America*, 84, pp. 6919–6923. doi: 10.1073/pnas.84.19.6919.

Innerarity, T. L., Mahley, R. W., Weisgraber, K. H., *et al.* (1990) 'Familial defective apolipoprotein B-100: a mutation of apolipoprotein B that causes hypercholesterolemia.', *Journal of lipid research*, 31, pp. 1337–1349.

Jenkins, T., Nicholls, E., Gordon, E., *et al.* (1980) 'Familial hypercholesterolaemia--a common genetic disorder in the Afrikaans population.', *South African Medical Journal*, 57(943), pp. 943–947.

Jialal, I. and Barton Duell, P. (2016) 'Diagnosis of Familial Hypercholesterolemia', *American Journal of Clinical Pathology*, 145(4), pp. 437–439. doi: 10.1093/ajcp/aqw001.

Kathiresan, S., Willer, C. J., Peloso, G. M., *et al.* (2009) 'Common variants at 30 loci contribute to polygenic dyslipidemia.', *Nature genetics*, 41(1), pp. 56–65. doi: 10.1038/ng.291.

Keates, A. K., Mocumbi, A. O., Ntsekhe, M., *et al.* (2017) 'Cardiovascular disease in Africa: epidemiological profile and challenges', *Nature Reviews Cardiology*, 14(5), pp. 273–293. doi: 10.1038/nrcardio.2017.19.

Kim, D. S., Burt, A. A., Ranchalis, J. E., *et al.* (2013) 'Novel gene-by-environment interactions: APOB and NPC1L1 variants affect the relationship between dietary and total plasma cholesterol.', *Journal of lipid research*, 54(5), pp. 1512–20. doi: 10.1194/jlr.P035238.

Kircher, M., Witten, D. M., Jain, P., *et al.* (2014) 'A general framework for estimating the relative pathogenicity of human genetic variants', *Nature Genetics*, 46(3), pp. 310–317. doi: 10.1038/ng.2892.

Klug, E., Raal, F., Marais, A., *et al.* (2015) 'South African Dyslipidaemia Guideline Consensus Statement.pdf', *South African Family Practice*, 57(2), pp. 22–31.

Koethe, J. R., Jenkins, C. A., Petucci, C., *et al.* (2016) 'Superior Glucose Tolerance and Metabolomic Profiles, Independent of Adiposity, in HIV-Infected Women Compared With Men on Antiretroviral Therapy.', *Medicine*, 95(19), pp. 1–8. doi: 10.1097/MD.00000000000003634.

- Kotze, M. J., Warnich, L., Langenhoven, E., *et al.* (1990) 'An exon 4 mutation identified in the majority of South African familial hypercholesterolaemics.', *Journal of medical genetics*, 27(5), pp. 298–302. doi: 10.1136/jmg.27.5.298.
- Kromberg, J. G., Sizer, E. B. and Christianson, A. L. (2013) 'Genetic services and testing in South Africa', *Journal of Community Genetics*, 4(3), pp. 413–423. doi: 10.1007/s12687-012-0101-5.
- Kulanuwat, S., Tungtrongchitr, R., Billington, D., *et al.* (2015) 'Prevalence of plasma small dense LDL is increased in obesity in a Thai population.', *Lipids in health and disease*, 14, p. 30. doi: 10.1186/s12944-015-0034-1.
- Kumar, P., Henikoff, S. and Ng, P. C. (2009) 'Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm', *Nature Protocols*, 4(7), pp. 1073–1082. doi: 10.1038/nprot.2009.86.
- Lange, L. A., Hu, Y., Zhang, H., *et al.* (2014) 'Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol.', *American journal of human genetics*, 94(2), pp. 233–245. doi: 10.1016/j.ajhg.2014.01.010.
- Langsted, A., Nordestgaard, B. G., Benn, M., *et al.* (2016) 'PCSK9 R46L Loss-of-Function Mutation Reduces Lipoprotein(a), LDL Cholesterol, and Risk of Aortic Valve Stenosis', *The Journal of Clinical Endocrinology & Metabolism*, 101(9), pp. 3281–3287. doi: 10.1210/jc.2016-1206.
- Lee, C. J., Lee, Y., Park, S., *et al.* (2017) 'Rare and common variants of APOB and PCSK9 in Korean patients with extremely low low-density lipoprotein-cholesterol levels', *PLOS ONE*, 12(10), p. e0186446. doi: 10.1371/journal.pone.0186446.
- Leitersdorf, E., Van Der Westhuyzen, D. R., Coetzee, G. A., *et al.* (1989) 'Two common low density lipoprotein receptor gene mutations cause familial hypercholesterolemia in Afrikaners', *Journal of Clinical Investigation*, 84, pp. 954–961. doi: 10.1172/JCI114258.
- Li, D., Lewinger, J. P., Gauderman, W. J., *et al.* (2011) 'Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies', *Genetic Epidemiology*, 35(8), pp. 790–799. doi: 10.1002/gepi.20628.
- Magkos, F., Mohammed, B. S. and Mittendorfer, B. (2008) 'Effect of obesity on the plasma lipoprotein subclass profile in normoglycemic and normolipidemic men and women', *International Journal of Obesity*, 32(11), pp. 1655–1664. doi: 10.1038/ijo.2008.164.
- Marais, A. D., Kim, J. B., Wasserman, S. M., *et al.* (2015) 'PCSK9 inhibition in LDL cholesterol reduction: Genetics and therapeutic implications of very low plasma lipoprotein levels', *Pharmacology & Therapeutics*, 145, pp. 58–66. doi: 10.1016/j.pharmthera.2014.07.004.
- Martín-Morales, R., García-Díaz, J. D., Tarugi, P., *et al.* (2013) 'Familial hypobetalipoproteinemia: Analysis of three Spanish cases with two new mutations in the APOB gene', *Gene*, 531(1), pp. 92–96. doi: 10.1016/j.gene.2013.08.049.
- 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. doi: 10.1073/pnas.0402133101.

- Mayosi, B. M., Flisher, A. J., Lalloo, U. G., *et al.* (2009) 'The burden of non-communicable diseases in South Africa', *The Lancet*, 374(9693), pp. 934–947. doi: 10.1016/S0140-6736(09)61087-4.
- McLaren, W., Pritchard, B., Rios, D., *et al.* (2010) 'Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor', *Bioinformatics*, 26(16), pp. 2069–2070. doi: 10.1093/bioinformatics/btq330.
- McNutt, M. C., Lagace, T. A. and Horton, J. D. (2007) 'Catalytic Activity Is Not Required for Secreted PCSK9 to Reduce Low Density Lipoprotein Receptors in HepG2 Cells', *Journal of Biological Chemistry*, 282(29), pp. 20799–20803. doi: 10.1074/jbc.C700095200.
- Meiner, V., Landsberger, D., Berkman, N., *et al.* (1991) 'A common Lithuanian mutation causing familial hypercholesterolemia in Ashkenazi Jews.', *American Journal of Human Genetics*, 49, pp. 443–449.
- Michaely, P., Li, W.-P., Anderson, R. G. W., *et al.* (2004) 'The modular adaptor protein ARH is required for low density lipoprotein (LDL) binding and internalization but not for LDL receptor clustering in coated pits.', *The Journal of Biological Chemistry*, 279(32), pp. 34023–34031. doi: 10.1074/jbc.M405242200.
- Miller, S. A., Hooper, A. J., Mantiri, G. A., *et al.* (2016) 'Novel APOB missense variants, A224T and V925L, in a black South African woman with marked hypocholesterolemia', *Journal of Clinical Lipidology*, 10(3), pp. 604–609. doi: 10.1016/J.JACL.2016.01.006.
- Miller, S. A., Dykes, D. D. and Polesky, H. F. (1988) 'A simple salting out procedure for extracting DNA from human nucleated cells', *Nucleic Acids Research*, 16(3), p. 1215. doi: 10.1093/nar/16.3.1215.
- NCBI (2017) *NCBI*. Available at: <https://www.ncbi.nlm.nih.gov/> (Accessed: 25 June 2017).
- Nordestgaard, B. G., Chapman, M. J., Humphries, S. E., *et al.* (2013) 'Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: Guidance for clinicians to prevent coronary heart disease', *European Heart Journal*, 34(45), pp. 3478–3490. doi: 10.1093/eurheartj/ehd273.
- Pagani, F. and Baralle, F. E. (2004) 'Opinion: Genomic variants in exons and introns: identifying the splicing spoilers', *Nature Reviews Genetics*, 5(5), pp. 389–396. doi: 10.1038/nrg1327.
- Palmer, C. S., Anzinger, J. J., Zhou, J., *et al.* (2014) 'Glucose Transporter 1–Expressing Proinflammatory Monocytes Are Elevated in Combination Antiretroviral Therapy–Treated and Untreated HIV+ Subjects', *The Journal of Immunology*, 193(11), pp. 5595–5603.
- Paquette, M., Chong, M., Thériault, S., *et al.* (2017) 'Polygenic risk score predicts prevalence of cardiovascular disease in patients with familial hypercholesterolemia', *Journal of Clinical Lipidology*, 11(3), p. 725–732.e5. doi: 10.1016/j.jacl.2017.03.019.
- Pickrell, J. K., Patterson, N., Barbieri, C., *et al.* (2012) 'The genetic prehistory of southern Africa.', *Nature communications*, 3(1143), pp. 1–6. doi: 10.1038/ncomms2140.
- Pirillo, A., Garlaschelli, K., Arca, M., *et al.* (2017) 'Spectrum of mutations in Italian

patients with familial hypercholesterolemia: New results from the LIPIGEN study', *Atherosclerosis Supplements*, 29, pp. 17–24. doi: 10.1016/j.atherosclerosissup.2017.07.002.

Purcell, S., Neale, B., Todd-Brown, K., *et al.* (2007) 'PLINK: a tool set for whole-genome association and population-based linkage analyses.', *American journal of human genetics*, 81(3), pp. 559–75. doi: 10.1086/519795.

R Core Team (v 3.4) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.r-project.org>.

Raal, F. J., Honarpour, N., Blom, D. J., *et al.* (2015) 'Inhibition of PCSK9 with evolocumab in homozygous familial hypercholesterolaemia (TESLA Part B): A randomised, double-blind, placebo-controlled trial', *The Lancet*, 385(9965), pp. 341–350. doi: 10.1016/S0140-6736(14)61374-X.

Ramsay, M., Crowther, N., Tambo, E., *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(e20), pp. 1–13. doi: 10.1017/gheg.2016.17.

Reeder, B. A., Senthilselvan, A., Després, J. P., *et al.* (1997) 'The association of cardiovascular disease risk factors with abdominal obesity in Canada. Canadian Heart Health Surveys Research Group.', *Canadian Medical Association Journal*, 157(1), pp. S39-45.

Rimbert, A., Pichelin, M., Lecointe, S., *et al.* (2016) 'Identification of novel APOB mutations by targeted next-generation sequencing for the molecular diagnosis of familial hypobetalipoproteinemia', *Atherosclerosis*, 250, pp. 52–56. doi: 10.1016/j.atherosclerosis.2016.04.010.

Robinson, J. G., Farnier, M., Krempf, M., *et al.* (2015) 'Efficacy and Safety of Alirocumab in Reducing Lipids and Cardiovascular Events', *New England Journal of Medicine*, 372(16), pp. 1489–1499. doi: 10.1056/NEJMoa1501031.

Rubinsztein, D. C., Coetzee, G. a, Marais, a D., *et al.* (1992) 'Identification and properties of the proline664-leucine mutant LDL receptor in South Africans of Indian origin.', *Journal of lipid research*, 33, pp. 1647–1655.

Saavedra, Y. G. L., Dufour, R., Davignon, J., *et al.* (2014) 'PCSK9 R46L, lower LDL, and cardiovascular disease risk in familial hypercholesterolemia: a cross-sectional cohort study.', *Arteriosclerosis, thrombosis, and vascular biology*, 34(12), pp. 2700–2705. doi: 10.1161/ATVBAHA.114.304406.

Sánchez-Hernández, R. M., Prieto-Matos, P., Civeira, F., *et al.* (2018) 'Autosomal recessive hypercholesterolemia in Spain', *Atherosclerosis*, 269, pp. 1–5. doi: 10.1016/j.atherosclerosis.2017.12.006.

Sandler, N. G., Zhang, X., Bosch, R. J., *et al.* (2014) 'Sevelamer Does Not Decrease Lipopolysaccharide or Soluble CD14 Levels But Decreases Soluble Tissue Factor, Low-Density Lipoprotein (LDL) Cholesterol, and Oxidized LDL Cholesterol Levels in Individuals With Untreated HIV Infection', *Journal of Infectious Diseases*, 210(10), pp. 1549–1554. doi: 10.1093/infdis/jiu305.

Sauna, Z. E. and Kimchi-Sarfaty, C. (2011) 'Understanding the contribution of synonymous mutations to human disease', *Nature Reviews Genetics*, 12(10), pp.

683–691. doi: 10.1038/nrg3051.

Schonfeld, G., Lin, X. and Yue, P. (2005) 'Familial hypobetalipoproteinemia: genetics and metabolism', *Cellular and Molecular Life Sciences*, 62(12), pp. 1372–1378. doi: 10.1007/s00018-005-4473-0.

Seftel, H. C., Baker, S. G., Jenkins, T., *et al.* (1989) 'Prevalence of familial hypercholesterolemia in Johannesburg Jews', *American Journal of Medical Genetics*, 34, pp. 545–547. doi: 10.1002/ajmg.1320340418.

Seidman, J. G. and Seidman, C. (2002) 'Transcription factor haploinsufficiency: When half a loaf is not enough', *Journal of Clinical Investigation*, 109(4), pp. 451–455. doi: 10.1172/JCI200215043.

Shen, H., Damcott, C. M., Rampersaud, E., *et al.* (2010) 'Familial defective apolipoprotein B-100 and increased low-density lipoprotein cholesterol and coronary artery calcification in the old order amish.', *Archives of internal medicine*, 170(20), pp. 1850–1855.

Sidney, S., Quesenberry, C. P., Jaffe, M. G., *et al.* (2016) 'Recent Trends in Cardiovascular Mortality in the United States and Public Health Goals', *JAMA Cardiology*, 1(5), p. 594. doi: 10.1001/jamacardio.2016.1326.

Sims, D., Sudbery, I., Illott, N. E., *et al.* (2014) 'Sequencing depth and coverage: key considerations in genomic analyses', *Nature Reviews Genetics*, 15(2), pp. 121–132. doi: 10.1038/nrg3642.

Snieder, H., van Doornen, L. J. and Boomsma, D. I. (1999) 'Dissecting the genetic architecture of lipids, lipoproteins, and apolipoproteins: lessons from twin studies.', *Arteriosclerosis, thrombosis, and vascular biology*, 19(12), pp. 2826–2834. doi: 10.1161/01.ATV.19.12.2826.

Soko, N. D., Masimirembwa, C. and Dandara, C. (2016) 'Pharmacogenomics of Rosuvastatin: A Glocal (Global+Local) African Perspective and Expert Review on a Statin Drug', *OMICS: A Journal of Integrative Biology*, 20(9), pp. 498–509. doi: 10.1089/omi.2016.0114.

Soto-Ramírez, N., Arshad, S. H., Holloway, J. W., *et al.* (2013) 'The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years', *Clinical Epigenetics*, 5(1), pp. 1–8. doi: 10.1186/1868-7083-5-1.

Soutar, A. K., Knight, B. L. and Patel, D. D. (1989) 'Identification of a point mutation in growth factor repeat C of the low density lipoprotein-receptor gene in a patient with homozygous familial hypercholesterolemia that affects ligand binding and intracellular movement of receptors.', *Proceedings of the National Academy of Sciences of the United States of America*, 86, pp. 4166–4170. doi: 10.1073/pnas.86.11.4166.

Soutar, A. K. and Naoumova, R. P. (2004) 'Autosomal Recessive Hypercholesterolemia', *Seminars in Vascular Medicine*, 4(03), pp. 241–248. doi: 10.1055/s-2004-861491.

Soutar, A. K., Naoumova, R. P. and Traub, L. M. (2003) 'Genetics, Clinical Phenotype, and Molecular Cell Biology of Autosomal Recessive Hypercholesterolemia', *Arteriosclerosis, Thrombosis, and Vascular Biology*, 23(11), pp. 1963–1970.

StataCorp (2015) *Stata Statistical Software: Release 14*, College Station, TX: StataCorp LP. Available at: <https://www.stata.com/support/faqs/resources/citing-software-documentation-faqs/>.

Stecher, R. M. and Hersh, A. H. (1949) 'Note on the Genetics of Hypercholesterolemia.', *Science*, 109(2821), pp. 61–62. doi: 10.1126/science.109.2821.61-a.

Stein, E. A. (2012) 'Effect of a monoclonal antibody to PCSK9 on LDL cholesterol', *New England Journal of Medicine*, 366(12), pp. 1108–1118.

Steyn, K. and Fourie, J. M. (2007) 'Heart Disease in South Africa', *The Heart and Stroke Foundation South Africa*, pp. 1–29. Available at: [http://www.heartfoundation.co.za/sites/default/files/Heart Disease in SA MRC Report.pdf](http://www.heartfoundation.co.za/sites/default/files/Heart%20Disease%20in%20SA%20MRC%20Report.pdf) (Accessed: 25 June 2017).

Szumilas, M. (2010) 'Explaining odds ratios.', *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3), pp. 227–9.

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

Talmud, P. J., Futema, M. and Humphries, S. E. (2014) 'The genetic architecture of the familial hyperlipidaemia syndromes : rare mutations and common variants in multiple genes', *Current Opinion in Lipidology*, 25(4), pp. 274–281. doi: 10.1097/MOL.0000000000000090.

Teo, Y. Y., Small, K. S. and Kwiatkowski, D. P. (2010) 'Methodological challenges of genome-wide association analysis in Africa', *Nature Reviews Genetics*, 11(2), pp. 149–160. doi: 10.1038/nrg2731.

Teslovich, T. M., Musunuru, K., Smith, A. V., *et al.* (2010) 'Biological, Clinical, and Population Relevance of 95 Loci for Blood Lipids', *Nature*, 466(7307), pp. 707–713. doi: 10.1038/nature09270.

The 1000 Genomes Project Consortium, Auton, A., Abecasis, G. R., *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

Tian, C., Gregersen, P. K. and Seldin, M. F. (2008) 'Accounting for ancestry: population substructure and genome-wide association studies', *Human Molecular Genetics*, 17(2), pp. R143–R150. doi: 10.1093/hmg/ddn268.

Toth, P. P. (2005) 'The "Good Cholesterol" High-Density Lipoprotein', *Circulation*, 111, pp. e90–e91. doi: 10.1161/01.CIR.0000126889.97626.B8.

Varret, M., Rabès, J. P., Saint-Jore, B., *et al.* (1999) 'A third major locus for autosomal dominant hypercholesterolemia maps to 1p34.1-p32.', *American journal of human genetics*, 64(5), pp. 1378–1387. doi: 10.1086/302370.

Versmissen, J., Oosterveer, D. M., Yazdanpanah, M., *et al.* (2008) 'Efficacy of statins in familial hypercholesterolaemia: a long term cohort study.', *BMJ (Clinical research ed.)*, 337, pp. 1–6. doi: 10.1136/bmj.a2423.

Vos, A., Devillé, W., Barth, R., *et al.* (2017) 'HIV infection and cardiovascular risk profile in a rural South African population: The Ndlovu cohort study', *BMJ Global*

Health, 2(Suppl 2), pp. A1–A67. doi: 10.1136/bmjgh-2016-000260.22.

Walker, S. H. and Duncan, D. B. (1967) 'Estimation of the Probability of an Event as a Function of Several Independent Variables', *Biometrika*, 54(1–2), pp. 167–179. doi: 10.2307/2333860.

Walley, A. J., Asher, J. E. and Froguel, P. (2009) 'The genetic contribution to non-syndromic human obesity', *Nature Reviews Genetics*, 10(7), pp. 431–442. doi: 10.1038/nrg2594.

Watts, G. F., Gidding, S., Wierzbicki, A. S., *et al.* (2014) 'Integrated guidance on the care of familial hypercholesterolemia from the International FH Foundation', *Journal of Clinical Lipidology*, 8(2), pp. 148–172. doi: 10.1016/j.jacl.2014.01.002.

Weiss, L. A., Pan, L., Abney, M., *et al.* (2006) 'The sex-specific genetic architecture of quantitative traits in humans', *Nature Genetics*, 38(2), pp. 218–222. doi: 10.1038/ng1726.

Wiegman, A., Gidding, S. S., Watts, G. F., *et al.* (2015) 'Familial hypercholesterolaemia in children and adolescents : gaining decades of life by optimizing detection and treatment', *European Heart Journal*, 36(36), pp. 2425–2437. doi: 10.1093/eurheartj/ehv157.

Willer, C. J., Schmidt, E. M., Sengupta, S., *et al.* (2013) 'Discovery and refinement of loci associated with lipid levels.', *Nature genetics*, 45(11), pp. 1274–1283. doi: 10.1038/ng.2797.

Williams, R. R., Hunt, S. C., Hopkins, P. N., *et al.* (2008) 'Evidence for single gene contributions to hypertension and lipid disturbances: definition, genetics, and clinical significance', *Clinical Genetics*, 46(1), pp. 80–87. doi: 10.1111/j.1399-0004.1994.tb04207.x.

Wilson, P. A., Gardner, S. D., Lambie, N. M., *et al.* (2006) 'Characterization of the human patatin-like phospholipase family.', *Journal of lipid research*, 47(9), pp. 1940–1949. doi: 10.1194/jlr.M600185-JLR200.

Wilson, P. W., D'Agostino, R. B., Levy, D., *et al.* (1998) 'Prediction of coronary heart disease using risk factor categories.', *Circulation*, 97(18), pp. 1837–1847. doi: 10.1161/01.CIR.97.18.1837.

Wooster, R., Bignell, G., Lancaster, J., *et al.* (1995) 'Identification of the breast cancer susceptibility gene BRCA2', *Nature*, 378(6559), pp. 789–792. doi: 10.1038/378789a0.

World Health Organisation (2014) *Noncommunicable Diseases Country Profiles 2014*. Available at: http://apps.who.int/iris/bitstream/10665/128038/1/9789241507509_eng.pdf?ua=1 (Accessed: 19 February 2018).

World Health Organisation (2015) 'WHO | Projections of mortality and causes of death, 2015 and 2030', *World Health Organisation*. Available at: http://www.who.int/healthinfo/global_burden_disease/projections/en/ (Accessed: 19 September 2017).

World Health Organisation (2017) 'Cardiovascular diseases (CVDs)', *World Health Organisation*. World Health Organization. Available at: <http://www.who.int/mediacentre/factsheets/fs317/en/> (Accessed: 29 August 2017).

Xu, C., Fang, J., Shen, H., *et al.* (2018) 'EPS-LASSO: Test for High-Dimensional Regression Under Extreme Phenotype Sampling of Continuous Traits.', *Bioinformatics*, p. epu ahead of print. doi: 10.1093/bioinformatics/bty042.

Yang, J., Jiang, H., Yeh, C.-T., *et al.* (2015) 'Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel', *The Plant Journal*, 84(3), pp. 587–596. doi: 10.1111/tpj.13029.

Zhou, D. T., Kodogo, V., Chokuona, K. F. V., *et al.* (2015) 'Dyslipidemia and cardiovascular disease risk profiles of patients attending an HIV treatment clinic in Harare, Zimbabwe.', *HIV/AIDS (Auckland, N.Z.)*, 7, pp. 145–155. doi: 10.2147/HIV.S78523.

Zidar, D. A., Juchnowski, S., Ferrari, B., *et al.* (2015) 'Oxidized LDL Levels Are Increased in HIV Infection and May Drive Monocyte Activation.', *Journal of acquired immune deficiency syndromes (1999)*, 69(2), pp. 154–160. doi: 10.1097/QAI.0000000000000566.

Van Zyl, T., Jerling, J. C., Conradie, K. R., *et al.* (2014) 'Common and rare single nucleotide polymorphisms in the LDLR gene are present in a black South African population and associate with low-density lipoprotein cholesterol levels', *Journal of Human Genetics*, 59(2), pp. 88–94. doi: 10.1038/jhg.2013.123.

7. Appendices

7.1. Appendix A: Plagiarism report and document

603401:turnitin.docx			
ORIGINALITY REPORT			
9%	2%	8%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
PRIMARY SOURCES			
1	Kathiresan, Sekar, and Daniel J. Rader. "Lipoprotein Disorders", Genomic and Personalized Medicine, 2013. Publication		<1%
2	Imes, C. C., and M. A. Austin. "Low-Density Lipoprotein Cholesterol, Apolipoprotein B, and Risk of Coronary Heart Disease: From Familial Hyperlipidemia to Genomics", Biological Research for Nursing, 2012. Publication		<1%
3	Ana L. Torres, Sital Moorjani, Marie-Claude Vohl, Claude Gagné et al. "Heterozygous familial hypercholesterolemia in children: low-density lipoprotein receptor mutational analysis and variation in the expression of plasma lipoprotein-lipid concentrations", Atherosclerosis, 1996 Publication		<1%
4	Submitted to University of Queensland Student Paper		<1%

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Mahtaab Hayat (Student number: 603401) am a student
registered for the degree of MSc (Med) in the academic year 2018.

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: M Hayat

Date: 26 February 2018

7.2. Appendix B: Ethics clearance certificate



R14/49 Miss Mahtaab Hayat et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M160833

NAME: Miss Mahtaab Hayat et al
(Principal Investigator)
DEPARTMENT: Division of Human Genetics
Sydney Brenner Institute for Molecular Bioscience

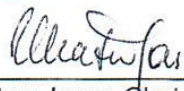
PROJECT TITLE: Genetic variants in Four Genes Associated with
Lipid Levels: A Study in African Populations

DATE CONSIDERED: 26/08/2016

DECISION: Approved unconditionally

CONDITIONS:

SUPERVISOR: Prof Michele Ramsay

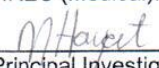
APPROVED BY: 
Professor P Cleaton-Jones, Chairperson, HREC (Medical)

DATE OF APPROVAL: 20/01/2017

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 in Room 301, Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. 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 resubmit the application to the Committee. **I agree to submit a yearly progress report.** The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in August and will therefore be due in the month of August each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).


Principal Investigator Signature

23/01/2017
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

7.3. Appendix C: Various tables

Table 1: Variants chosen according to deleteriousness, type of variant and MAF

Low Density Lipoprotein Receptor (<i>LDLR</i>)								
Genomic location	rs number	PolyPhen2 score	SIFT score	CADD score	African MAF	Global MAF	Type of variant	Notes
19:11210921	rs72658855	-	-	15.14	T = 0.04	T = 0.01	synonymous	
19:11222300	rs11669576	benign (0.08)	Tolerated (1)	1.68	A = 0.20	A = 0.07	missense	
19:11226800	rs5929	-	-	12.1	T = 0.12	T = 0.12	synonymous	
19:11242133	rs3826810	Unknown (0)	-	4.198	A = 0.12	A = 0.08	missense	
Apolipoprotein B (<i>APOB</i>)								
Genomic location	rs number	PolyPhen2 score	SIFT score	CADD score	African MAF	Global MAF	Type of variant	Notes
2:21225753	rs1042031	Benign (0.00)	-	1.71	T = 0.16	T = 0.12	missense	Failed MassARRAY test
2:21229860	rs12720855	possibly damaging (0.64)	-	23.60	G = 0.08	G = 0.02	missense	
2:21231524	rs676210	probably damaging (0.99)	-	27.10	A = 0.12	A = 0.37	missense	Failed MassARRAY test
2:21250914	rs679899	possibly damaging (0.64)	Tolerated (0.12)	26.60	A = 0.13	A = 0.49	missense	
2:21260934	rs6752026	probably damaging (0.92)	Deleterious (0.03)	25.30	A = 0.11	A = 0.03	missense	
Proprotein convertase subtilisin/kexin type 9 (<i>PCSK9</i>)								
Genomic location	rs number	PolyPhen2 score	SIFT score	CADD score	African MAF	Global MAF	Type of variant	Notes
1:55518370	rs7552471	-	-	20.40	T = 0.08	T = 0.02	synonymous	
1:55523855	rs28362263			9.68	A = 0.09	A = 0.03	missense	Failed MassARRAY test
Low Density Lipoprotein Receptor Adaptor Protein 1 (<i>LDLRAP1</i>)								
Genomic location	rs number	PolyPhen2 score	SIFT score	CADD score	African MAF	Global MAF	Type of variant	Notes
1:25889539	rs12071264	-	-	4.70	G = 0.14	G = 0.04	intronic	
1:25893927	rs35910270	-	-	4.73	-0.42	-0.49	frameshift	

Genomic locations are reported using NCBI Build 37 (hg19)

Table 2: Variants chosen according to MAF only

Low Density Lipoprotein Receptor (<i>LDLR</i>)								
Genomic location	rs number	PolyPhen2 score	SIFT score	CADD score	African MAF	Global MAF	Type of variant	Notes
19:11230881	rs5925	-	-	0.51	C = 0.15	C = 0.34	synonymous	
19:11238239	rs2569540	Probably damaging (0.96)	Deleterious (0)	1.20	C = 0.42	C = 0.32	missense	
19:11237772	rs2569542	-	-	0.34	A = 0.54	A = 0.31	stop lost	Failed MassARRAY test
19:11239618	rs17242635			1.73	A = 0.24	A = 0.17	splice region	
19:11239701	rs201052824	-	-	3.20	Del = 0.69	Del = 0.41	frameshift	Failed MassARRAY test
19:11232199	rs2569546	Possibly damaging (0.66)	Tolerated (0.12)	1.41	A = 0.67	A = 0.38	missense	Failed MassARRAY test
Apolipoprotein B (<i>APOB</i>)								
Genomic location	rs number	PolyPhen2 score	SIFT score	CADD score	African MAF	Global MAF	Type of variant	Notes
2:21239661	rs12720820	-	-	1.82	C = 0.24	C = 0.15	regulatory region	
2:21246306	rs12714102	-	-	0.62	C = 0.43	C = 0.20	regulatory region	
2:21245367	rs3791981	-	-	2.20	G = 0.43	G = 0.20	regulatory region	
2:21223763	rs58411594	-	-	0.40	T = 0.34	T = 0.12	downstream gene	Failed MassARRAY test
Proprotein convertase subtilisin/kexin type 9 (<i>PCSK9</i>)								
Genomic location	rs number	PolyPhen2 score	SIFT score	CADD score	African MAF	Global MAF	Type of variant	Notes
1:55509872	rs4927193	-	-	3.89	C = 0.22	C = 0.15	downstream gene	
1:55518622	rs45613943	-	-	3.55	C = 0.29	C = 0.12	regulatory region	
1:55526428	rs11206517	-	-	1.29	G = 0.21	G = 0.08	regulatory region	Failed MassARRAY test
1:55521313	rs472495	-	-	2.39	T = 0.32	G = 0.42	regulatory region	Failed MassARRAY test
Low Density Lipoprotein Receptor Adaptor Protein 1 (<i>LDLRAP1</i>)								
Genomic location	rs number	PolyPhen2 score	SIFT score	CADD score	African MAF	Global MAF	Type of variant	Notes
1:25876516	rs74425832	-	-	6.89	G = 0.25	G = 0.08	regulatory region	Failed MassARRAY test
1:25890373	rs13373894	-	-	1.74	A = 0.38	A = 0.11	intron	

Genomic locations are reported using NCBI Build 37 (hg19)

Gene	rs number	Missingness	N	Minor Allele (A1)	All data HWE	All MAF	High LDL					Low LDL				
							A1/A1 n(f)	A1/A2 n(f)	A2/A2 n(f)	MAF	HWE	A1/A1 n(f)	A1/A2 n(f)	A2/A2 n(f)	MAF	HWE
LDL RAP 1	rs12071264	<0.01	995	G	0.70	0.09	3	60	435	0.07	0.47	6	101	390	0.11	1.00
	rs35910270	<0.01	995	G	0.84	0.40	65	242	191	0.37	0.44	96	232	169	0.43	0.31
PCSK 9	rs4927193	0.01	990	C	0.16	0.24	40	157	297	0.24	0.01	24	185	287	0.24	0.46
	rs7552471	<0.01	991	T	0.69	0.09	2	84	408	0.09	0.41	4	75	418	0.08	0.77
	rs45613943	<0.01	994	C	0.58	0.28	31	174	292	0.24	0.46	49	217	231	0.32	0.92
APOB	rs12720855	<0.01	995	G	0.51	0.08	4	63	431	0.07	0.30	3	75	419	0.08	1.00
	rs3791981	<0.01	994	G	0.66	0.48	103	253	141	0.46	0.65	125	236	136	0.49	0.28
	rs679899	<0.01	995	A	0.03	0.12	18	110	370	0.15	0.01	5	89	403	0.10	1.00
	rs6752026	<0.01	995	A	0.34	0.14	4	93	401	0.10	0.81	10	147	340	0.17	0.26
LDLR	rs72658855	<0.01	995	T	0.23	0.03	0	20	478	0.02	1.00	2	37	458	0.04	0.20
	rs5929	<0.01	994	T	0.88	0.12	2	98	398	0.10	0.15	12	106	378	0.13	0.17
	rs5925	<0.01	996	C	0.14	0.15	15	134	350	0.16	0.63	14	111	372	0.14	0.13
	rs2569540	<0.01	994	G	0.56	0.43	89	249	159	0.43	0.65	101	229	167	0.43	0.17
	rs3826810	<0.01	995	A	0.31	0.11	5	80	413	0.09	0.58	9	102	386	0.12	0.40

Table 3: 29 variants initially selected for genotyping in 1000 individuals

Table 4: MAF differences between geographical regions

Gene	rs number	E/W	E/S	W/S
LDLR	rs72658855	0.01	<0.01	<0.01
	rs5929	0.25	<0.01	0.01
	rs2569540	0.26	<0.01	0.01
	rs3826810	<0.01	<0.01	<0.01
	rs5925	<0.01	<0.01	5.16 x10⁻⁵
APOB	rs12720855	0.49	<0.01	<0.01
	rs679899	1.20 x10⁻⁷	1.03 x10⁻⁷	0.01
	rs6752026	2.41 x10⁻¹⁶	<0.01	3.01 x10⁻¹⁰
	rs3791981	0.28	<0.001	0.01
PCSK9	rs7552471	0.01	<0.01	4.34 x10⁻⁵
	rs4927193	0.26	2.22 x10⁻⁵	<0.01
	rs45613943	4.07 x10⁻⁷	3.60 x10⁻⁵	<0.01
LDLRAP1	rs12071264	4.00 x10⁻⁴	1.00 x10⁻³	2.78 x10⁻⁶
	rs35910270	3.89 x10⁻⁵	1.00 x10⁻³	3.89 x10⁻⁷

E/W: p value of frequencies between East and West Africa

E/S: p value of frequencies between East and South Africa

W/S: p value of frequencies between South and West Africa

Significant difference p<0.05 (highlighted in bold)