APOE, *PCSK9*, and *CETP* genetic variants as potential biomarkers of dyslipidaemia in black South Africans with Type 2 Diabetes Mellitus

---

Jonathan Evans

Supervisor: Prof. Collet Dandara

Presented for M.Sc. (Med) in Human Genetics

Division of Human Genetics

Faculty of Health Sciences

University of Cape Town

May 2018

Declaration

I, Jonathan Evans, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: 29/05/2018

Abstract

Dyslipidaemia is a commonly encountered clinical condition and is a major risk factor for cardiovascular diseases. Although there are many factors associated with dyslipidaemia, a strong genetic component is evident. Apolipoprotein E (APOE), proprotein convertase subtilisin/kexin type 9 (PCSK9), and cholesteryl ester transfer protein (CETP) are key regulators of plasma cholesterol levels. Thus, genetic variation in the genes coding for these proteins contributes to dyslipidaemia. In this study, a cohort of black South African Type 2 Diabetes Mellitus (T2DM) patients was characterized for mutations in genes coding for APOE, PCSK9, and CETP, and the possible effects of these variants on their lipid profiles was evaluated.

Participants (n=417) were recruited from the Chris Hani Baragwaneth Hospital Diabetes Clinic, Johannesburg from whom blood samples were obtained for DNA extraction. The cohort was further stratified into two groups; individuals on statin treatment (Sim+, n=291), and the second that was not on treatment (Sim-, n=87). Lipid profiles were determined by enzymatic methods. DNA was genotyped for *APOE*, *PCSK9*, and *CETP* variants using PCR-RFLP and Sanger sequencing. Analysis of the effects of the genetic variants was carried out in two ways. Firstly, for all the participants combined, and then by separating those on statin treatment from those without (Sim+ vs. Sim-). Genotype and allele frequencies were calculated followed by genotype-phenotype correlations with lipid profiles.

Univariate analysis showed a significant association between the APOE4 isoform and lower HDL-c levels in the combined cohort (p=0.034). The effects were more pronounced in the Sim- group (p=0.004) but were absent in the Sim+ group. Contrary to above, APOE2 was significantly associated with lower total cholesterol (TC) (p<0.001) and lower LDL-c (p<0.001) when compared to APOE3 in the combined cohort. Upon analysing treatment groups, the correlations were observed in the Sim+ group (p=0.027 and p=0.003, respectively), while there were no observed correlations in the Sim- group. The *CETP rs34065661C/G* and *G/G* genotypes were significantly associated with increased HDL-c levels (p=0.017; when applying a dominant genetic model) in the combined cohort, as well as in the Sim+ group (p=0.026). Multivariate analysis, using a generalized linear model, confirmed associations between *APOE rs429358C* and lower HDL-c (OR=0.881, p=1.64e-04), and *APOE rs7412T* and decreased LDL-c (OR=0.759, p=0.012). No significant associations were observed for *PCSK9* polymorphisms.

We report significant associations between *APOE* and *CETP* genetic variations and altered lipid levels in this black South African T2DM population. These genetic variants could be biomarkers for dyslipidaemia among Africans. However, it is imperative that the *APOE*, *PCSK9*, and *CETP* genes are fully characterized for additional polymorphisms in order to come up with a better genetic profile that explains the variance in lipid levels observed in the black South African population. The impact of these genetic variants could be relevant to other black African populations as well.

Acknowledgments

I would like to acknowledge the following people and institutions for playing a major role in the completion of this MSc project and dissertation.

I would firstly like to thank the University of Cape Town for providing the opportunity for young South African students to engage in cutting edge fields of science by providing world class facilities and faculty. This is backed up by generous student support through excellent administration and funding opportunities. I would also like to thank the Division of Human Genetics, and especially the research group under Prof. Collet Dandara, for being my home for four years. It is through my daily interactions with some of the finest minds in Human Genetics in South Africa that I have learnt so much about field, as well as academia as a whole. The support provided by these people is greatly appreciated. A special thanks to Dr. Nicholas Thomford for taking the time to mentor me in the areas of academia that lie behind the scenes. Thank you to Dr. Donald Tanyanyiwa for all his hard work and for making this project a possibility.

To my supervisor, Prof. Collet Dandara, I must express the greatest of gratitude. He has guided me through my postgraduate career with a body of knowledge, sense of empathy, and genuine concern for my development and growth that is rare to find in an academic supervisor. Prof. Dandara is a true asset to the division and the university.

On a more personal note, I would like to thank my family for supporting me through this process, unconditionally. To my mother Kathleen Evans, this acknowledgement cannot do justice to level of gratitude I have for the support, care, and interest given by you every day. Thank you to my sister Sarah for always being there to listen to me vent and for being so keen to hear and brag about the most mundane aspects of my project. To the newest member of our family, my niece Georgia Mary, thank you for putting a smile on my face every day through the writing process. Thank you to Dr. Simone Jenneke, for being a source of light and happiness in a time when it was much needed. I will always appreciate you.

Lastly, I would like to thank the National Research Foundation and the University of Cape Town for funding my studies, this project, and the travel opportunities that arose as a result, thereof. I am truly grateful for the support these institutions have given me throughout my academic career so far.

Contents

Abbreviations

°C – degree Celsius

A – adenine

ABCA1 - adenosine triphosphate binding cassette A1

ACAT- acyl-CoA: cholesterol acyltransferase

ACM- acute myocardial infarction

APO- apolipoprotein

APOE - apolipoprotein E

BMI – body mass index

bp – base pair

C – Cysteine

C – cytosine

CDC - Centres for Disease Control and Prevention

CE - cholesterol esterase

CETP - cholesteryl ester transfer protein

CEU - Utah residents with Northern and Western European ancestry

CHB - Han Chinese from China

CHOD - cholesterol oxidase

CI - confidence interval

cm – centimetre

CM- chylomicron

CVD- cardiovascular diseases

DD - diabetic dyslipidaemia

ddNTPs – dideoxynucleotides

dF - degree of freedom

dL – decilitre

DNA – deoxyribonucleic acid

dNTPs - deoxynucleotides

E – Glutamic acid

EAS - European Atherosclerosis Society

EDTA - ethylenediaminetetraacetic acid

ER - endoplasmic reticulum

ESC - European Society of Cardiology

*Exo*I- *Exonuclease* I

FE - Fischer Exact test

G – Glycine

g – gravity

G – guanine

GK – glycerokinase

GLM - generalised linear model

GPO - glycerol phosphate oxidase

GWAS - genome wide association studies

Hb – haemoglobin

HbA1c - haemoglobin A1C

HDL-c - high-density lipoprotein cholesterol

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

HSDA - Sodium N-(2-hydroxy-3-sulfopropyl) -3,5-dimethoxyaniline

HSPG- heparan sulphate proteoglycans

HWE - Hardy-Weinberg Equilibrium

IDF - International Diabetes Federation

IDL - intermediate density lipoprotein

kb – kilobase

kDa – kilodalton

kg – kilogram

L – litre

LCAT - lecithin cholesteryl acyl transferase

LD - linkage disequilibrium

LDL-c - low-density lipoprotein cholesterol

LDLR - low-density lipoprotein receptors

LPA -lysophosphatidic acid

LPL- lipoprotein lipase

LWK - Luyha from Kenya

m – metre

MAF - minor allele frequency

mg – milligram

$MgCl_2$ - magnesium chloride

MI - myocardial infarction

mL – millilitre

mL – millilitre

mM – millimolar

mmol – millimoles

mRNA – messenger ribonucleic acid

MTP - microsomal triglyceride transfer protein

NaCl - sodium chloride

NaOAc – sodium acetate

NCD- non-communicable diseases

ng – nanogram

nm – nanometre

NRF - National Research Foundation

OD - optical density

OR – odds ratio

PBS- phosphate buffered saline

PCR - polymerase chain reaction

PCSK9 - pro-protein convertase subtilisin/kexin type 9

PEG - polyethylene glycol

pH – potential of hydrogen

PLTP- phospholipid transfer protein

PNPLA5 - patatin like phospholipase domain containing 5

POD – peroxidase

Q- Glutamine

RE - restriction enzymes

RFLP - restriction fragment length polymorphism

SD - standard deviation

sdH_20 - sterile distilled water

SDS- sodium dodecyl sulphate

Sim- - not on statin therapy

Sim+ - on statin therapy

SNP – single nucleotide polymorphism

SRB1 - scavenger receptor protein BI

SREBP- sterol-regulated transcription protein

T – thymine

T2DM - type 2 diabetes mellitus

Ta - annealing temperature

TBE - Tris-Borate EDTA

TC- total cholesterol

TG – triglycerides

TINIA - turbidimetric inhibition immunoassay

U – units

ug – microgram

uL – microlitre

uM – micromolar

UV – ultraviolet

V – volt

v/v – volume/volume

VLDL - very low-density lipoprotein

w/v – weight/volume

WHO- World Health Organization

X – stop codon

Y – Tyrosine

YRI - Yoruba from Nigeria

χ^2 - Chi^2

List of Figures

List of Tables

Summary of thesis outline

- In **Chapter 1**, we deal with the motivation and justification of this study. We review the recent literature on the genetics of dyslipidaemia in order to gain perspective on the topic and how this can used to guide our own study. A clear understanding of potential aspects of dyslipidaemia in type 2 diabetes mellitus patients and how this problem can be tackled in our own study is further explored. We outline the research topic, namely the genetics of dyslipidaemia in black South Africans, in the context of its increasing burden to the global and local health systems. In doing so, we will put the research problem into perspective, outlining its broad impact on society, and why it is necessary to study the genetics of complex non-communicable diseases such as dyslipidaemia in South Africans at all.
- **Chapter 2** outlines a comprehensive explanation of the methods and materials used to achieve the aims and objectives set out in Chapter 1. We outline the exact approach used to investigate the impact of genetic variants on dyslipidaemia in a T2DM cohort.
- **Chapter 3** presents the results of the experiments performed to complete the objectives outlined in Chapter 2. A comprehensive overview of the data analysis employed to achieve the aim of the study is presented.
- In **Chapter 4** we further discuss the results obtained from the experiments we employed to investigate the contribution of genetic variations in *APOE, PCSK9,* and *CETP* on the aberrant lipid profiles of black South Africans with T2DM. We evaluate these results and discuss any inferences or conclusions which can be taken from them, whilst comparing this to the literature. We also discuss the technical and analytical issues that are specific to this project, and how these might be overcome in the future.

Chapter 1: Literature Review

1.1) Non-communicable disease (NCD)

Non-communicable diseases (NCD) are slowly progressing disorders, which are typically of long duration, and are not transferred by a pathogen host. They fall into four main groups: cardiovascular diseases (CVD), cancers, chronic respiratory diseases and diabetes. NCDs are rapidly becoming a global health epidemic. The World Health Organization (WHO) estimated that in 2015, out of a total of 56.4 million reported global deaths, 39.5 million (70.1 %) were a direct result of NCDs[1,2]. This is up from the 31.4 million (60.2 %) global deaths which were reported in 2000[2].

The reports show that the number of NCD related deaths has increased in every global region since 2000, and is expected to rise to a staggering 52 million by 2030[1,3]. A commonly held view of NCDs is that it is a "Western" phenomenon. The burden of NCD deaths is highest in high-income countries, with 88 % of recorded deaths a result of NCDs compared to 37 % in low-income countries[3]. However, 78 % of total global deaths due to NCDs occurred in low- and middle-income countries in 2015[3]. **Figure 1.1**, shows the probability of dying from the four main NCDs by country, highlighting the increasing burden in these regions.

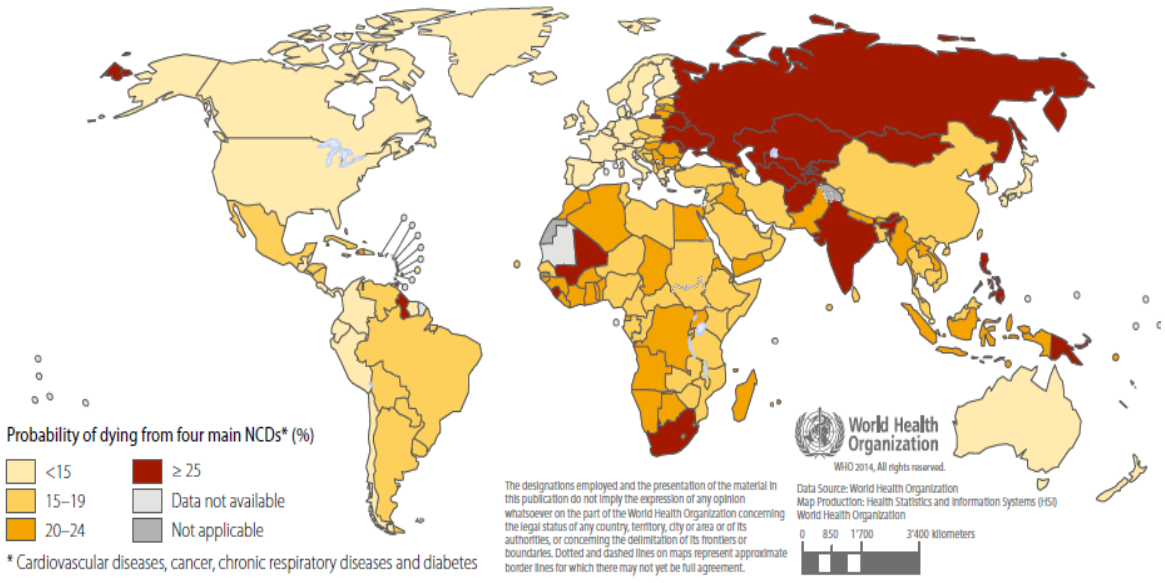


Figure 1.1: Probability of dying from the four main NCD between the ages of 30 and 70 years in 2012. **Adapted from**: WHO. Global status report on noncommunicable diseases. (2014).

The burden of NCDs in Africa have increased substantially as well. In 2000, the number of NCD related deaths in the African region was 2.17 million (22.2 %). By 2015, the number had risen to 3.08 million deaths (33.5 %)[2]. The scale of the rising crisis hits home with six out of the top ten causes of natural deaths in South Africa in 2015 being a direct result of NCD, accounting for 27.7 % of deaths in females and 32.5 % of deaths in males[4]. The leading cause of NCD deaths in South Africa and, indeed, the rest of the world are CVDs.

1.2) Cardiovascular disease (CVD)

CVDs are a group of disorders affecting the circulatory system, namely the blood vessels and the heart. Of the reported 17.6 million global CVD deaths in 2015, 8.75 million were due to ischaemic heart disease (e.g. heart attacks), and 6.24 million deaths were caused by cerebrovascular disease (e.g. stroke)[2]. CVDs caused by atherosclerosis are responsible for the largest proportion of CVD related deaths.

Atherosclerosis is a complex inflammatory disease process in which the walls of medium-sized and large- sized arteries are mainly affected[5,6]. Atherosclerosis occurs over time, can be present throughout a person's lifetime, and may present as early as infancy[5,7]. The process of atherosclerosis can be initiated by the exposure of the endothelium of the blood vessels to elevated levels of low-density lipoprotein cholesterol (LDL-c) or free radicals (e.g. caused by cigarette smoking), resulting in a dysfunctional endothelium which becomes permeable to monocytes and lymphocytes. These cells then emigrate from the blood to the deeper layers of the artery cell walls, multiplying in a forming cell lesion. Activation of these cells results in the release of cytokines, chemokines, growth factors, and hydrolytic enzymes which continue to cause further damage[5]. LDL-c can be attracted to and trapped in the lesion site where it is modified by oxidation in the artery cell walls through a number of reactions[8]. The LDL-c particles are engulfed by monocytes, which are in turn transformed into macrophages. Smooth muscle cells migrate to the lesion site, and fibrous tissue is formed, further enlarging the lesion. It eventually becomes covered by a fibrous cap, consisting of smooth muscle cells and collagen that covers a core of lipid and necrotic tissue[5,6]. The lesions are capable of accumulating even more lipids and cells and enlarge to the point where they protrude into the artery lumen. If the lesions burst, cellular debris and lipid fragments are released which can form a thrombus if exposed to thrombogenic agents on the endothelial surface. A large

enough thrombus can block a coronary artery or cerebral blood vessel, resulting in a heart attack or stroke, respectively[5,6].

1.3) Individual behavioural CVD risk factors:

There are number of factors which have been reported to modify the risk of atherosclerosis induced CVD. These include preventable environmental and behavioural lifestyle factors, as well as CVD disease comorbidities[9]. The exact role and effect size of these risk factors has been thoroughly investigated in a number of large-scale studies, due to the overwhelming global burden of CVD. **Figure 1.2** shows the intricate relationship between associated CVD risk factors and comorbidities[10].

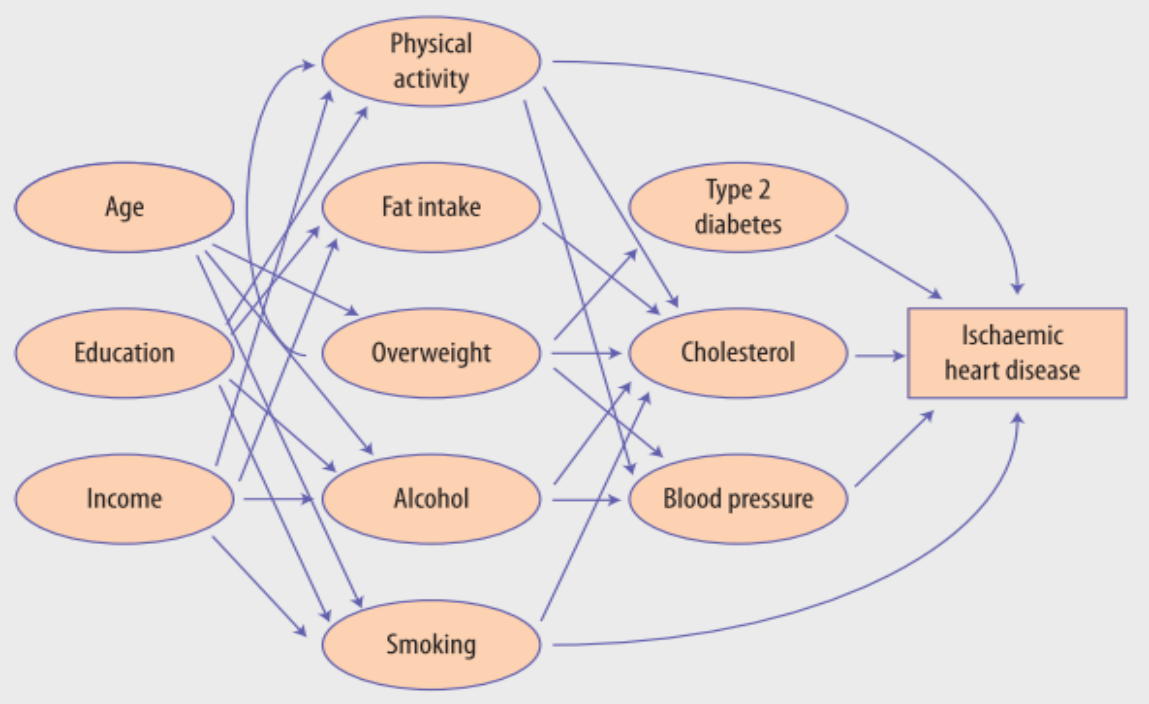


Figure 1.2: The major causal chain of ischaemic heart disease which shows the intricate relationship between CVD risk factors and comorbidities. **Adapted from:** WHO. Global Health Risks: Mortality and burden of disease attributable to selected major risks. *Bull. World Health Organ.* **87,** 646–646 (2009).

1.3.1) Smoking

It is currently estimated that 31 % of men and 6 % of women smoke some form of tobacco product every day and although the use of tobacco based products has been steadily declining in most countries since 1980, smoking remains one of the highest reported CVD risk factors[9,11]. Patterns of smoking can vary between different global regions, sex, and socioeconomic status. The WHO estimates that the smoking of tobacco products causes approximately 10 % of reported CVD cases[10]. The INTERHEART case-control study examined modifiable risk factors of acute myocardial infarction (ACM) events in 15 152 cases and 14 820 sex- and age- matched controls across 52 countries[12]. What they found is that regular smoking was significantly associated with non-fatal ACM (odds ratio (OR) 2.95, 95 % confidence interval (CI) 2.77–3.14, $p<0.0001$) when compared to non-smoking controls, and that smoking even one cigarette a day increased your risk of ACM substantially[12]. They also reported on a linear relationship between the number of cigarettes smoked per day and increasing ACM risk, with the risk increased by 5.6 % per additional cigarette smoked[12]. A meta-analysis and literature review by Barnoya and Glantz found that the effects of second hand smoking on the cardiovascular system were, on average, between 80 % and 90 % as large as chronic active smoking[13]. The evidence is clear that any form of tobacco smoking greatly increases your risk of CVD and as such, it should always be accounted for when studying CVD related co-morbidities.

1.3.2) Alcohol consumption

The excessive use, or abuse of alcohol is associated with a multitude of negative health outcomes including severe liver cirrhosis, pancreatitis, and neuropathy[6]. However, the relationship between general alcohol use, and CVD risk is more complex. There have been reports in which different ranges of CVD risks are found depending on the amount of alcohol consumed[14]. It has also been thought that low levels of alcohol consumption are associated with lower risk of CVD, however, an extensive review of alcohol and health by the United Kingdom's Chief Medical Officers found that any protective effect was limited to the elderly, who are already at high risk for ischaemic CVD[15]. Additionally, they report that there may be confounding factors which were unaccounted that may contribute to the associated protection[15]. Conversely, heavy consumption of alcohol has been consistently associated with an increased risk for CVD[9,14].

1.3.3) Obesity

There has been a steady increase in the global prevalence of obesity (Body Mass Index ≥30 kg/m^2), with the incidence of obesity rising from 3 % to 11 % in men and 6 % to 15 % in women between 1975 and 2014[16]. The incidence of obesity is highest in high-income countries, but it is not isolated to these regions. In South Africa, approximately 25 % of adults are obese with women being the worst affected and the rate of obesity in children rising[1,17]. Obesity is strongly associated with other reported CVD risk factors, and can cause adverse metabolic effects on blood pressure and lipid profiles[6]. The INTERHEART study found a linear association between Body Mass Index (BMI), waist circumference, and waist–hip ratio and an increased the risk of myocardial infarction[18]. However, the association with BMI was not a predictor of myocardial infarction in those with a history of hypertension or raised APOB/APOA ratio[18]. Conversely, waist-to-hip ratio remained consistently associated with myocardial infarction risk when adjusting for hypertension and raised APOB/APOA, and was found to be a better predictor of risk than waist circumference alone[18]. This suggests that waist-to-hip ratio is a better index of obesity which is as a predictor of CVD in most populations. However, a study done by Gelber *et al.* found that the magnitude of associations of BMI, waist circumference, waist–to-hip ratio and waist– to- height ratio with CVD risk were similar, but that using BMI as an obesity index might miss individuals with increased CVD due to central fat distribution[19].

It must be noted that there are ethnic differences in waist-to-hip ratios. It has been shown that South African black women have significant interindividual variability in BMI at a given body fat percentage[20]. Additionally, they were shown to have less visceral abdominal adipose tissue at a similar waist-to-hip ratio when compared to European women[20,21]. However, the cut-off points for waist-to-hip ratios is largely based on European derived data and may not be valid for all ethnicities. As such, population specific cut-offs must be used as a measure of obesity for CVD risk[20].

1.4) Dyslipidaemia

Dyslipidaemias, are disorders of lipoprotein metabolism and cover a wide spectrum of lipid abnormalities defined by aberrant lipid profiles. Dyslipidaemias are commonly encountered in clinical settings and been consistently identified as a major risk factors for the occurrence of CVD[22–25]. Dyslipidaemias may be primary or secondary. Primary dyslipidaemias are

caused by the interactions of environmental factors with genetic predispositions. Secondary dyslipidaemias are related to other diseases. In order to better understand the pathology of dyslipidaemias, we have to investigate the various lipids components and the molecular mechanisms of the lipid metabolism pathway.

1.4.1) Overview of the lipid metabolism pathway

Cholesterol is an essential constituent of the cell needed for normal cell growth and development, whilst providing the building blocks for several cell processes. A constant supply of cholesterol is needed for these cell processes, which include cell growth, steroid hormone synthesis and the structural maintenance of cellular membranes[26]. Lipids can be made available to the cell through three processes: i) the dietary exogenous pathway, ii) the endogenous pathway, and iii) through the process of reverse cholesterol transport pathway (**Figure 1.3**)[26]. Cholesterol uptake and cholesterol biosynthesis are interdependent, so any changes to dietary uptake or changes to cellular requirements influence the rate of cholesterol biosynthesis through complex feedback mechanisms[26]. The human body synthesises approximately 700-900 mg of cholesterol every day while dietary sources of cholesterol account for 300-600 mg of cholesterol/day[26,27]. The primary sites of cholesterol metabolism are in the intestines and liver.

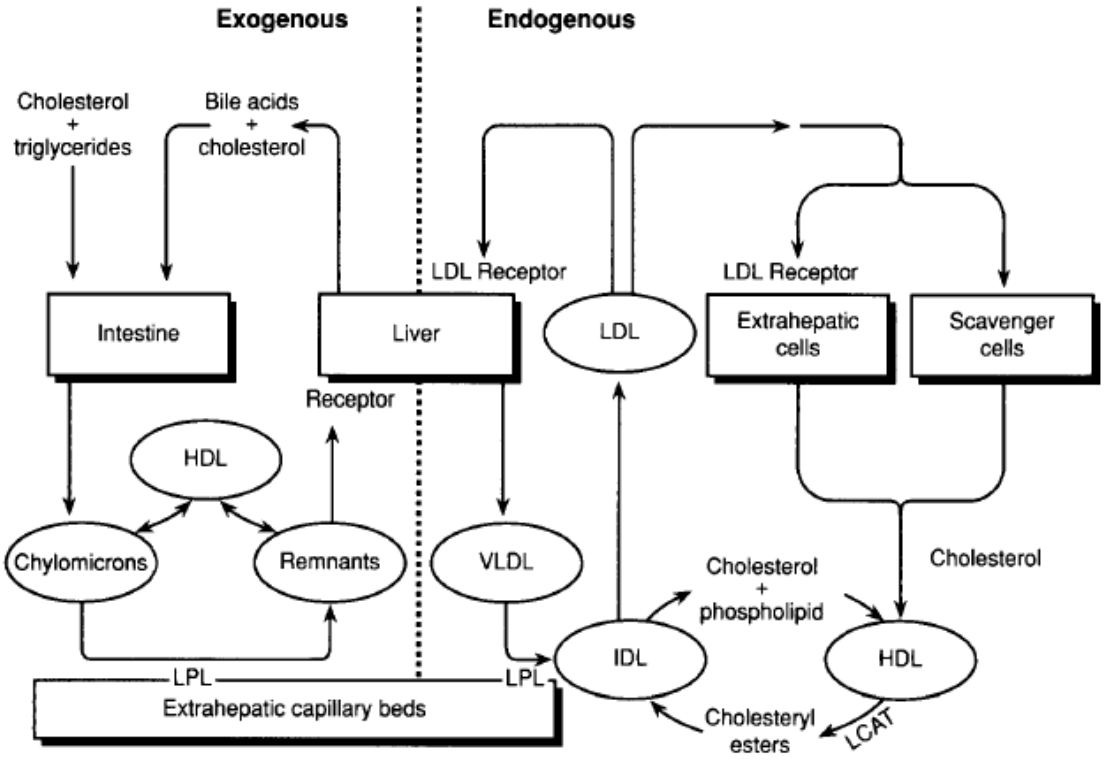


Figure 1.3: A representative image of the lipoprotein transport pathways in humans where the exogenous and endogenous uptake pathways are briefly outlined. **Adapted from:** Russell, D. W. Cholesterol biosynthesis and metabolism. *Cardiovasc. Drugs Ther.* **6,** 103–10 (1992).

1.4.1.1) The exogenous lipid metabolism pathway

In the exogenous pathway, dietary fats consisting of mostly triglycerides (TG), phospholipids, and fat soluble vitamins cholesterols are emulsified by bile salts making them susceptible to hydrolysis by pancreatic lipase[28]. The TGs are hydrolysed into 2-monoacylglycerols and free fatty acids which are then absorbed into intestinal enterocytes. In the intestinal wall, TGs are resynthesized by the 2-monoacylglycerol pathway[29]. Cholesterols are esterified by the endoplasmic reticulum (ER) localised enzymes cholesterol esterase, and acyl-CoA: cholesterol acyltransferase (ACAT) and are packaged along with newly synthesised TG, phospholipids, apolipoprotein B's (APOB), and apolipoprotein A's (APOA) into particles called chylomicrons (CM)[28,29].

CMs circulate throughout the intestinal lymphatic system and in the blood. It is here where they acquire APOCI, APOCII, APOCIII and apolipoprotein E (APOE). As the free CMs circulate they interact with endothelial bound APOCII activated lipoprotein lipase (LPL), which mediates the hydrolysis of CM bound TGs. This results in a reduction of the CM core

volume, and the release of free fatty acids which enter muscle cells for energy production and adipocytes for storage[28]. Some CM components such as phospholipids, free cholesterol, APOCII and APOCIII are repackaged into high-density lipoproteins (HDL). The remaining CM remnants are TG- depleted, and cholesterol ester and APOE enriched. They are then able to interact with the receptors on hepatocytes and are subsequently removed from circulation[28,30].

1.4.1.2) The endogenous lipid metabolism pathway

Endogenous lipoproteins are synthesised and released into circulation where they move to the peripheral tissues in the form of very low-density lipoproteins (VLDL). The VLDL synthesis process is very similar to that of CM synthesis, and VLDL production is synchronised with CM secretion[28]. This is essential to avoid hepatic steatosis and unfavourable lipid concentrations in the blood[31]. TGs are synthesised in the liver in response to free fatty acid influxes from adipocytes, CM remnants, and the intestine[31]. These TGs, along with phospholipids, cholesterol esters, and certain apolipoproteins, are the core components of VLDL particles (**Figure 1.4**)[31,32]. VLDL is synthesised in the lumen of the ER in a two-step process. Newly translated APOB is translocated across the ER into the lumen, where it is partially lipidated to pre-VLDL particles by microsomal triglyceride transfer protein (MTP). It then moves into the Golgi apparatus where it fuses with TG rich particles into mature VLDL[28,31]. Mature VLDL is secreted into the plasma where it undergoes hydrolysis by LPL, generating the smaller denser particles called intermediate density lipoproteins (IDL) and VLDL remnants. Free fatty acids are once again released and used for energy production in muscle cells and storage in adipocytes. During this process, the remnants become enriched with APOE, which is a high affinity ligand for LDL receptors (LDLR)[28]. VLDL remnants and some IDL particles are taken up by the liver via LDLRs, however, IDL particles can undergo further hydrolysis to become LDL particles[28].

Circulating LDL is the principal transporter of endogenous cholesterols and active phospholipids synthesized in the liver (such as lysophosphatidic acid (LPA)) to the rest of the body[32]. LDLs contain fewer lipids in their core than VLDLs but have more lipids than HDLs[32]. The structural composition of LDL can be observed in **Figure 1.4**. LDL particles are catabolised by two pathways: i) A receptor-dependant pathway in the liver, and ii) a receptor independent pathway in non-hepatic tissues[32]. The receptor dependant pathway is mediated

by the LDLR. LDL particles bind with high affinity to the LDLR, which is mediated by APOB100 and APOE. This results in the hydrolysis of LDL leading to the release of cholesterol esters, and degradation of apolipoproteins in the lysosome[33]. LDL derived cholesterol in the lysosome suppresses 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMG-CoA) activity, which is responsible for the rate of cholesterol synthesis by inhibiting the sterol-regulated transcription proteins (SREBPs) pathway[33]. LDL blocks SREBP transcription factor transport and the SREBP's targets'(HMG-CoA and LDLR) transcription is inhibited which results in reduction of cellular cholesterol production[33,34]. In this manner, the amount of cholesterol being produced at any given time is regulated by the total amount of cholesterol present in the liver[33,34].

LDL-c processing by the receptor-independent pathway, in non-hepatic cells, is regulated by LDL-c plasma concentration, however, even at low LDL-c concentrations LDLRs are saturated[32,35]. When LDL-c plasma concentrations are high, receptor-independent uptake of LDL is greater than the amount of LDL catabolised by the receptor-dependant pathway in the liver and LDLs accumulate in non-hepatic cells.[32,35] This is clinically important as the deposited LDL-c can become oxidised and atherosclerosis can be initiated as explained in **Chapter 1.2**.

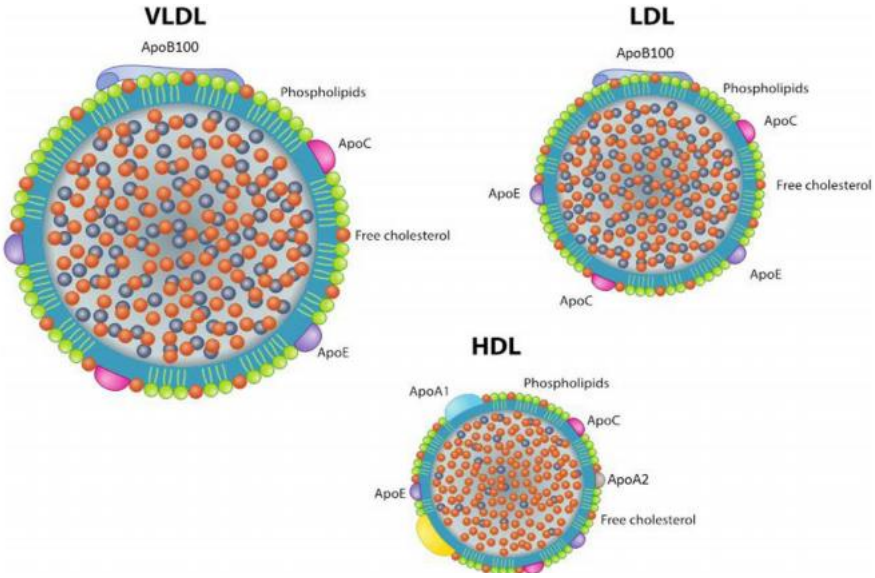


Figure 1.4: A representative image of the lipid content and surface protein makeup of very low-density lipoprotein (VLDL), low-density lipoprotein (LDL), and high-density lipoprotein (HDL). **Adapted from**: Helkin, A. *et al.* Dyslipidaemia Part 1—Review of Lipid Metabolism and Vascular Cell Physiology. *Vasc. Endovascular Surg.* **50,** 107–118 (2016).

1.4.1.3) Reverse cholesterol transport and HDL-c synthesis

HDL particles are lipoproteins which regulate plasma cholesterol and TG levels by mediating the transport of cholesterols from peripheral tissues back to the liver in a process called reverse cholesterol transport[28,36]. HDL particles are the smallest forms of lipoproteins and have a similar structural composition to VLDL and LDL particles (**Figure1.4**). The accumulation of cholesterol in non-hepatic peripheral tissues is toxic as it cannot be metabolised, thus, removal of cholesterol via HDL particles in the reverse cholesterol transport pathway contributes to the homeostasis of cholesterol[28]. HDL particles activate a number of anti-oxidant and anti-inflammatory pathways in endothelial cells and macrophages[32]. Thus, in the context of CVD, HDL is often referred to as "good" cholesterol.

APOAI is the major protein constituent of HDL, and is produced by enterocytes and hepatocytes[37]. It is in these cells that free cholesterol is added to APOAI particles to make nascent HDL-c via adenosine triphosphate binding cassette A1 (ABCA1)[32,38]. The mechanism of the interaction between APOAI and ABCA1 is not yet fully understood, and an accepted model for the mechanism for the creation of nascent HDL-c through this interaction does not exist, though several models have been proposed[28,39]. Nascent HDL-c particles exist in different particle sizes. HDL_2 and HDL_3 form the two major spherical subpopulations of circulating HDL. HDL_3 is smaller and denser, with fewer cholesterol esters and more proteins, while HDL_2 is larger with more cholesterol esters[28,40]. Nascent HDL-c particles in the plasma is modified by lecithin cholesteryl acyl transferase (LCAT)[41]. LCAT esterifies cholesterol present in HDL to form the more spherical and mature HDL particles[41].

In the indirect reverse cholesterol transport pathway, HDL cholesterol esters are exchanged for TG with VLDL and LDL particles, which is then followed by the uptake of LDL by LDLR in hepatocytes[28]. The phospholipid transfer protein (PLTP) mediates the transferal of phospholipids between VLDL and HDL, while also remodelling HDL particles to smaller and larger sizes[28,42]. Lipids can also be transferred between the cell membranes of hepatocytes, as well as other cholesterol utilizing tissues, and HDL via the HDL scavenger receptor protein BI (SRB1) in a bidirectional manner[28,32,43].

1.4.2) Dyslipidaemia and CVD risk

Dyslipidaemia has consistently been shown to be a major risk factor for the development of CVDs. Elevated TC and LDL-c, a combination which can lead to atherosclerosis, were

shown to increase the risk of ischaemic heart disease regardless of patient age or blood pressure levels in The Prospective Study Collaboration[23]. Additionally, the WHO estimates that raised cholesterol levels results in approximately 2.6 million deaths globally and is responsible for a third of reported ischaemic heart disease cases[10]. Further evidence of the critical role of elevated LDL-c levels in the development of CVD can be found in the analysis of LDL-c lowering therapies. Statins, or 3-hydroxy-3-methylglutarylcoenzyme A (HMG-CoA) reductase inhibitors, have a well-documented effect on lipid profiles, especially with prolonged use. LDL-c is the primary target of statin therapy. Statins target hepatocytes and inhibit HMG-CoA reductase by competitively binding to the active site, altering protein conformation[44]. Mevalonate, the product of HMG-CoA reductase activity, has a variety of cellular functions. Thus, statin therapy leads to inhibition of cholesterol biosynthesis, less secretion of lipoproteins, and increased rates of LDL uptake and degradation[44,45]. Meta-analysis of randomized trials analysing statin therapy suggest that the frequency of annual major vascular events can be reduced by approximately a fifth for each 1 mmol/L reduction in LDL-c[24]. In another review of the efficacy of statin therapy, it is reported that lowering LDL-c by 2 mmol/L in 10 000 patients with statin therapy over a 5 year period would prevent the onset of major CVD events in about 1000 patients with pre-existing CVD, as well as in 500 patients who are at high risk but who have not yet had a CVD event[46].

The meta-analysis also found that a similar risk reduction could be achieved in patients with significantly elevated HDL-c levels[24]. HDL-c plays a protective role in the development of CVD through its action in the reverse cholesterol transport pathway, but also through its reported anti-inflammatory, anti-oxidative and anti-apoptotic effects[47]. Strong evidence from epidemiological studies, as we as the known physiological role of HDL, has contributed to the long held notion that raising HDL-c through pharmacological therapies should lead to decreased risk of CVD[48]. However, large scale clinical trials of three cholesteryl ester transfer protein (CETP) inhibitors, which raise HDL-c levels, failed to produce reduced risk of CVD and were terminated prematurey[49–51]. It has also been reported, however that under certain circumstances, which are related to inflammation and oxidation, HDL-c can lose its protective effects, which may be a possible explanation for the trial's negative outcomes[47,48].

The role of elevated TG as a risk factor for atherosclerosis and ischemic heart disease is not as clear and is still under debate. Individuals with high TGs are often found with low HDL-c levels, and high LDL-c levels[52]. As such, it is difficult to distinguish between the effect of high TG concentrations on CVD from those of low HDL-c/high LDL-c. In one study,

seemingly positive associations found between fasting plasma TG and increased CVD risk were not repeated after adjusting for other risk factors, such as HDL-c levels[25]. Data from the Copenhagen City Heart Study showed a stepwise association between elevated non-fasting TG levels and increased myocardial infarction risk, most notably in women[53]. Large scale Mendelian randomization genetic studies provide further evidence that supports the hypothesis that increased TG levels is independently associated with increased CVD risk[54,55]. Additionally, meta-analysis of a large number of studies found a positive association between elevated TG and risk of stroke[56]. Thus, there is conflicting evidence on the role of triglycerides, but the consensus appears to be that significantly elevated triglyceride levels are detrimental to CVD outcomes.

1.4.3) The influence of environmental risk factors on dyslipidaemia

A number of environmental factors have been shown to influence lipid profiles. These risk factors are almost always also CVD risk factors/ co-morbidities. A risk factor such as an unhealthy diet (defined as a high intake of dietary fats and sugars) may affect dyslipidaemia by causing increased endogenous LDL-c levels[57]. Dyslipidaemia is then, itself, a risk of CVD. But an unhealthy diet also affects other CVD risk factors, such as increased body fat and blood pressure. There is an intricate relationship between these risk factors and CVD co-morbidities (**Figure 1.2**). Some additional key environmental and lifestyle factors which have been shown to have significant effects on lipid profiles include body weight and physical activity, alcohol usage, and smoking[58–60].

1.4.4) The genetics of dyslipidaemia

Dyslipidaemia has been shown to have a strong genetic component. A paired twin study conducted by Heller *et. al.* showed that approximately 50 % of the interindividual variation in plasma lipid levels could be explained by genetic variations[61]. There has been a concerted effort over the past few decades, with varying success, to identify genes and genetic variants associated with altered lipid profiles[62]. Rare genetic variants, usually encoding protein loss-of-function, are often the sole cause of dyslipidaemias such as familial hypercholesterolemia[63].

Common genetic variants, which usually have a smaller effect on gene function but are involved in the polygenic modulation of plasma lipid levels, have been identified in

population-based association studies[64]. The most common method used in this approach has been large-scale genome wide association studies (GWAS). In this approach large number of unrelated individuals belonging to a specific, defined population group, are genotyped for millions of candidate SNPs across the entire human genome[62]. Meta-analysis of 46 lipid GWAS studies carried out by Teslovich *et.al.* reported on 95 GWAS significantly loci which were associated with altered lipid profiles in more than 100 000 individuals of European descent[64]. The number of identified loci was then increased to 157 in a subsequent GWAS meta-analysis of 188 578 individuals of European descent and 7898 Non-Europeans (African = 3263)[65]. Genes with loci that have been associated with abnormal lipid profiles include *LDLR, PCSK9, CETP, APOE, LPA, LCAT, LPL, APOB*[65].

The advent of cheaper next generation sequencing technology has allowed for rare variant association studies in dyslipidaemia. These studies have identified rare coding variants, with large effect sizes, in large numbers of unrelated individuals[62]. Peloso *et al.* carried out exome wide analysis to genotype more than 200 000 low frequency, rare coding variants in 42 208 European ancestry and 14 330 African ancestry individuals[66]. Through this method they were able to replicate associations in a large number of variants in both populations, as well as identify 4 novel loci which had significantly large effects on HDL-c or TG[66]. A study by Lange *et al.* carried out whole exome sequencing of 1153 European American and 854 African American individuals and identified low frequency variants in patatin like phospholipase domain containing 5 (*PNPLA5*), a gene which was not previously associated with altered LDL-c levels[67].

Genetic studies have been used to bolster evidence from observational epidemiological studies that link abnormal lipid profiles to CVD risk. A large-scale Mendelian randomization study used a genetic score consisting of 14 SNPs commonly associated with HDL-c, and tested this score in 12482 cases of myocardial infarction (MI) and 41331 controls[68]. They used a generated score made from 13 SNPs commonly associated with LDL-c which was then tested in the cohorts as a positive control. They expected, from observational epidemiological studies, that a 1-standard deviation (1-SD) increase in LDL-c due to genetic scores (≈1.94 mmol/l increase) would be associated with a 54 % increase in MI risk. What they found was that a 1-SD increase due to the genetic score of the 13 SNPs resulted in a 113 % increase in MI risk ($P=2\times10^{-10}$) (**Figure 1.5**)[62,68]. Thus, there is clear evidence that there is a direct cause between the genetics of LDL-c alteration and CVD.

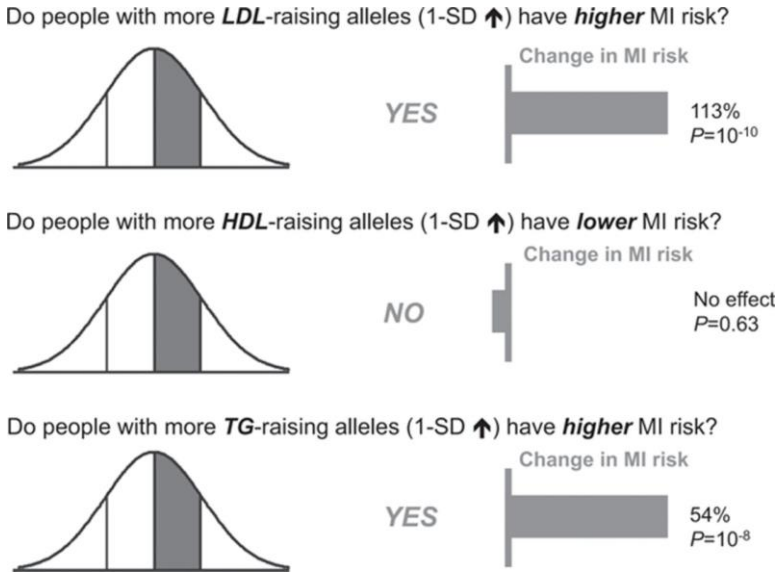


Figure 1.5: The cumulative effect of genetic variants that raise plasma lipid concentrations by 1-SD on MI risk. **Adapted from:** Musunuru, K. & Kathiresan, S. Surprises from Genetic Analyses of Lipid Risk Factors for Atherosclerosis. *Circ. Res.* **118,** 579–585 (2016).

The same study did not find any evidence of a similar causal role of HDL-c in CVD. A 1-SD (0.832 mmol/L) increase of HDL-c caused by the 14 SNPs was expected to confer a 38 % decreased risk of MI using data from observational epidemiological studies. They found that the genetic score conferred no significant change in MI risk (7 % decrease, *P*=0.63) (**Figure 1.5**)[62,68]. While the data does not rule out the possibility that there are biological mechanisms leading to increased plasma HDL-c that also protect against CVD, it does lend evidence to the clinical trials that found that pharmacologically raising HDL does not reduce CVD risk[62].

SNPs that have been associated with the modulation of plasma TG are pleiotropic, also being associated with other lipids. Of the 157 loci associated with lipid traits, 37 were associated with TG in some way, but only 16 were associated with TG only (**Figure 1.6**)[65]. While this pleiotropy makes interpreting data from Mendelian randomization studies difficult to do, it is not unexpected as TG are carried by multiple classes of lipoprotein particles and plasma TG concentrations are a reflection of the contribution of various physiological processes[62]. Multivariable Mendelian randomization was recently developed to try to separate the TG associated genetic effects from those of LDL-c and HDL-c on CVD risk[55]. The study

confirms the observations made by Peloso *et al.* that also show an increase in TG due to genetic effects, when adjusted for HDL-c and LDL-c, and an increase the risk of CVD in a manner similar to that of LDL-c (**Figure1.5**)[55,62].

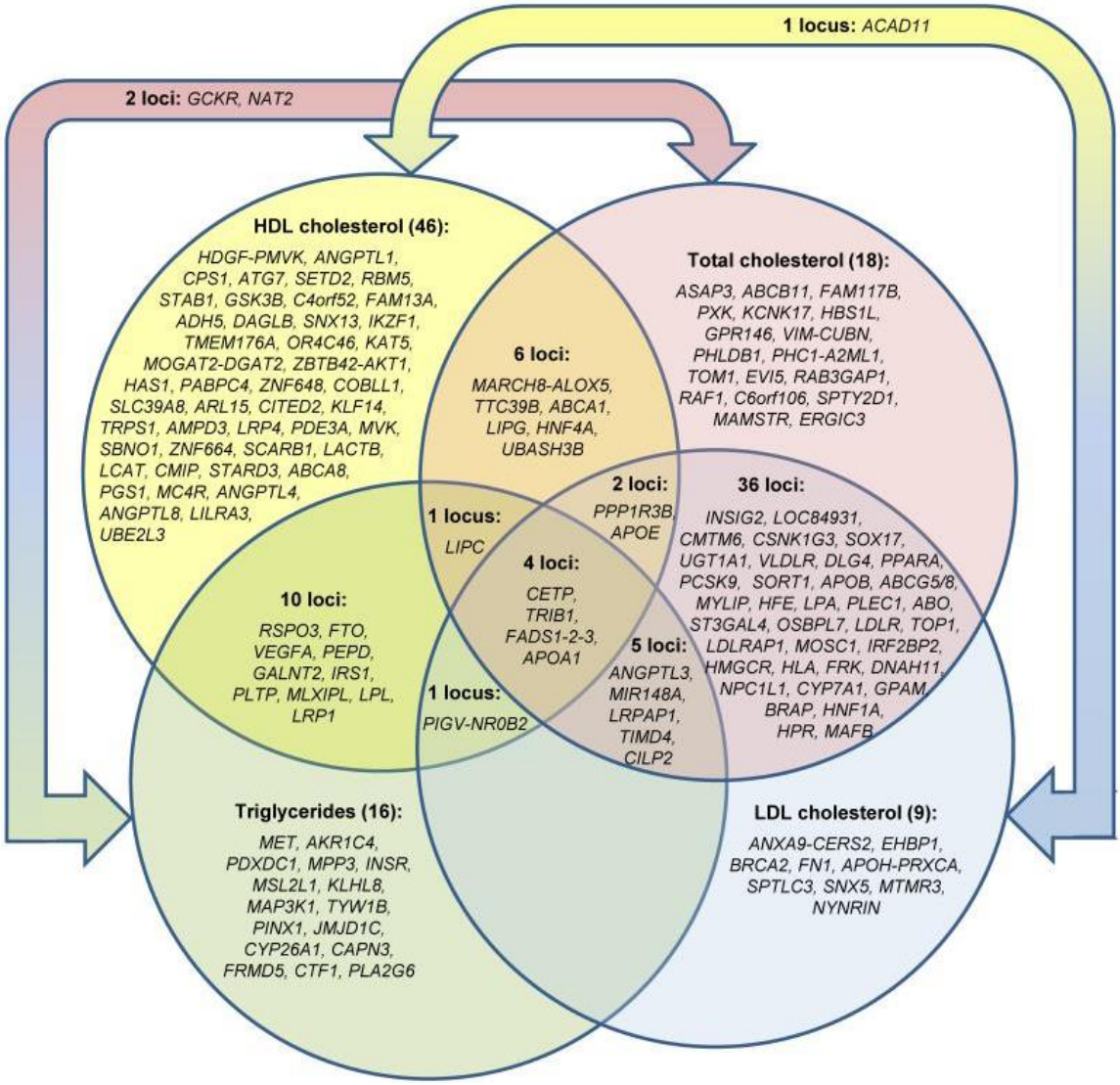


Figure 1.6: Venn diagram showing number of identified loci associated with multiple lipid traits. **Adapted from**: Willer, *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45,** 1274–1283 (2013).

Complex lipid metabolism phenotypes are often seen in the clinical setting; thus, it is important to take contribution of multiple loci into consideration in patients with less clearly defined dyslipidaemia disorders. There are a multitude of previously identified loci with varying degrees of reported association with altered plasma lipid profiles. In this study we look at genetic variation in three of these genes, *APOE*, *PCSK9*, and *CETP*, which have all

consistently been shown to be of great importance when looking at the genetics of dyslipidaemia in various population groups.

1.4.5) Apolipoprotein E (*APOE*)

The well-studied gene *APOE* is a key regulator of plasma lipid levels. *APOE* is located on chromosome 19q13.2, within an apolipoprotein gene cluster, and contains 4 exons. *APOE* is mainly expressed in the liver and encodes a 299 amino acid protein, with two independently folding protein domains[28,69]. APOE consists of an N-terminal domain of 22kDa (residues 1-191) and the 10kDa C-terminal domain (residues 222-299), with a short flexible hinge region connecting the two domains[70]. APOE functions by mediating the clearance of VLDL and CM remnants by acting as a ligand for cell surface receptors[28,69]. The N-terminal contains the LDLR binding site, whilst the C-terminal is responsible for the binding of lipids, and to LDLR, LRP1 heparan sulphate proteoglycans (HSPG).[71] APOE is a key component of various lipoproteins (**Figure 1.4**), thus, genetic variation in this gene is likely to be of concern in the context of aberrant lipid profiles.

The APOE protein primarily exists in three major isoforms named ε2, ε3 and ε4. ε3 is the most common isoform, being referred to as the 'wild type' in humans, and is characterised by the presence of cysteine and arginine at residues 112 and 158, respectively[69]. ε2 has cysteine at both sites, while ε4 has arginine at both sites[69]. The determination of APOE isoform can be achieved by genotyping for the *rs429358T>C* (ε4) and *rs7412C>T* (ε2) polymorphisms. Both ε3 and ε4 isoforms bind to LDLR with high affinity, but ε2 has severely defective LDLR binding and is associated with type III hyperlipoproteinemia[72]. Additionally, ε4 binds with better affinity to VLDL than ε3, while ε2 and ε3 are found to have a higher affinity for HDL[73]. ε4 expression was found to increase VLDL levels and decrease HDL-c levels relative to ε3 in APOE-null mice[74]. In contrast, ε2 was to be associated with lower LDL-c and higher HDL-c in children, and ε2 homozygotes were found to be associated with increased plasma TG and lower LDL in mice[75,76]. Further evidence has been shown in a comprehensive meta-analysis of 82 lipid studies and 121 studies of coronary outcomes where an approximately linear relationship between *APOE* genotypes (ε2/ ε2, ε2/ ε3, ε2/ ε4, ε3/ ε3, ε3/ ε4, ε4/ ε4), LDL-c and CVD risk has been seen[77]. They also reported a weak inverse relationship with *APOE* genotypes and HDL-c, as well as an increase in TG levels when comparing ε2/ ε2 individuals to ε3/ ε3 individuals[77]. *APOE* has been studied in a South African cohort before,

but as it is seen as a major predictor of altered lipid levels, it will be important to evaluate its effect on our specific cohort.

1.4.6) Pro-protein convertase subtilisin/kexin type 9 (*PCSK9*)

The *PCSK9* gene, located at 1p32.3, consists of 12 exons and encodes a 692 amino acid protein belonging to the mammalian serine proprotein convertase family[78]. This family of proteins is involved in the proteolytic maturation of secreted proteins such as cytokines, receptors and cell surface proteins[79]. PSCK9 consists of a signal sequence (residues 1–30), a prodomain (residues 31–152), a catalytic domain (residues 153-449), and a 243–amino acid cysteine rich and histidine-rich C-terminal region (**Figure 1.7**)[78,80]. *PCSK9* is primarily expressed in the liver, but is also found at lower levels of expression in the kidneys, intestine and brain[81]. PSCK9 has been shown to play a vital role in lipoprotein metabolism through its functional relationship with LDLR. LDL bound to LDLR is internalized in clatherin coated pits where it eventually undergoes lysosomal degradation. LDLR dissociates from the LDL and is then recycled back to the plasma membrane where it is available to bind more LDL, shuttling back and forth in a continuous process[69]. Circulating PCSK9, in sera, binds to the extracellular domain of LDLR resulting in the redistribution of LDLR from the cell surface to lysosomes where the LDLR undergoes degradation. This action also prevents the receptor from recycling to the surface[82–84]. Thus, high PCSK9 activity promotes increased levels of LDL-c and vice versa.

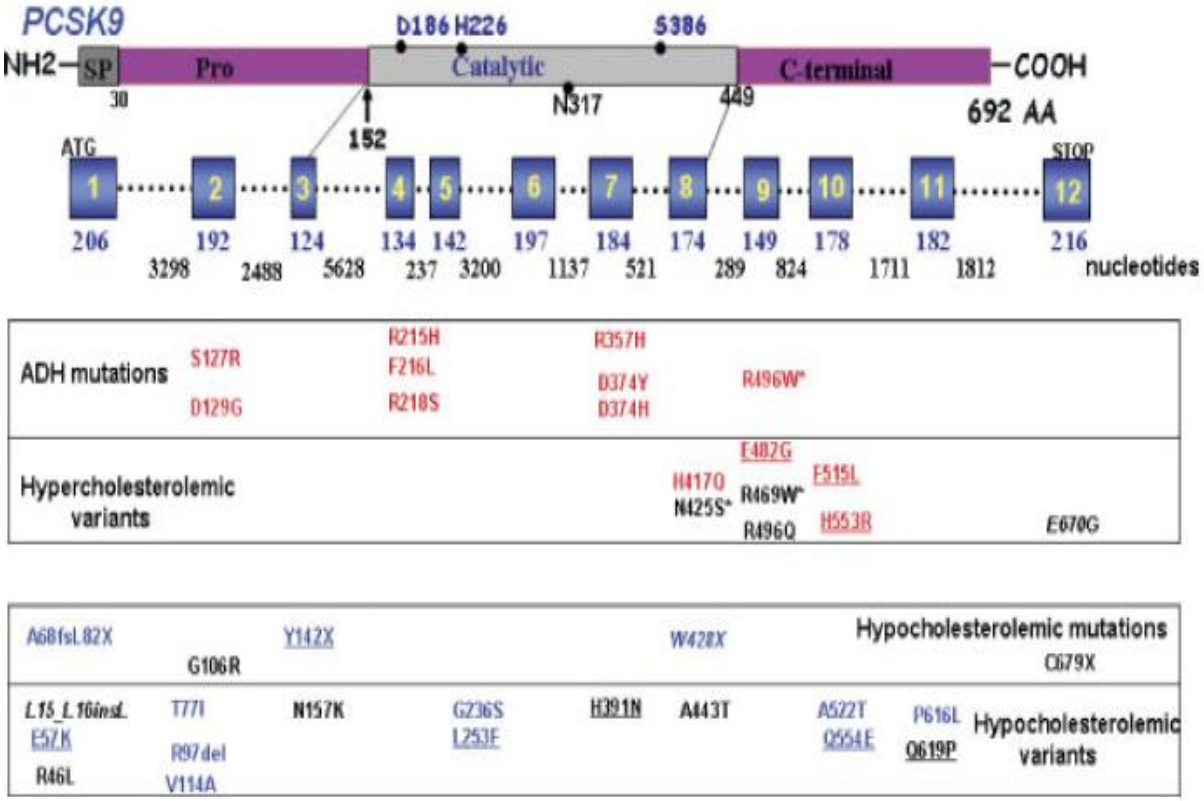


Figure 1.7: A representative image of the *PSCK9* gene structure, with corresponding protein domain, and relative positions of commonly identified genetic variants which are associated with increased or decreased LDL-c. Variants in red are found only in patients with high LDL-c. Variants in blue are found only in patients with low LDL-c. Variants found in only African populations are underlined. Variants found with high frequency in both African and Caucasian populations are in italic. **Adapted from**: Abifadel, M. *et al.* Mutations and polymorphisms in the proprotein convertase subtilisin kexin 9 (PCSK9) gene in cholesterol metabolism and disease. *Hum. Mutat.* **30,** 520–529 (2009).

To date a number of common gain-of-function mutations in *PCSK9* have been reported in several population groups, which were associated with increased LDL-c levels (**Figure 1.7**)[78]. These are good candidates to study in a South African cohort of patients with high LDL-c levels. The E670G (*PCSK9 rs505151A>G*) polymorphism in *PCSK9* has been associated with altered LDL-c levels in a number of studies. Chen, *et al.* reported that a haplotype group defined by the E670G polymorphism was significantly associated with elevated LDL-c, accounting for 3.5 % of plasma LDL-c variability[85]. This association was repeated by Evans and Beil, who reported an increase in LDL-c levels in men with the 670G variant, however, the association was not repeated in women[86]. These results have not been consistent. A study done in a Chinese cohort found that carriers of the 670G allele had significantly lower LDL-c levels than non-carriers but it was not found to be associated with CVD risk[87]. In order to

resolve this inconsistency, a meta-analysis was carried out by Cai, *et al.* on all the studies with published results to date[88]. Their analysis confirmed that the 670G allele was indeed associated with higher LDL-c levels in all reported population groups analysed both separately and combined[88].

Nonsense mutations in *PCSK9* were found to be common in African American individuals in the Atherosclerosis Risk in Communities study, with approximately 2.6 % of the African American cohort having at least one allele with a Nonsense mutation *PCSK9* (Y142X or C679X)[89]. These variants were found to reduce LDL-c levels by 28 % and also reduced the incidence of CVD by 88 %[89]. Additionally, C679X (*PCSK9 rs28362286C>A*) carriers were found to significantly reduce LDL-c levels by 36.9 % in African Americans in the Dallas Heart Study[90]. In Africans, the C679X variant was found at an incidence of 3.7 % in Zimbabwean women, and was associated with a decrease in LDL-c levels by 27 %[91]. The C679X truncation disrupts the folding of the C-terminal domain, which is thought to be important for the successful secretion of the protein[80]. PCSK9 protein with the C679X truncation is efficiently expressed and processed in human liver cells and retains its catalytic activity, but is localised in the ER and is not secreted into the cell medium[92]. As such, the prevalence and role of both gain-of- and loss-of-function mutations in *PCSK9* should be explored in our cohort in order to understand the nature of *PCSK9* genetic variation in the South African population.

1.4.7) Cholesteryl ester transfer protein (*CETP)*

CETP is located at chromosome 16q12-21 and spans 25 kb, encoding a 476 amino acid protein, and is expressed predominantly in the liver, spleen and adipose tissue[40,93]. The CETP protein plays an essential role in the reverse cholesterol transport pathway by mediating the exchange of cholesteryl esters for TGs, from HDL to APOB-rich LDL and VLDL particles[40,93]. Increased CETP activity contributes to adverse plasma lipoprotein profiles by lowering HDL-c levels and increasing LDL-c levels[40,93,94]. CETP activity was shown to increase the total cholesterol content in the liver, while lowering the levels of LDLR mRNA and protein, in CETP transgenic mice studies[95].

A number of common and rare *CETP* variants have been associated with altered lipid levels, and CETP activity through sequencing efforts, GWAS and Mendelian randomization studies[94,96,97]. However, the association of many of these SNPs with lipids and CVD, has not

been consistent. The vast majority of research on *CETP* genetic variation and its effects on lipids has been done in European and Asian populations. Studies conducted in Japanese cohorts failed to detect CETP mass and activity in individuals who were homozygous for the 5′ splice donor site of intron 14 (Int14+1G→A) or Q309X nonsense mutations and these variants were also significantly associated with increased plasma HDL-c levels[98–100].

One of the most common *CETP* polymorphisms that has been widely studied is the *TaqIB* (*CETP rs708272G>A*) variant. *CETP rs708272G>A* is a silent SNP affecting the 277th nucleotide of the intron 1[69]. The *TaqIB* B2 allele (*rs708272A* allele) has repeatedly been associated with increased HDL-c[101,102]. Meta-analysis carried out by Boekholdt and Thompson, in over 10 000 individuals, reported that B2/B2 homozygotes had lower CETP activity and an average increase in HDL-c of 12mmol/L when compared to B1/B1 homozygotes[96]. These results are supported by an even larger meta-analysis of 47 studies, which investigated 23,928 cases and 27,068 controls, where it was reported that the B2 allele was weakly associated with reduced CVD risk[103]. Further analysis of the B2 allele's role in HDL-c modulation, in 3,600 cases and 5,929 controls, found an increase of roughly 0.25 mmol/L in HDL-c in a Caucasian cohort[103]. However, as *CETP rs708272G>A* is a synonymous mutation with no predicted effect on CETP function and no predicted effects on regulation or splicing site alteration, it is hypothesised that this association is caused by another SNP in strong linkage disequilibrium (LD) with *CETP rs708272G>A.*

One of the proposed candidate SNPs is *CETP rs1800775C>A* which lies in the *CETP* promoter region and has been found in strong LD with *CETP rs708272G>A* in Caucasians. The *CETP rs1800775A* allele was associated with increased HDL-c levels and was also found to reduce CETP protein mass and *CETP* promoter activity by 28 % and 26 %, respectively[104]. However, this variation may not be the causative allele in the *CETP rs708272G>A* haplogroup as demonstrated in a recent study by Pirim *et al.* which re-sequenced the *CETP* gene and flanking regions in Caucasians and Africans with extreme plasma HDL-c levels[97]. They reported on 12 SNPs which were associated with altered HDL-c in both the Caucasian and African cohorts, including *CETP rs708272G>A* and *CETP rs1800775C>A*[97]. However, the *CETP rs708272G>A* and *CETP rs1800775C>A* variants were not in strong LD in the African cohort ($r^2 = 0.19$)[97]. This suggests that there could be other additional functional genetic variants, other than *CETP rs1800775C>A,* that are in LD with *CETP rs708272G>A,* and which are responsible for its reported association to increased HDL-c. They also report on large proportion of HDL-c altering *CETP* haplotypes that are

located in the 5'region, the promoter, and the first 12kb of *CETP*, making it a prime region to investigate in our cohort[97]. The relationship between *CETP*, dyslipidaemia, and CVD requires further elucidation in more population groups in order determine the true role of *CETP* genetic variation on aberrant lipid profiles.

1.5) Diabetic dyslipidaemia (DD)

Type 2 diabetes mellitus (T2DM) is a complex metabolic disorder which is characterized by chronic hyperglycaemia and unfavourable alterations to carbohydrate, fat and protein metabolism[105]. These metabolic irregularities are caused by defects in insulin secretion, and/or insulin activity[105]. T2DM is especially important in the context of CVD as T2DM patients are at significantly increased risk of developing atherosclerotic CVD[106]. A population-based retrospective cohort study investigating CVD events in 379,003 adults with T2DM and in 9,018,082 adults without T2DM, found that the T2DM patients typically developed more severe CVD approximately 14.6 years sooner than people without T2DM[107]. CVD is the leading cause of mortality in T2DM patients with about two thirds of T2DM deaths being attributed to CVD, of which approximately 40 % is due to ischemic heart disease[106,108]. Thus, the incidence of T2DM compounds the effect of other CVD risk factors such as race, hypertension, smoking, diet and obesity.

Dyslipidaemia is very common in T2DM patients, with an estimated 60-70% presenting with abnormal lipid profiles[109]. Both glucose and lipids play important interrelated roles in energy metabolism, with both groups being regulated in the liver. DD is characterized by elevated TG and LDL-c levels, and lowered HDL levels[110]. It has been found that LDL particles are more atherogenic in T2DM patients, regardless of LDL-c concentrations[111]. The molecular mechanisms of this phenomenon are not completely understood.

Insulin is a key regulator of the metabolism of VLDL, which is the main carrier of TG particles[110]. Insulin mediates the transferal of TG to APOB and is a regulator of lipoprotein lipase (LPL) activity, which is responsible for the delipidation of VLDL[110]. LPL activity can be suppressed by circulating free fatty acids. In individuals with insulin resistance, free fatty acids are elevated, APOB degradation is decreased leading to overproduction of VLDL, LPL activity is impaired and VLDL hepatic uptake is decreased. This results in increased TG (Hypertriglyceridemia)[106,110]. Hypertriglyceridemia stimulates CETP activity and leads to lipoproteins with high TG content. HDL with enriched TG content is more susceptible to

catabolism, lowering HDL levels in T2DM patients, while LDL particles are hydrolysed and are subsequently smaller and denser[106,110]. Thus, it is important to investigate factors which could compound the already increased risk of CVD in T2DM patients, such as genetic variations which cause dyslipidaemia.

1.6) The importance of studying the genetics of dyslipidaemia in black South African T2DM patients

Historically, there have been low rates of CVD and T2DM reported in individuals of African descent, which has been attributed to low levels of TC, LDL-c, and TGs and high levels of HDL-c found in these populations[112,113]. However, as lifestyle factors have changed, so too have the CVD risk profiles of black South Africans. Results published by the Heart of Soweto Study showed that of the 1832 black South Africans observed, 39 % had raised cholesterol profiles, 63 % had low HDL-c, 44 % had high LDL-c and 23 % had high TG levels[114]. Approximately 25% of South African adults are obese (BMI $\geq$30 kg/m^2), with women being the worst affected, and the rate of childhood obesity rapidly rising[1,17,115]. The International Diabetes Federation (IDF) estimates that the global incidence of DM will increase by 55 % by 2040[116]. The greatest increase is expected in Africa where impaired glucose tolerance is estimated to rise by 126.4 % in Sub-Saharan Africa[116]. An estimated 2.286 million South African adults are living with DM, a prevalence rate of 7 %[105]. The incidence of T2DM in the urban Xhosa population has increased from 7 % in 1990, to 11.7 % in 2009[117]. CVD co-morbidities and associated risk factors are also on the rise.

Therefore, it is becoming increasingly important to determine the prevalence and impact of population specific genetic variations in black South Africans, which may predispose them to developing dyslipidaemia as their lifestyles and diets change. These variations can be used as biomarkers to predict future dyslipidaemia. This is even more pertinent in T2DM patients who are already at a higher risk of dyslipidaemia and CVD, and in whom the changes to lipid profiles caused by genetic predisposition could have an even bigger impact on disease outcomes. Complex lipid metabolism phenotypes are often seen in the clinical setting; thus, it is important to take the contribution of multiple loci into consideration in patients with less clearly defined dyslipidaemia disorders. The impact of the above-mentioned genetic variations has been well characterized in individuals of European and Asian descent, but little work has been done on the indigenous black South African population for the majority of

these genes. There is insufficient data on its role in African populations, where the burden of dyslipidaemia and its related co-morbidities is rapidly increasing. Thus, we carried out genetic research to identify population-specific genetic markers that are associated with altered lipid metabolism in indigenous black South African patients with T2DM. It is crucial to understand the aetiology of dyslipidaemia, in a setting where the burden of CVD, obesity and diabetes is rapidly increasing, to apply appropriate management and preventative strategies.

1.7) Aims and Objectives

This study aims to characterize the impact of genetic variations in *APOE*, *PCSK9* and *CETP* on altered lipid metabolism in an indigenous black South African population with T2DM. Ultimately, we aim to identify a panel of relevant genetic variations that predispose black South Africans to dyslipidaemia and may be used as biomarkers to inform and apply appropriate management and preventative strategies.

Primary end-points:

Total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol and triglyceride concentrations (mmol/L) taken from the blood of T2DM patients

Secondary end-points:

HbA1c levels, correlation of genetic markers and phenotypes

The study aim was achieved through completion of a number of objectives:

- Recruitment of participants presenting for diabetes screening or treatment at the Chris Hani Baragwanath Academic Hospital
- Determination of patient lipid profiles
- Evaluation of patient's medical records for demographic and clinical information related to diabetes
- Evaluating the contribution of additional clinical variables on altered lipid profiles
- Determination of the prevalence of known *APOE*, *PCSK9* and *CETP* variations in black South Africans with T2DM by genotyping methods

- Investigating the contribution of these variations to altered TC, HDL-c, LDL-c, TG and HbA1c levels by genotype-phenotype associations using univariate and multivariate analysis
- Identifying potential biomarkers for dyslipidaemia in the black T2DM cohort.

Chapter 2: Methodology

2.1) Study design

A quantitative cross-sectional study design was used for our approach to the research question. This allowed us to clearly define the pathogenicity of our selected genetic variants in our study population, as well as allowing us to better explain some of the reported heterogeneity of these genes.

2.2) Research site and resources

Recruitment and determination of clinical parameters for this study was completed by Dr. Donald Tanyanyiwa at the Diabetes Clinic at Chris Hani Baragwanath Academic Hospital as part of a larger project on dyslipidaemia in diabetic South Africans. Medical information obtained at the Chris Hani Baragwanath Academic Hospital is kept confidential and curated according to their own protocols. Collected whole blood and DNA samples were stored in freezers at the University of Cape Town according to ethics clearance (**HREC REF 231/2010**), which allows for the storage biological samples for research. Biological samples were given a lab code number and anonymised. All further downstream analysis, including genotyping, was carried out at the premises of the University of Cape Town's division of Human Genetics, or relevant authorised service providers.

Clinical platforms for measuring lipid profiles and monitoring blood glucose levels, as well as medical staff who carried out blood sample collection, were made available by the Chris Hani Baragwaneth Academic Hospital Diabetes clinic in Johannesburg. These resources were coordinated by Dr. Donald Tanyanyiwa. Genotyping platforms and relevant research supplies and equipment were available through Professor Collet Dandara at the University of Cape Town. Data analysis and statistical analysis was performed by researchers involved in the study at the University of Cape Town. Funding for the study was provided by the National Research Foundation (NRF) and the University of Cape Town through grants awarded to Professor Dandara. Funding of students was also awarded by the NRF and the University of Cape Town.

2.3) Ethics clearance

This project forms part of a broader study on dyslipidaemia which has approved ethical clearance (**HREC REF 089/2013**) from the University of Cape Town Health Sciences Ethics Committee. Ethical clearance for our linked sub-study was approved for ethical clearance (**HREC REF 750/2016**) from the University of Cape Town Health Sciences Ethics Committee (**Appendix VII**).

2.4) Recruitment and Enrolment of Study Cohort

Recruitment was carried out in the following manner; Patients were invited to participate in the study through an information package that explained the aims of the study (**Appendix VI**). The information package was explained by the recruitment team in a language that the patient could easily understand. The patients were asked to sign consent forms allowing for the extraction of 5 ml whole blood samples for DNA extraction for use in genomic analysis. Patients were asked to consent to the researchers accessing their medical records for information relating to dyslipidaemia and medication use. The patients were asked a series of additional questions relevant to the project at the time of recruitment, such as their demographic information, family history and lifestyle choices. To date approximately 1000 participants have been recruited from which the study case cohort was drawn.

2.5) Study Population

Our study population consisted solely of individuals from the indigenous Bantu black South African ethnicities. We focused on this population group in particular as there is little data on the exact role of genetics in dyslipidaemia for black South Africans. The few genetic studies available, which include black South African populations, have relatively small sample sizes and focus on a few commonly reported variants in clinically significant genes. We aimed to greatly improve on the sample size and number of variants investigated, as the need for comprehensive data on the effects of genetics on non-communicable disease becomes increasingly relevant in the African context.

A number of the pathogenic genetic variants, published in Caucasian populations, are present at low frequencies (>1 %) in individuals of African descent. Comprehensive screening of our genes of interest in Caucasian and African American cohorts has revealed that a multitude of

variants are specific to each group. There is large genetic variability in Africans, thus, it is probable that there are unidentified novel variants in our population that may be pertinent in understanding the rise in dyslipidaemia and its comorbidities in South Africa.

2.6) Inclusion and exclusion criteria

The study included a cohort of dyslipidaemic patients who were attending the diabetes clinic at Chris Hani Baragwanath Academic Hospital. All of the patients included in this study have been diagnosed as having T2DM by a physician according to the Society for Endocrinology, Metabolism and Diabetes of South Africa 2017 guidelines for the management of T2DM[105]. Patients from the larger sample group of approximately 1000 individuals were stratified by lipid levels into 4 distinct groups which were included in this study's cohort. Group 1 (high TC), group 2 (high TG), group 3 (mixed dyslipidaemia) and group 4 (normal lipid levels). Lipid levels were defined to be dyslipidaemic based on the European Atherosclerosis Society (EAS) guidelines for non-diabetic individuals[52]. Abnormal levels were, thus, defined as follows:

- TC > 5.2 mmol/L,
- LDL-c > 4.1 mmol/L,
- Triglycerides > 2.3 mmol/L,
- HDL-c < 1 mmol/L.

Selection of samples to include by stratification by lipid profile into the 4 groups was done to ensure that a there was a wide range of lipid traits present in the cohort. The samples were combined into one case cohort during analysis. Thus, potential alterations to lipid levels caused by genetic variations were likely to be teased out.

The exclusion criteria were a) all patients with known primary causes of dyslipidaemia, and b) those with non-diabetic related secondary cause of dyslipidaemia. These included thyroid disorders, treatment with antiretroviral drugs protease inhibitors associated with dyslipidaemia, and liver failure.

2.7) Study cohort sample size

The lowest frequency of 2% for a SNP was be used for sample size determination with the method outlined by Naing, *et al.*[118]. We wish to estimate the population proportions of black South Africans with the abnormal lipid levels (high TC, high TG, mixed dyslipidaemia and normal levels). To obtain a 95 % confidence interval for each of these parameters with a width of 2Δ and a sample proportion of *p*, we require a sample size of

$$n = \left(\frac{1.96}{\Delta}\right)^2 p(1-p)$$

If we assume that 5% of the sampled patients will have abnormal lipid levels in each group and we want a 95% confidence interval of width 0.05 (i.e. $\Delta = 0.05$), then the required sample size is $n = (1.96/0.05)^2\ 0.05\ (1 - 0.05) = 73$ (rounding up). Thus, the study needed 73 participants in each of the four groups: group 1 (high TC), and group 2 (high TG), group 3 (mixed dyslipidaemia) and group 4 (normal levels) which is 73×4 =292.

After selection of relevant samples based on inclusion criteria, adding at least 100 samples from each group for added power, we ended up with a **final sample size** of **417** participants.

2.8) Lipid Profile Determination

The lipid profiles of each participant were measured using the Roche/Hitachi **cobas c** systems (Roche, Mannheim, Germany) according to the manufacturer's instructions, at the Chris Hani Baragwanath Academic Hospital laboratories. A brief overview of the enzymatic methods employed by the system is summarised here.

2.8.1) Total Cholesterol determination

The CHOL2 kit uses an *in vitro* enzymatic/colorimetric test to quantitively determine TC levels in human serum and plasma. Cholesterol esters are cleaved by cholesterol esterase (CE) activity yielding fatty acids and free cholesterol. Cholesterols are, in turn, oxidised to form cholest-4-en-3-one and hydrogen peroxide by the action of cholesterol oxidase (CHOD). The hydrogen peroxide formed effects the oxidative coupling of phenol and 4-aminophenazone to form a red quinone-imine dye, in the presence of peroxidase (POD).

The colour intensity of the red quinone-imine dye is determined by measuring an increase in absorbance and is directly proportional to TC concentration.

2.8.2) HDL-c determination

The Roche direct-HDLC3 assay is an *in vitro* homogeneous enzymatic colorimetric diagnostic test for determining human serum and plasma HDL-c concentrations. LDL, VLDL, and chylomicrons form water-soluble complexes with dextran sulphate in the presence of magnesium ions. These complexes are resistant to polyethylene glycol (PEG)-modified enzymes, which are have selectively catalytic activities to lipoproteins in the following order LDL < VLDL ≈ chylomicrons < HDL[119].

HDL-c esters are cleaved by PEG modified-CE activity yielding fatty acids and free HDL-c. HDL-c's are, in turn, oxidised to form Δ4-cholestenone and hydrogen peroxide by the action of PEG-modified CHOD. The generated hydrogen peroxide reacts with 4-amino-antipyrine and Sodium N-(2-hydroxy-3-sulfopropyl) -3,5-dimethoxyaniline (HSDA) to form a purple-blue dye. The colour intensity of this dye is measured photometrically and is also directly proportional to HDL-c concentration.

2.8.3) Triglyceride determination

The Roche direct-TRGL assay is an *in vitro* homogeneous enzymatic colorimetric diagnostic test that was used to determine human serum and plasma TG concentrations. TG are hydrolysed to glycerol through the action of LPL. This is followed by oxidation of glycerol to dihydroxyacetone phosphate and, subsequently, hydrogen peroxide through the catalytic actions of glycerokinase (GK) and glycerol phosphate oxidase (GPO) respectively. POD catalyses the reaction of 4-aminophenazone and 4-chlorophenol to form 4-(p-benzoquinone-monoimino) -phenazone, which is a red dyestuff that can be measured photometrically, the intensity of which is directly proportional to TG concentration.

2.8.4) LDL-c determination

LDL-c levels were estimated using the Friedewald equation. This method for the indirect estimation of LDL-c concentrations requires the direct measurement of TC, TG and HDL-c. In this method the assumption that the TC is made up of LDL-c, HDL-c and VLDL-c. Thus,

in order to estimate LDL-c concentrations**: LDL = TC - HDL - TG/5.0 (mg/dL),** where TG/5 is an estimation of VLDL-c[120].

The assumption of VLDL:TG = 1:5 is convenient as most of circulating TG is carried in the VLDL fraction, however, the method does not take into account the contribution of CMs. In most plasma samples CMs are not detectable, thus, it is has been proven to be a reliable method to use[121]. The Friedewald equation is not without limitations. In cases of dysbetalipoproteinemia (Type Ill hyperlipoproteinemia), VLDL-c concentrations are underestimated, and LDL-c concentrations are overestimated. Additionally, in cases where TG ≥ 4.5 mmol/l the equation overestimates VLDL-c concentrations and underestimates LDL-c concentrations[121]. Downstream analyses of LDL-c associations were excluded for patients with TG ≥ 4.5 mmol/l in our study.

2.9) Haemoglobin A1C (HbA1c)

Haemoglobin (Hb) is a protein located in the erythrocytes and consists of four subunits containing a haem moiety. Hb functions by mediating the transport of oxygen and carbon dioxide in the blood. Hb is heterogeneous, with 90% of Hb in human erythrocytes being HbA[122]. There are a number of other Hb derivatives, including HbA1c which is a post translational modification of HbA. HbA1c is formed by a non-enzymatic reaction of glucose to the N-terminal amino group of the β-chain of normal adult HbA[122]. HbA1c% as a fraction of Hb levels has been shown to be correlated with response to glucose tolerance tests and fasting blood sugar levels[123,124]. Stable HbA1c formation is limited to approximately 100-120 days, the lifespan of erythrocytes, thus, HbA1c% is a reflection of the average blood glucose level of the previous 2 to 3 months and is a useful parameter for monitoring long-term blood glucose control in T2DM patients[122]. HbA1c% has also been shown to be a predictor of diabetic complications in T2DM patients, with T2DM patients having a 2-3 fold increase in HbA1c[123,125,126].

2.9.1) HbA1c% determination

HbA1c levels were measured using the Tina-quant Haemoglobin A1c Gen.2 whole blood application on the Roche/Hitachi **cobas c** systems (Roche, Mannheim, Germany) according

the manufactures instructions, at the Chris Hani Baragwanath Academic Hospital laboratories.

The turbidimetric inhibition immunoassay (TINIA) for haemolysed whole blood is the bases for HbA1c determination. HbA1c reacts with anti-HbA1c antibodies and forms soluble complexes. Polyhaptens are added to the solutions, which react with the anti-HbA1c antibodies. This results in the formation of insoluble antibody-polyhapten complexes which can be measured turbidimetrically. Hb in the samples is converted to a derivative with a characteristic absorption spectrum which is then measured bichromatically. The HbA1c% calculated from the HbA1c/Hb ratio:

HbA1c (%) = (HbA1c/Hb) × 91.5 + 2.15

2.10) DNA isolation

DNA was extracted from whole blood samples using a modified version of the salting out protocol adapted from Gustafson *et al.*[127]. Whole blood samples (5 mL), frozen in storage, were thawed at room temperature and diluted with 2 X volume of phosphate buffered saline (PBS) (**Appendix I**). The PBS solution ensures a constant pH which preserves the white blood cells during maceration. The samples were then centrifuged at 2200 g for 10 minutes, after which the supernatant was carefully discarded, and the pellet re-suspended in Sucrose Triton X-100 lysis buffer (10 mL) (**Appendix I**) and vortexed. The Tris-HCl in solution maintains a favourable pH of 8, the magnesium chloride ($MgCl_2$) provides Mg^{2+} ions which are co-factors for the enzymes in the solution, the sucrose increases osmolarity of the cells which results in the intake of water aiding in cell lysis, and Triton X-100 is a non-ionic surfactant which weakens cell membranes while not denaturing proteins. The tubes were then placed on ice for 5 minutes, allowing the completion of cell lysis and hindering the action of DNA cleaving enzymes in solution. The tubes were again centrifuged at 2200 g for 10 minutes, and the supernatant was discarded. The pellet was re-suspended in 0.6 X volume of blood sample T20E5 (3 mL) containing ethylenediaminetetraacetic acid (EDTA) (**Appendix I**), which captures the divalent cations needed by DNase enzymes, inhibiting DNA degradation. 200 uL of 10 % sodium dodecyl sulphate (SDS) (**Appendix I)**, which breaks down membranes, and 200 ug/mL Proteinase K was added to degrade any remaining proteins in solution. The samples were then incubated at 45°C overnight in a water bath.

Saturated sodium chloride (NaCl) (1 mL) was added to each sample the following day, followed by vigorous mixing. Increased NaCl concentration results in some of the H_2O molecules being attracted by Na^+ ions, which do not interact with hydrophilic residues of proteins. The result is that protein: protein interactions are stronger than protein: water interactions, causing the proteins to coagulate. The tubes were then centrifuged at 2400 g for 30 minutes, pelleting out the protein fragments and leaving DNA in solution. The DNA containing supernatant was transferred to a new tube, and 2X volumes of room temperature 100 % ethanol was added. DNA is not soluble in ethanol; thus, the addition of 100 % ethanol allows for the precipitation of DNA out of the solution. The tubes were gently agitated, and the precipitated DNA was spooled and transferred to an Eppendorf tube, and washed in ice cold 70 % ethanol (1 mL) to remove any remaining salts. The DNA was then air dried and re-suspended in 1 X Tris-HCl, EDTA (TE) buffer for storage.

2.11) Determination of DNA yield, purity, and integrity

Quality control of the extracted DNA samples was carried out to ensure the success of subsequent downstream experiments. The methods used are outlined below. An example of DNA yield and purity, as well as DNA integrity can be found in **Appendix II.**

2.11.1) DNA yield and purity quantification

The yield and purity of extracted DNA was quantified by spectrophotometry using a NanoDrop™ 1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, USA).

Sample concentrations were calculated for each sample from optical density (OD) according to the following equation:

A = -log (Intensity sample/Intensity blank)

A is absorbance is measured at 260nm. A modified version of the Beer-Lambert equation is then employed to correlate absorbance to concentration:

c = (A * e)/b

Where **c** is the nucleic acid concentration in ng/μL, **A** is the absorbance in AU, **e** is the wavelength-dependent extinction coefficient in ng-cm/μL (50 ng-cm/μL for double stranded DNA) and **b** is the pathlength in cm.

DNA purity was determined by measuring the A_{260} nm/A_{280} nm ratio. A DNA sample with a A_{260} nm/A_{280} nm ratio of greater than 1.8 is generally accepted as high purity. Contaminants that absorb at 280nm, such as proteins, will lower the ratio. The A_{260} nm/A_{230} nm ratio is another parameter used to determine purity and is used to check for contaminants such as phenol, carbohydrates and EDTA which absorb at 230 nm. A DNA sample with A_{260} nm/A_{230} nm ratio of between 1.8 and 2.2 is considered pure.

2.11.2) Determination of DNA integrity using agarose gel electrophoresis

DNA samples were evaluated for potential degradation using agarose gel electrophoresis. In this method, DNA fragments are visualised and separated based on size in an agarose gel. Powdered agarose (Seakem® LE Agarose, Lonza, Rockland, USA) is added to 1 X Tris-Borate EDTA (TBE) buffer at a concentration of 1 % (w/v). The solution is mixed, superheated in a microwave until the solution is clear and all the agarose is fully melted in the 1XTBE. The solution was allowed to cool and stained with 40 uL EZ-VISION™ DNA Dye (VWR Life Science, AMRESCO, Philadelphia, USA) /100 mL gel, a non-mutagenic agent that inter-chelates with DNA and fluoresces under ultraviolet (UV) light. The cooling solution was set in a plastic cast to form a solid gel. Agarose polymers form a gel-matrix with pore sizes proportional to the concentration of agarose in the gel solution. DNA samples were mixed with loading dye (2 uL) and added to wells formed in the fully set gel, along with a lane containing GeneRuler® 100 bp Plus DNA Ladder (Fermentas, Ontario, Canada) (**Appendix** I) consisting of fragments of known size. The DNA ladder was used to visually estimate the size of DNA fragments of interest.

The gel is placed into a tank containing 1 X TBE attached to an Enduro™ power pack (Labnet, Woodbridge, USA). An electric current (100 V) was then passed through the gel which for 60 minutes, which caused DNA with a negatively charged phosphate backbone and uniform mass/charge ratio to move towards the positively charged anode. DNA fragments of a smaller size were able to pass through larger pores in the agarose matrix more easily than larger fragments and move faster than larger fragments. The distance that a fragment moves in a defined period of time is inversely proportional to the log of its molecular weight. DNA was visualized by exposure to UV light in the a UVItec FireReader V4 UV gel documentation machine (UVItec, Cambridge, UK) using the UVItec FireReader imaging and editing software (UVItec, Cambridge, UK). The integrity of the DNA is determined by the

banding pattern seen in the captured gel image. Intact genomic DNA will appear as a large, bright band. Degraded DNA will appear as a smear, or series of smaller fragmented bands. The greater the intensity of the smearing, the greater extent of the degradation.

2.12) Working samples

Working samples of 100 ng/uL were made from stock DNA samples to reduce free-thawing of the DNA and ensure the availability of DNA for downstream applications, as well as to minimize the risk of contamination of the stock DNA samples. DNA samples with a concentration that was greater than 100 ng/uL were diluted in Sabax sterile distilled water (sdH_20) (Adcock Ingram, Johannesburg, South Africa) to a final concentration of 100 ng/uL. Samples with concentrations less than 100 ng/uL were not diluted, rather a volume of 30 uL of the stock was aliquoted into new tubes for downstream experiments. Working solutions were stored at 4 °C to prevent degradation due to repeated thawing.

2.13) Characterization of *APOE, PCSK9,* and *CETP* genetic variations

DNA samples from the selected patients in the study cohort were genetically characterised for selected SNPs in the *APOE, PCSK9,* and *CETP* genes which are known to play crucial roles in the lipid metabolism pathway. Gene annotation was performed using PerlV5 based Annottv programme written by Dr. George Rebello (April 2006, Division of Human Genetics, UCT, SA). The *CETP* gene was analysed for known SNPs positions and intron/exon lengths were determined. The 5'region was included in this analysis. Gene sequences for primer designing were annotated from information in the reference sequences provided by the National Centre for Biotechnology Information (NCBI) Gene Entrez website (URL: http://www.ncbi.nih.gov) and the Ensembl website (URL: http://www.ensembl.org).

2.13.1) Selection of genetic variations for analysis

Genetic variations in the *APOE, PCSK9,* and *CETP* genes were selected for characterization based on previous reports in the literature on their functional significance in the lipid metabolism pathway, or association with altered lipid profiles. Many loci have previously been shown to be associated with altered lipid metabolism (**Figure 1.6**), thus a number of selection criteria were employed in order to select SNPs likely to be relevant to our cohort for

cost-effective genotyping. In addition to being linked to dyslipidaemia in other studies, the SNPs had to have been reported in other African populations so as to be frequent enough to have observable effects in our cohort. SNPs with previously reported and validated genotyping assays that allowed for timely and cost-effective analysis of two or more SNPs in one assay were given preference and are summarised in **Table 2.1**.

Based on these criteria 2 SNPs (*APOE rs429358T>C* and *APOE rs7412C>T*) were selected for genotyping in *APOE.* The genotyping of these SNPs, in combination, allows for the characterization of the APOE protein isoform. These SNPs have been widely studied in other population groups, as well as South Africans, in diabetic and non- diabetic cohorts, with variable results[77,128]. However, as the APOE Ɛ4 and Ɛ2 isoforms have been consistently linked to dyslipidaemia, they were included in this study as they were likely to play a major role in our cohort. Thus, they were essential to investigate in the context of our study.

The *PCSK9* gene was included as it is hot topic in the genetics of lipid metabolism, with several African specific variations being strongly associated with altered LDL-c metabolism[88,89,90]. Of these potential targets, two SNPs in the *PCSK9* gene (*rs505151A>G* and *rs28362286C>A*) were selected for genotyping for the following reasons. Firstly, *rs505151A>G* encodes the E670G gain-of-function amino acid change, while *rs28362286C>A* encodes the African specific loss-of-function C679X variation. Secondly, these SNPs are in close proximity, so they can be genotyped together with a single sequencing assay.

CETP is another gene that is currently being widely investigated for its role in HDL-c metabolism, with little work being done in African cohorts[94,96,97]. The *CETP rs708272G>A* (*Taq1B*) polymorphism was selected as a potential target for genotyping despite it encoding a silent SNP with no predicted effects on regulation or splicing site alteration[69]. It has also been consistently associated with altered lipid metabolism which may be due to it being in LD with another unknown functionally important SNP[103]. Pirim *et al.* reported a strong association between the *CETP rs708272A* allele and slightly increased HDL-c levels in an African cohort[97]. Upon annotation of the region encompassing *CETP rs708272G>A*, it was found that a total of 12 CETP SNPs (**Table 2.1**) were in close enough proximity (761bp) to be genotyped in a single sequencing assay. Of these SNPs, 5 (*CETP rs17231520G>A, CETP rs34065661C>G, CETP rs711752G>A, CETP rs708272G>A* and *CETP rs3816117C>T*) were reportedly associated with altered HDL-c metabolism in the same study[97]. Thus, a

sequencing assay of the encompassing DNA region would provide data on 12 SNPs, and also possibly reveal novel SNPs in a greatly variable region of *CETP*.

Table 2.1: *APOE*, *PCSK9* and *CETP* genetic variants selected for genotyping in the study cohort.

Gene	SNP ID	Reported Association	MAF African*	Genotyping Method	Assay Reference
APOE	*rs429358T>C*	Increased LDL-c	0.268	PCR-RFLP	[128]
	rs7412C>T	Decreased LDL-c	0.103	PCR-RFLP	[128]
PCSK9	*rs505151A>G*	Increased LDL-c	0.288	Sanger seq.	[86]
	rs28362286C>A	Decreased LDL-c	0.008	Sanger seq.	[86]
CETP	*rs17231520G>A*	Increased HDL-c	0.082	Sanger seq.	[97]
	rs34065661C>G	Increased HDL-c	0.083	Sanger seq.	[97]
	rs711752G>A	Increased HDL-c	0.247	Sanger seq.	[97]
	rs708272G>A	Increased HDL-c	0.247	Sanger seq.	[97]
	rs5884C>A	None	0.056	Sanger seq.	[97]
	rs34680782C>A	None	0.012	Sanger seq.	[97]
	rs17231534C>A	None	0.133	Sanger seq.	[97]
	rs3816117C>T	Decreased HDL-c	0.404	Sanger seq.	[97]
	rs561260717C>T	None	0.001	Sanger seq.	[97]
	rs34119551A>T	None	0.005	Sanger seq.	[97]
	rs5030708C>T	None	0.000	Sanger seq.	[97]
	rs183782798G>T	None	0.003	Sanger seq.	[97]

Minor allele frequencies (MAF); * average of African population frequencies; frequencies obtained from the 1000 genomes project

2.14) Molecular methods for the characterization of *APOE*, *PCSK9,* and *CETP* genetic variations

Following the identification of genetic variants for analysis in this study, the appropriate genotyping methods were selected and optimized for use with the reagents and equipment that were available. The *APOE* and *PCSK9* SNPs were characterized using modified versions of the methods and primers available from the literature, while the *CETP* SNPs were characterized with a new assay designed for this study (**Table 2.1**). The principal of the methods used are described here.

2.14.1) Principal of polymerase chain reaction (PCR) amplification

Polymerase chain reaction (PCR) amplification is a widely used technique that allows for the exponential amplification of a specific genomic DNA region of interest. The technique relies on the design of oligonucleotide primers which hybridise to the region of interest with high specificity, allowing the replicating action of *Taq* DNA polymerase to take place inside a thermocycler. The reaction consists of repeated thermal cycling steps, starting with an initial denaturation cycle at 94°C during which the double-stranded DNA strands are separated into two complimentary strands. This is followed by repeated cycling of: a 94°C denaturation step, an annealing step at a temperature which is primer dependant and allows the specific binding of forward and reverse primers to the complementary template sequence, and an extension at 72°C which is the optimum temperature for the action of *Taq* DNA polymerase.

Taq DNA polymerase is a thermo-stable DNA polymerase, which catalysis the addition of deoxynucleotides (dNTPs) 3' downstream of the bound primers in the presence of the co-factor $MgCl_2$. Two new daughter strands complementary to the parent template DNA strands are formed in each cycle and thus, the theoretical number of amplicons per cycle is represented by 2^n where n = number of PCR cycles. Several PCR amplification-based techniques have been developed for the genotyping of selected genetic variants including: restriction fragment length polymorphism (RFLP) analysis, and Direct cycle dye-terminator (Sanger) sequencing.

2.14.2) Principal of PCR-RFLP analysis

PCR-RFLP analysis is one of the more widely used methods for the detection of variation at the DNA sequence level. The PCR-RFLP technique makes use of restriction enzymes (RE), isolated from bacteria and archaea, which bind to and cleave DNA substrates at specific sites based on the recognition of specific DNA sequences. These cleaving sites are called restriction sites and are often palindromic. A SNP in the DNA sequence can create or abolish a restriction site. This is useful as it allows for the determination of wild-type or variant alleles based on the digestion of a PCR amplification product with appropriate RE's.

PCR-RFLP analysis makes use of RE that only cleave at a specific site, creating DNA fragments of a defined length. This is determined by the size of the PCR amplification product, and the position of the restriction sites on the DNA sequence. Digestion with an

appropriate RE will result in DNA fragments of defined and deferent lengths, that dependent on wild-type or variant allele being present. These fragments can then be separated and visualised by agarose gel electrophoresis. Thus, the genotype of the SNP of interest can be elucidated. RFLP digest reactions were designed using the online program NEBcutter v2.0 (New England BioLabs, Inc., Ipswich, USA) (http://tools.neb.com/NEBcutter2/index.php).

2.14.3) Direct Cycle Dye-terminator Sequencing

Direct Cycle Dye-terminator Sequencing is a powerful tool used for the direct determination of the nucleotide sequence PCR amplified product. This method is based on the incorporation of fluorescently-labelled dideoxynucleotides (ddNTP's) and dNTP's by *Taq* DNA polymerase to newly synthesized complementary DNA strands. The post-PCR reaction takes place in a BioRad T100 Thermocycler™ (BioRad Laboratories Inc, USA) and follows much of the same steps as a regular PCR amplification. Denaturing of the double-stranded DNA fragments is followed by hybridisation of oligonucleotide primers to their specific binding sites. dNTPs are incorporated 3' downstream of the bound primers into the growing chain of the complimentary strand by the action of *Taq* DNA polymerase. *Taq* DNA polymerase simultaneously randomly incorporates fluorescently-labelled ddNTP's into the nascent strand as. ddNTPs are similar to dNTPs except they contain a hydrogen group on the 3' carbon instead of a hydroxyl moiety. This modification prevents the incorporation of further nucleotides, as phosphodiester bond cannot form between the ddNTP and the incoming dNTPs, resulting in chain termination.

The reaction is repeatedly cycled allowing for synthesis of complimentary fragments of every possible length. The DNA fragments are then heat denatured and separated according to size by capillary electrophoresis, with resolution at a single nucleotide. The ddNTPs are labelled with fluorescent dyes which emit light at different wavelengths under laser excitation. Light emissions are detected and recorded as they pass through a genetic analyser detector. This can be analysed using specialised software and visualised as a chromatogram.

2.15) Genotyping study cohort for selected SNPs

2.15.1) PCR Primer design

PCR primer sequences were available for the 218 bp *APOE* fragment and the 168 bp *PSCK9* fragment from two previous studies by Marrzoq *et al*. (2011) and Evans and Beil (2006), respectively[86,128]. Primers for the characterization of the 12 *CETP* SNPs were designed for this study. A primer pair was designed to amplify a 761 bp region flanking part of the 5' UTR-Exon1-Intron1 region of *CETP*, which contains all 12 relevant *CETP* SNPs in **Table2.1.** Details of the primers used for the three assays are described in **Table 2.2** below.

CETP primers were designed using the annotated sequence as a guide. Primer pairs were designed using the PrimerQuest online tool (Integrated DNA Technologies Inc, 2016) (URL: https://eu.idtdna.com/PrimerQuest/). Optimal primers were chosen to have lengths between 17-28 bp, a *GC* content of approximately 50 %, and were checked so that no SNPs were present in the sequence. Primers were assessed for these parameters, as well as for primer melting temperatures, and significant secondary structures and dimer formations using the OligoAnalyzer v3.1 software (Integrated DNA Technologies Inc, 2017) (URL: https://www.idtdna.com/analyzer/Applications/OligoAnalyzer/).

The primer pair that met the aforementioned criteria were then analysed using the National Centre for Biotechnology (NCBI) Primer-BLAST search tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) to determine the specificity of the primers binding. Primers that were shown to bind with low specificity to many sites in the human genome were excluded. Primers were designed to be highly specific to only the region of interest to ensure that no non-specific PCR products were amplified in the reaction.

Primers were ordered from and synthesised by Integrated DNA Technologies, Inc. (Coralville, IA, USA) and resuspended in 1 X TE buffer with a final concentration of 100 uM. Working solutions of 10 uM were made of all primer pairs and stored at 4 °C.

Table 2.2: Oligonucleotide primer pair summary of three PCR amplification assays.

A) *APOE* PCR Fragment

Primer	**Forward**	**Reverse**
Sequence (5'-3')	TCCAAGGAGCTGCAGGCGGCGCA	GCCCCGGCCTGGTACACTGCCA
Length (bp)	23	22
Melting temperature °C	70.1	68.7
GC content (%)	69.6	72.7
Product size (bp)	218	

B) *PCSK9* PCR Fragment

Primer	**Forward**	**Reverse**
Sequence (5'-3')	TACGCCGTAGACAACACG	TCCCCAGACACCCATCCTGG
Length (bp)	18	20
Melting temperature °C	54.6	61.4
GC content (%)	55.6	65
Product size (bp)	168	

C) *CETP* PCR Fragment

Primer	**Forward**	**Reverse**
Sequence (5'-3')	GGGAGACAAGTAGAAGTTGGG	TCCCCAGACACCCATCCTGG
Length (bp)	21	19
Melting temperature °C	54.8	55.7
GC content (%)	52.4	52.6
Product size (bp)	761	

Primer pairs that were obtained from the literature, namely those for the *APOE* and *PSCK9* fragments, were then re-optimised for PCR amplification in this study. Successful amplification of a target sequence required optimisation of the reaction conditions. These conditions are dependent on the nature of the primers, the condition of the template DNA, and the thermocycler being used for the reaction. Common PCR optimisation steps include: optimisation of the annealing temperature (T_a), the amount of initial template DNA added $MgCl_2$ concentration, extension cycle length, and the amount of *Taq* DNA polymerase added per reaction. Agarose gel electrophoresis, as discussed in **Chapter 2.11.2**, was used to visualise PCR amplification results.

2.15.2) Genotyping for the *APOE rs429358T>C* and *APOE rs7412C>T* single nucleotide polymorphisms

2.15.2.1) Annealing temperature optimisation and PCR amplification of the *APOE* fragment flanking *APOE rs429358T>C* and *APOE rs7412C>T*

A temperature gradient PCR amplification was performed to determine the optimal T_a for the successful amplification of the 218 bp *APOE* fragment. The temperature gradient PCR amplification was performed on a BioRad T100 Thermocycler™ (BioRad Laboratories Inc, USA).

Following T_a optimization, PCR amplification of the 218 bp fragment of *APOE* fragment flanking *APOE rs429358T>C* and *rs7412C>T* was performed on the same BioRad T100 Thermocycler™ with the following cycling conditions: an initial denaturation step at 95 °C for 3 minutes, followed by 35 cycles of 95 °C for 30 seconds, annealing at 65.6 °C for 30 seconds and elongation at 72 °C for 30 seconds. This was followed by a final extension at 72 °C for 10 minutes. Each reaction contained 1 X Green GoTaq® Reaction Buffer (Promega, Madison, USA) with added loading dye, 0.4 mM dNTPs (BioLine, London, UK), 1.5 mM $MgCl_2$, 0.1 uM of each primer and 0.5 Units (U) GoTaq® DNA polymerase (Promega, Madison, USA). Each reaction was made up to its final volume of 25 uL with sdH_2O.

PCR products from both the optimisation and amplification reactions were electrophoresed on 1.5 % (w/v) agarose gels run at 100 V for 60 minutes. Agarose gels were visualized by EZ-VISION™ DNA Dye (VWR Life Science, AMRESCO, Philadelphia, USA) gel staining followed by exposure to UV light in a UVItec FireReader V4 UV gel documentation machine (UVItec, Cambridge, UK) using the UVItec FireReader imaging and editing software (UVItec, Cambridge, UK). GeneRuler® 100 bp Plus DNA Ladder (Fermentas, Ontario, Canada) was used as a size standard (**Appendix I**).

2.15.2.2) PCR-RFLP genotyping for the characterisation of *APOE rs429358T>C* using *AflIII* restriction enzyme digest

The 218 bp *APOE* fragment was digested using the *AflIII* RE to genotype the *APOE rs429358T>C* SNP. Briefly, 10 ul of *APOE* PCR product was digested with 3 U *AflIII* RE (New England BioLabs, Inc., Ipswich, USA) in 1 X NEB3.1 buffer (New England BioLabs, Inc., Ipswich, USA), made up to a final volume of 30 uL with sdH_2O. *AflIII* digestions were

incubated in a water bath at 37 °C overnight. The products of the digest were visualised by electrophoresis at 60 V for 6 hours on a 3.5 % (w/v) Nusieve™ 3:1 agarose gel (Lonza, Rockland, USA) stained with GelRed® nucleic acid dye (Biotium, Inc., Fremont, CA, USA). The GeneRuler® 100 bp Plus DNA Ladder (Fermentas, Ontario, Canada), and GeneRuler® 50bp DNA Ladder (Fermentas, Ontario, Canada) were used as a size standard (**Appendix I**).

The *AflIII* RE recognises the 5'... A↓C R Y G T ...3' restriction site. *APOE rs429358TT* homozygotes have one *AflIII* restriction site, resulting in two fragments sized 168 bp and 50 bp after complete digestion. Introduction of the *APOE rs429358C* allele abolishes the restriction site. Thus, *APOE rs429358CC* homozygotes present with one 218 bp fragment, while *APOE rs429358TC* heterozygotes present with three fragments sized 50 bp, 168 bp, and 218 bp, following digestion with *AflIII*.

2.15.2.3) PCR-RFLP genotyping for the characterisation of *APOE rs7412C>T* using *HaeII* restriction enzyme digest

The 218 bp *APOE* fragment was digested using the *HaeII* RE to genotype the *APOE rs7412C>T* SNP. Briefly, 10 uL of *APOE* PCR product was digested with 3 U *HaeII* RE (New England BioLabs, Inc., Ipswich, USA) in 1 X CutSmart® buffer (New England BioLabs, Inc., Ipswich, USA), made up to a final volume of 30 uL with sdH_2O. *HaeII* digestions were incubated in a water bath at 37 °C overnight. The products of the digest were visualised by electrophoresis at 60 V for 6 hours on a 3.5 % (w/v) Nusieve™ 3:1 agarose gel (Lonza, Rockland, USA) stained with GelRed® nucleic acid dye (Biotium, Inc., Fremont, CA, USA). The GeneRuler® 100 bp Plus DNA Ladder (Fermentas, Ontario, Canada), and GeneRuler® 50 bp DNA Ladder (Fermentas, Ontario, Canada) were used as a size standard (**Appendix I**).

The *HaeII* RE recognises the 5'... G G↓C C ...3' restriction site. *APOE rs7412CC* homozygotes have one *HaeII* restriction site, resulting in two fragments sized 195 bp and 23 bp after complete digestion. Introduction of the *APOE rs7412T* allele abolishes the restriction site. Thus, *APOE rs7412TT* homozygotes present with one 218 bp fragment, while *APOE rs7412CT* heterozygotes present with three fragments sized 23 bp, 195 bp, and 218 bp, following digestion with *HaeII*.

2.15.2.4) Validation of PCR-RFLP genotyping of APOE fragment flanking APOE rs429358T>C and APOE rs7412C>T by Direct Cycle Dye-terminator Sequencing

The PCR-RFLP genotyping of *APOE rs429358T>C* and *APOE rs7412C>T* variants was validated using direct cycle Sanger sequencing. This is done to ensure the identification of banding patterns is correct, so that genotyping of the study cohort can be carried out *en masse*. The digest of *APOE rs7412C>T* by *HaeII* resulted in banding patterns of 23 bp, 195 bp, and 218 bp for *APOE rs7412CT* heterozygotes, and a single 218 bp fragment for *APOE rs7412CC* homozygotes. The 23 bp difference in band size is difficult to see in high percentage gel pictures which are not perfectly clear, making the calling of *APOE rs7412CC* homozygotes more difficult. For this reason, all samples which were thought to be *APOE* r*s7412CC* by means of RE digest were sequenced to confirm their genotype.

The post-PCR sequencing process starts with a clean-up protocol which includes adding 1 U Fast alkaline phosphatase (FastAP) (Thermo Fisher Scientific, Wyham, USA) and 2 U *Exonuclease* I (*Exo*I) (Fermentas, Ontario, Canada) to 10 uL PCR product which is made up to a final reaction volume of 20 uL with sdH_2O. FastAP dephosphorylates unincorporated dNTPs and ddNTPs rendering them inert. The *Exo*I degrades single stranded DNA, including remaining primers. The reaction was incubated at 37 °C for 60 minutes, followed by 72 °C for 15 minutes to denature both enzymes.

Direct cycle sequencing reaction was carried out using the ABI Prism® BigDye® Terminator Cycle Sequencing v3.1 Kit (Applied Biosystems, Carlsbad, CA, USA). 3 ul of cleaned-up PCR product was added to: 2 uL BigDye® Terminator v 3.1 mix containing fluorescently-labelled ddNTPs and *Taq* DNA polymerase, 1 X sequencing buffer to ensure optimal conditions for *Taq* DNA polymerase activity, and 1 uM forward or reverse primer. This was made up to a total volume of 10 uL with sdH_2O.Thermal cycling amplification was performed on a BioRad T100 Thermocycler^{TM} (BioRad Laboratories Inc, USA), starting with an initial denaturation at 96 °C for 5 minutes. This was followed by 35 cycles of 96 °C for 30 seconds, primer annealing at 50°C for 15 seconds, and lastly elongation at 60 °C for 4 minutes.

Sequencing products were purified for capillary electrophoresis by ethanol precipitation. In this process 1 uL 3 M NaOAc and 22 uL of ice-cold absolute ethanol are added to each tube of 10 uL sequenced PCR product, which is then vortexed and left overnight at -20 °C. The following day, the samples were centrifuged at 10000 g for 10 minutes and the supernatant

was discarded. 70 % (v/v) ice-cold ethanol (35 uL) was added to each tube, followed by centrifugation at 10000 g for 10 minutes. The supernatant was discarded, and the pellets left to air dry. The pellets were then re-suspended in 10uL sdH_2O. Sequencing sample (5 uL) was added to 5 ul Hi-Di (highly de-ionised) formamide in a MicroAmp® 96-well reaction plate (Applied Biosystems®, Life Technologies, California, USA), and denatured at 94°C for 5minutes. The plate was loaded onto an ABI Prism 3130*xl* Genetic Analyser (Applied Biosystems®, Life Technologies, California, USA) for capillary electrophoresis. Analysis and alignment of the sequenced PCR products was carried out on the DNASTAR® Lasergene Software (DNASTAR, Inc., Madison, USA) at the University of Cape Town.

2.15.3) Genotyping for the *PCSK9 rs505151A>G* and *PCSK9 rs28362286C>A* single nucleotide polymorphisms

2.15.3.1) Annealing temperature optimisation and PCR amplification of the *PCSK9* fragment flanking *PCSK9 rs505151A>G* and *PCSK9 rs28362286C>A*

A temperature gradient PCR amplification was performed to determine the optimal T_a for the successful amplification of the 168 bp *PCSK9* fragment. The temperature gradient PCR amplification was performed on a BioRad T100 Thermocycler™ (BioRad Laboratories Inc, USA).

Following T_a optimization, PCR amplification of the 168 bp *PCSK9* fragment flanking *PCSK9 rs505151A>G* and *PCSK9 rs28362286C>A* was performed on the same BioRad T100 Thermocycler™ (BioRad Laboratories Inc, USA) with the following cycling conditions: an initial denaturation step at 95 °C for 3 minutes, followed by 34 cycles of 95 °C for 30 seconds, annealing at 56.6 °C for 30 seconds and elongation at 72 °C for 30 seconds. This was followed by a final extension at 72 °C for 5 minutes. Each reaction contained 50 ng of sample DNA, 1 X Green GoTaq® Reaction Buffer (Promega, Madison, USA) with added loading dye, 0.4 mM dNTPs (BioLine, London, UK), 1.5 mM $MgCl_2$, 0.4 uM of each primer and 0.5 Units (U) GoTaq® DNA polymerase (Promega, Madison, WI, USA). Each reaction was made up to its final volume of 25 uL with sdH_2O.

PCR products from both the optimisation and amplification reactions were electrophoresed on 1.5 % (w/v) agarose gels run at 100 V for 60 minutes. Agarose gels were visualized by EZ-VISION™ DNA Dye (VWR Life Science, AMRESCO, Philadelphia, USA) gel staining

followed by exposure to UV light in a UVItec FireReader V4 UV gel documentation machine (UVItec, Cambridge, UK) using the UVItec FireReader imaging and editing software (UVItec, Cambridge, UK). GeneRuler® 100 bp Plus DNA Ladder (Fermentas, Ontario, Canada) was used as a size standard (**Appendix I**).

2.15.3.2) Direct Cycle Dye-terminator Sequencing for the characterisation of *PCSK9 rs505151A>G* and *PCSK9 rs28362286C>A*

The 168 bp *PCSK9* PCR amplified product was genotyped for *PCSK9 rs505151A>G* and *PCSK9 rs28362286C>A* by direct cycle dye-terminator sequencing. The post-PCR sequencing process of all the samples was done at the University of Cape Town with the same clean-up protocol mentioned in **Chapter 2.15.2.4**. Samples were sequenced at the University of Cape Town using this protocol, and at the Central Analytical Facilities service at Stellenbosch University using their own optimised protocols.

The direct cycle sequencing reaction was carried out using the ABI Prism® BigDye® Terminator Cycle Sequencing v3.1 Kit (Applied Biosystems, Carlsbad, CA, USA), using the same protocol as described in **Chapter 2.15.2.4** above. Analysis and alignment of the sequenced PCR products was carried out on the DNASTAR® Lasergene Software (DNASTAR, Inc., Madison, USA) at the University of Cape Town.

2.15.4) Genotyping for the *CETP* single nucleotide polymorphisms

2.15.4.1) Annealing temperature optimisation and PCR amplification of the 761 bp *CETP* fragment flanking 12 *CETP* polymorphisms

A temperature gradient PCR amplification was performed to determine the optimal T_a for the successful amplification of the 761 bp *CETP* fragment. The temperature gradient PCR amplification was performed on a BioRad T100 Thermocycler^{TM} (BioRad Laboratories Inc, USA)

Following T_a optimization, PCR amplification of the 761 bp *CETP* fragment was performed on the same BioRad T100 Thermocycler^{TM} (BioRad Laboratories Inc, USA) with the following cycling conditions: an initial denaturation step at 94 °C for 3 minutes, followed by

40 cycles of 94 °C for 30 seconds, annealing at 66.4 °C for 30 seconds and elongation at 72 °C for 30 seconds. This was followed by a final extension at 72 °C for 10 minutes. PCR amplification reactions of 100 ng of sample DNA contained 1 X Green GoTaq® Reaction Buffer (Promega, Madison, USA), 0.4 mM dNTPs (BioLine, London, UK), 1.5 mM $MgCl_2$, 0.4 uM of each primer and 0.5 U GoTaq® DNA polymerase (Promega, Madison, USA). Each PCR was made up to its final volume of 25 uL with sdH_2O.

PCR products from both the optimisation and amplification reactions were electrophoresed on 1.5 % (w/v) agarose gels run at 100 V for 60 minutes. Agarose gels were visualized by EZ-VISION™ DNA Dye (VWR Life Science, AMRESCO, Philadelphia, USA) gel staining followed by exposure to UV light in a UVItec FireReader V4 UV gel documentation machine (UVItec, Cambridge, UK) using the UVItec FireReader (UVItec, Cambridge, UK) imaging and editing software. GeneRuler® 100 bp Plus DNA Ladder (Fermentas, Ontario, Canada) was used as a size standard **(Appendix I).**

2.15.4.2) Direct Cycle Dye-terminator Sequencing for the characterisation of the *CETP* fragment flanking 12 *CETP* polymorphisms

The 761 bp *CETP* PCR amplified product was genotyped for 12 *CETP* SNPs by direct cycle sequencing. The post-PCR sequencing process of all the samples was done at the University of Cape Town with the same clean-up protocol mentioned in **Chapter 2.15.2.4**. Samples were sequenced at the University of Cape Town using this protocol, and at the Central Analytical Facilities service at Stellenbosch University using their own optimised protocols.

The direct cycle sequencing reaction was carried out using the ABI Prism® BigDye® Terminator Cycle Sequencing v3.1 Kit (Applied Biosystems, Carlsbad, CA, USA), using the same protocol as described in **Chapter 2.15.2.4** above, with 45 cycles of the sequencing reaction as the *CETP* PCR product was larger than the *APOE*, and *PCSK9* products. Analysis and alignment of the sequenced PCR products was carried out on the DNASTAR® Lasergene Software (DNASTAR, Inc., Madison, USA) at the University of Cape Town.

2.16) Statistical analysis

2.16.1) Descriptive statistics

The GraphPad Prism® v5.04 (GraphPad Software, California, USA) software package was used to determine the descriptive statistics of the study cohort. Demographic characteristics such as age and gender, as well as the clinical data of the study cohort such as BMI, smoking status, statin therapy regimen, HbA1c% levels and lipid profile levels were summarized for further analysis.

2.16.2) Stratification of the study cohort by Statin usage status

Nearly 77 % of the study cohort were on statin (mostly Simvastatin) treatment regimens. This was not unusual as CVD accounts for 70% of deaths in T2DM patients[105], so one would expect a large proportion of dyslipidaemic T2DM patients to be on lipid lowering therapy regimens from the outset. Statin usage did have a profound impact on the analysis of this study. LDL-c is the primary target of statin treatment regimens and was, thus, likely to have a significant effect on lipid profiles of the cohort based on published literature (see **Chapter 1.4.2**). Statin usage was also likely to confound any association tests between genetic profiles and lipid profiles.

Therefore, the study cohort was further stratified into individuals that were on statin treatment regimens at the time of enrolment (Sim+) and those who were not on statin treatment regimens (Sim-) for further analysis. This was done in order to assess the impact of statin usage on lipid profiles, as well as to try to account for the effect it would have on further analysis of altered lipid profiles and genetic variation. The GraphPad Prism® v5.04 (GraphPad Software, California, USA) software package was used for this analysis.

2.16.3) Hardy-Weinberg Equilibrium (HWE)

Hardy-Weinberg Equilibrium (HWE) was calculated by comparing the observed and expected frequencies of the study cohort *APOE, PCSK9,* and *CETP* genotype frequencies. The HWE principal states that genetic variations in a randomly selected, randomly mating, and indefinitely large population containing discrete generations will remain constant from one generation to the next in the absence of factors such as mutations, natural and artificial

selection, inbreeding, genetic drift, or random sampling from a population of finite size[129]. For example, in individuals with alleles *A* and *a*, p = expected frequency of *A*, and q = expected frequency of allele *a*. The equilibrium frequencies for the genotypes *AA*, *Aa*, and *aa* are calculated using the equation:

$\mathbf{p^2 + 2pq + q^2 = 1}$, where:

$\mathbf{p^2}$ = expected frequency of *AA* homozygotes

$\mathbf{q^2}$ = expected frequency of *aa* homozygotes

2pq = expected frequency of *Aa* heterozygotes.

Deviation from this equilibrium gives us a standard from which to detect and estimate the effects of these ever-present factors. Technical error during genotyping is most likely the explanation for deviation from HWE, however, nonconforming loci may need to be followed up on with further analysis[129]. In this study, HWE was calculated using the required equations inputted in a Microsoft Excel (Microsoft Corporation, Redmond, WA, USA) spread sheet. Observed and expected allele/genotype frequencies were compared using the Chi^2 (χ^2) test with one degree of freedom (dF) and a two-sided Fischer Exact (FE) test was used to calculate p values. Genotype frequencies with p values <0.05 were considered a statistically significant deviation from HWE.

2.16.4) *APOE*, *PSCK9* and *CETP* genetic variation distribution

The genotype frequencies of genotyped *APOE*, *PSCK9* and *CETP* genetic variants were compared between patients in the Sim+ and Sim- cohorts using the χ^2 test, and the FE test where categorical numbers were <5. Genetic variants with frequencies with p value <0.05 were determined to have statistically significant differences in genotype frequencies. The STATA® v11.2 (StataCorp®, Texas, USA) programme was used to perform genotype frequency analysis.

2.16.5) Linkage disequilibrium (LD) analysis

Linkage disequilibrium (LD) is the non-random association of two or more alleles at different loci. LD analysis is a powerful tool as it gives us genome level insight into a population's history, pattern of geographical subdivision, history of natural selection, genetic drift and mutations which change allele frequencies and are directly influenced by recombination.[130] It also has a more practical use in environments with time and money constraints. If two alleles are in complete LD, then one of the alleles can be used as a genetic marker of the other. Two common measures of LD are the D' and r^2 statistics. The D' statistic is constrained by allele frequencies, thus the r^2 statistic is a better measure of LD. r^2 can be between 0-1. With $r^2 = 1$ indicating complete LD and vice versa. Separate LD analysis for each gene was performed using the SHEsis online software (http://analysis2.bio-x.cn/myAnalysis.php), as the *APOE, PSCK9,* and *CETP* genes are all on different chromosomes (see **Chapter 1.4**).

2.16.6) Univariate association analysis of clinical and genetic variables

Continuous variables were evaluated for normal distribution by the Shapiro-Wilk test. Attempts were made to normalise the variables by various transformations with no success. No samples were excluded for being outliers, as the sample size of the study cohort was already small. Thus, nonparametric tests of association, which do not assume Gaussian distributions, were used to evaluate the associations between clinical parameters and genetic profiles.

Spearman's rank test of correlation was used to determine the possible associations between the continuous clinical variables and lipid profiles. The χ^2 test was used to test for association between multiple categorical variables. The Mann-Whitney U-test (for comparing two groups) and Kruskal-Wallis test with Dunn's multiple comparison test (for comparing 3 or more groups) were used to determine the association between *APOE, PCSK9* and *CETP* genetic variants and median lipid profile values, as well as the associations between stratified groups and lipid profiles. Genetic variants with minor allele frequency (MAF) <0.01 were excluded from analysis as there were too few variant-carrying patients to investigate any meaningful associations. The GraphPad Prism® v5.04 (GraphPad Software, California, USA) and STATA® v11.2 (StataCorp®, Texas, USA) software packages were used to perform all univariate analysis.

2.16.7) Multivariate analysis of clinical and genetic variables with lipid profiles

Following univariate association of genetic variants and lipid profiles, after stratifying for statin usage status, multivariate analysis was performed to investigate the effects of multiple clinical traits and genetic variations on lipid profiles in combination as a means to confirm and/or compare the previous univariate results. In the univariate analysis single genetic variants or clinical variables were investigated for their potential association with altered lipid profiles. The impact of multiple independent variables, in combination, on a single dependant variable was investigated by multivariable linear regression.

The method employed was a multivariable linear regression using a generalised linear model (GLM) with the default Identity link function (Gaussian family). In GLM the dependant variable does not need to be normally distributed. The R Studio V 1.0.153 (RStudio,Inc, Boston,MA, USA) statistical computing software package was used to perform multivariable linear regression using a GLM to investigate the relationship between the lipid profiles (TC, TG, LDL-c, and HDL-c) and possible confounding clinical variables (Gender, Age, Statin usage status, BMI, Smoking status, and HbA1c% as an indicator of T2DM control).

Lipid profiles were then regressed against *APOE, PCSK9,* and *CETP* genetic variants, using an additive model for variant alleles and adjusting for the clinical variables that were significantly associated for each lipid as determined in the previous GLM analysis. Genetic variants included in the model had MAF >0.01 as there were too few variant-carrying patients to investigate any meaningful associations. Where SNPs were in strong LD, only one of the pair was included in analysis.

Where a statistically significant association was observed, and a non-normally distributed independent variable was being accounted for, the p value was adjusted by performing a permutation test by random resampling 10000X. This test involves randomly re-pairing X and Y variables to create a distribution expected by the null hypothesis[131]. This method is recommended in analyses where sample sizes are small, and variables are nonnormally distributed. Assumptions from tests which assume normal distributions do not need to be met in order to draw valid inferences from permutation tests, and they do well at limiting Type 1 errors[131]. The minimum p value was then reported.

Chapter 3: Results

Dyslipidaemia in South Africans is on the rise and the data on the genetics of dyslipidaemia in our indigenous black population is insufficient. It is for this reason that we aimed to characterize the impact of genetic variations in the three important genes *APOE, PCSK9* and *CETP* on altered lipid metabolism in a population of T2DM patients who are high risk for developing CVD. We aimed to identify genetic biomarkers of dyslipidaemia in our population which can be used in the future to predict predisposition to high risk lipid profiles. This information could be used to better inform clinicians, who would be able apply appropriate management and preventative strategies.

We set out to achieve this aim by completing of a number of objectives. We recruited participants presenting for diabetes screening or treatment at the Chris Hani Baragwanath Academic Hospital and determined their lipid profiles. We combined this information with clinical and demographic data, to gain insight into the potential contribution of these variables to altered lipid profiles. We identified and characterised several genetic variants in *APOE, PCSK9* and *CETP* which had been shown to affect lipid metabolism in other study populations. We then investigated the contribution of these variations to altered TC, HDL-c, LDL-c, TG and HbA1c levels by genotype-phenotype associations using both univariate and multivariate analysis techniques. In doing so we identify here several biomarkers of dyslipidaemia in a black South African T2DM cohort.

3.1) Demographic and clinical features of the research cohort

A total of 417 participants were included in our research cohort and were genotyped for selected genetic variants in *APOE, PCSK9,* and *CETP*. These patients had data on TC, TG, LDL-c, HDL-c and HbA1c% levels, as well as data on other important clinical and demographic features. (**Table 3.1**). All of the study participants were black South Africans of Bantu origin. The median age of the cohort was 59 years old, with patients ranging from 19 to 91 years. The majority of the cohort was female (65.04 %). Only 23.02 % of patients self-reported as smokers. A number of (n=20) patients had TG levels >4.5 mmol/L and, as a result of the limitations of the Friedewald equation method, did not have estimated LDL-c levels.

Additionally, a small number of patients had various clinical/demographic data missing from the database, thus the number of observations listed in **Table 3.1** was used for subsequent analysis.

Table 3.1: Demographic and clinical features of the combined type 2 diabetes mellitus cohort.

Clinical/Demographic feature	**n=417**
Median Age (years) (n=378)	59 (19 – 91)
Median BMI (kg/m^2) (n=378)	29.5 (17.0 – 63.3)
Gender M/F (%) (n=409)	143 (34.96) / 266 (65.04)
Smoker Y/N (%) (n=376)	82 (21.11) / 294 (78.89)
Statin Treatment Y/N (%) (n=378)	291 (76.98) / 87 (23.02)
Median HbA1c (%) (n=411)	6.60 (2.20 – 28.90)
Median Total Cholesterol (mmol/L) (n=417)	4.92 (1.64 – 13.28)
Median Triglycerides (mmol/L) (n=416)	1.64 (0.33 – 14.53)
Median HDL-C (mmol/L) (n=417)	1.23 (0.11 – 3.42)
Median LDL-C (mmol/L) (n=397)	2.70 (0.48 – 6.08)

Continuous results reported as median (range); Categorical results reported as n (%); Total cholesterol = TC; Triglycerides = TG; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; BMI = Body mass index; HbA1c = Haemoglobin A1C.

Of the combined study cohort, 180 patients (43.17 %) had TC levels >5.2 mmol/L, 140 patients (33.57 %) had TG >2.3 mmol/L, 60 patients had LDL-c >4.1 mmol/L, and 125 (29.98 %) had HDL-c levels <1.0. TC, TG, and LDL-c levels above, and HDL-c levels below these conservative cut-off are indicators of a high risk of CVD events according to the ESC/EAS guidelines for non-diabetic individuals[52]. As these patients are T2DM, they would be considered as high CVD risk patients and were taken as highly dyslipidaemic individuals in this study. **Table 3.2** shows a comparison of the number of patients considered

dyslipidaemic from our study and and the Heart of Soweto study according to the cut-offs in their published study[114]**.** There is a significantly higher proportion of TC, TG, and LDL-c dyslipidaemic patients and lower proportion of HDL-c dyslipidaemic patients from our cohort, than from the Heart of Soweto Study ($p<0.0001$, in all cases).

Table 3.2: Comparison of number of black South African patients with dyslipidaemic traits in this study and those reported in the Heart of Soweto study.

Dyslipidaemic Trait	**This Study**	**Heart of Soweto Study**[114]	**Chi^2 P value**
TC >4.5 mmol/l	253 (61 %)	715 (39 %)	<0.0001
TG >1.7 mmol/l	200 (48 %)	386 (23 %)	<0.0001
LDL-c >2.5 mmol/l	229 (55 %)	718 (44 %)	<0.0001
HDL-c <1.0 mmol/l	125 (30 %)	1044 (63 %)	<0.0001

Results reported as n (%). Total cholesterol = TC. Triglycerides = TG. High-density lipoprotein cholesterol = HDL-c. Low-density lipoprotein cholesterol = LDL-c. $p<0.05$ is significant.

The incidence of dyslipidaemia is significantly higher in our cohort. We wanted to determine what clinical and demographic factors, if any, were associated with the aberrant lipid profiles seen in our study cohort.

3.2) Test of normality by Shapiro-Wilk test

The distribution of the lipid profiles and co-variates of our case cohort were tested for normality using the Shapiro Wilk test. BMI, HbA1c%, TC, TG, LDL-c, HDL-c are all non-normally distributed with $p<0.05$. Age is approaching a normal distribution ($p=0.033$) and Gender, a binary variable, is normally distributed. Thus, non-parametric tests of association were used in subsequent analysis.

3.3) Stratification by Statin use

A large percentage of the research cohort (**77 %**) were on statin (mostly simvastatin) treatment. Thus, the cohort was stratified by statin treatment status in further analysis. We report significant differences in demographic data and lipid levels between Sim+ and Sim- cohorts (**Table 3.3**). Patients on statins had significantly higher median age (59 years vs.50.5 years; p=0.0002), higher median HbA1c% levels (6.9 % vs. 5.5 %; p=0.0001), and generally, significantly increased lipid levels than those not on statins (p=0.0001 in all cases). A trend towards significantly higher median BMI (30.1 kg/m^2 vs 29.4 kg/m^2; p=0.058) was seen in the Sim+ cohort. There no differences observed for gender and tobacco smoking status. Statin treatment status has a significant effect on observed lipid profiles and needed to be accounted for in the further analysis of the potential genetic contribution to dyslipidaemia.

Table 3.3: Demographic characteristics of study cohort stratified by statin therapy usage.

	Sim+ (n=291)	**Sim- (n=87)**	**P Value**
Median Age (years)	59 (32 – 91)	50.5 (19 – 85)	0.0002
Median BMI (kg/m^2)	30.1 (17.0 – 63.3)	29.4 (18.6 – 49.9)	0.058
Gender M/F (%)	92 (35.12) / 198 (64.88)	36 (41.38) / 51 (58.62)	0.095
Smoker Y/N (%)	63 (21.65) / 229 (78.69)	19 (22.35) / 66 (77.65)	0.89
Median HbA1c (%)	6.9 (2.2 – 28.9)	5.5 (2.6 – 17.3)	0.0001
Median Total Cholesterol (mmol/L)	5.42 (1.64 – 13.28)	3.66 (1.78 – 7.64)	0.0001
Median Triglycerides (mmol/L)	1.86 (0.40 – 14.53)	1.23 (0.33 – 4.48)	0.0001
Median HDL-C (mmol/L)	1.28 (0.15 – 3.42)	1.01 (0.11 – 2.90)	0.0001
Median LDL-C (mmol/L)	3.08 (0.48 – 6.08)	1.9 (0.71 – 5.17)	0.0001

Sim+ = patients on statins; Sim- = patients not on any statins; Continuous results reported as median (range); Categorical results reported as n (%); Total cholesterol = TC; Triglycerides = TG; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; BMI = Body mass index; HbA1c = Haemoglobin A1C; $p<0.05$ is significant.

3.4) Investigating associations between clinical characteristics, demographic features, and altered lipid profiles

The most apparent confounding factor that was identified as significantly associated with aberrant lipid profiles was patients being treated with lipid lowering therapies in the form of statins. We wanted to determine if there were any other clinical or demographic factors that had a significant impact on lipid profiles, which might need to be accounted for when trying to identify potential genetic biomarkers of dyslipidaemia. We employed two methods of analysis to investigate these variables. The potential associations between single clinical variables and lipid profiles was evaluated by the non-parametric Spearman's rank test of correlation. This method allows for the non-parametric testing of correlations between continuous and categorical variables.

Spearman's rank test of correlation testing (**Table 3.4)** showed that there were significantly more women among the older participants than males (Rho=0.102, p=0.048). Statin therapy regimens were seen more frequently in the elderly (Rho=0.206, p=0.001), in individuals with higher BMI (Rho=0.110, p=0.034), and in patients with higher blood glucose concentrations (HbA1c%) (Rho=0.352, p<0.001). Patients with increased HbA1c% levels were also slightly older (Rho=0.117, p=0.025). No significant correlations were found between smoking and other clinical variables.

Table 3.4: Correlation of clinical and demographic variables by the Spearman's rank test.

Clinical Variable	Gender (Female)	Age	On Statins	Smoking	BMI	HbA1c%
Gender (Female)	-	-	-	-	-	-
Age	0.102 **p=0.048**	-	-	-	-	-
On Statins	0.078 p=0.133	0.206 **p=0.001**	-	-	-	-
Smoking	-0.083 p=0.110	0.0826 p=0.113	-0.016 p=0.763	-	-	-
BMI	0.026 p=0.619	0.033 p=0.533	0.110 **p=0.034**	0.053 p=0.313	-	-
HbA1c%	0.099 p=0.056	0.117 **p=0.025**	0.352 **p<0.001**	0.069 p=0.185	0.065 p=0.269	-

Table shows Rho coefficient and corresponding p value. A positive Rho indicates a positive correlation. $p<0.05$ is significant. Total cholesterol = TC; Triglycerides = TG; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; BMI = Body mass index; HbA1c = Haemoglobin A1C.

The Spearman's rank correlation test was used to investigate potential univariate associations between clinical variables and TC, TG, LDL-c and HDL-c lipid profiles in the combined study cohort, the results of which are seen in **Table 3.5**. We also employed multivariable linear regression, using a GLM framework, to investigate the effects of multiple clinical characteristics and demographic features on lipid profiles to better inform the subsequent analysis of selected genetic variants. The following variables were investigated for their combined effects on TC, TG, LDL-c, and HDL-c: Gender, Age, Statin therapy status, smoking status, BMI, and HbA1c% - as an indicator of how well controlled a patient's glucose metabolism is. The results of which are shown in **Table 3.6**. In instances where $p<0.001$ or where an independent variable was not normally distributed, permutation tests using random shuffling 10000x were performed to correct for non-normal distribution of the test statistic.

The first result that stands out is the confirmation of the association of statin therapy with increased levels of TC (OR=5.69, p=3.53e^{-05}), TG (OR=1.56, p=0.016), LDL-c (OR=3.40, p=1.76e^{-05}), and HDL-c (OR=1.40, p= 1.03e^{-05}). Thus, statin therapy status is an indicator of TC, TG and LDL-c dyslipidaemia.

From the combination of analysis, we see that females had significantly increased in HDL-c levels than males (Rho=0.110, p=0.034). This was confirmed by the multivariate analysis (OR=1.12, p=0.044). The correlation seen between older patients and increased HDL-c levels (Rho=0.136, p=0.009) was not repeated when adjusting for multiple variables (p=0.328), however older patients were significantly associated with slightly decreased TG levels (OR=0.98, p=0.003). Patients with greater BMIs were significantly correlated with increased TG levels in both univariate and multivariate analysis (Rho=0.159, p=0.002; OR=1.03, p=0.011). Poor blood glucose control was significantly correlated with increased TG (Rho=0.378, p<0.001; OR=1.11, p=3.26e^{-05}) and very slightly decreased HDL-c (OR=0.99, p=3.28e^{-05}). Associations between HbA1c%, TC and LDL-c were not repeated when adjusting for other confounding factors. Smoking was not associated with altered lipid profiles in our cohort.

Table 3.5: Spearman's rank test of correlation between clinical variables and lipid profiles.

Variable	Gender (Female)	Age	On Statins	Smoking	BMI	HbA1c%
Total Cholesterol	0.076 p=0.146	0.070 p=0.179	0.552 **p<0.001**	-0.0407 0.436	0.065 p=0.217	0.203 **p<0.001**
Triglycerides	0.063 p=0.231	-0.051 p=0.334	0.216 **p<0.001**	0.046 p=0.381	0.159 **p=0.002**	0.378 **p<0.001**
LDL-c	0.003 p=0.945	0.058 p=0.275	0.479 **p<0.001**	-0.039 p=0.459	-0.011 p=0.843	0.131 **p=0.014**
HDL-c	0.110 **p=0.034**	0.136 **p=0.009**	0.272 **p<0.001**	-0.064 p=0.221	0.066 p=0.208	-0.090 p=0.083

Table shows Rho coefficient and corresponding p value. A positive Rho indicates a positive correlation. p<0.05 is significant. Total cholesterol = TC; Triglycerides = TG; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; BMI = Body mass index; HbA1c = Haemoglobin A1C.

Table 3.6: Multivariate analysis of associations between confounding clinical variables and lipid profiles in the combined South T2DM cohort by multivariable linear regression using a generalised linear model

Variable	Total Cholesterol			Triglycerides			LDL-c			HDL-c		
	OR	**95%CI**	**P value**	**OR**	**95%CI**	**P value**	**OR**	**95%CI**	**P value**	**OR**	**95%CI**	**P value**
Gender	1.10	0.87 - 1.39	0.50	1.21	0.94- 1.55	0.216	0.95	0.79- 1.14	0.62	1.12	1.02- 1.22	**0.044**
Age	0.99	0.98 - 1.00	0.18	0.98	0.97- 0.99	**0.003**	0.99	0.99- 1.01	0.63	1.00	0.99- 1.01	0.328
On Statins	5.69	4.28 - 7.58	**$3.53e^{-05}$***	1.56	1.15- 2.11	**0.016**	3.40	2.72- 4.25	**$1.76e^{-05}$***	1.40	1.26- 1.56	**$1.03e^{-05}$***
Smoking	1.00	0.76 - 1.31	1.00	1.31	0.98- 1.75	0.124	0.93	0.75- 1.16	0.59	0.95	0.86- 1.06	0.438
BMI	1.00	0.98 - 1.02	0.80	1.03	1.01- 1.05	**0.011**	0.98	0.97- 1.00	0.12	1.00	0.99- 1.01	0.694
HbA1c%	1.03	0.99 - 1.06	0.20	1.11	1.07- 1.15	**$3.26e^{-05}$***	0.99	0.97- 1.02	0.77	0.99	0.97- 0.99	**$3.28e^{-05}$***

N=369 due to missing data; *= random shuffling of p value 10000x; Bold indicated statistically significant p values (<0.05). Total cholesterol = TC; Triglycerides = TG; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; BMI = Body mass index; HbA1c = Haemoglobin A1C. CI = Confidence Interval. OR = odds ratio

3.5) Genotyping results for characterisation of variants in *APOE*, *PCSK9*, and *CETP*

We now had a clear picture of the clinical and demographic variables which were playing significant roles in the variation of aberrant lipid profiles. We decided to adjust for statin use by stratification when investigating the univariate effects of *APOE, PCSK9*, and *CETP* genetic variants on lipid profiles. In the multivariate analysis, we adjusted for a unique combination of associated clinical variable for each lipid. We then characterised our cohort for the selected genetic variants.

3.5.1) Characterisation of *APOE rs429358T>C* and *APOE rs7412C>T*

The *APOE* 218 bp DNA fragment which flanks the *APOE rs429358T>C* and *APOE rs7412C>T* SNPS was successfully amplified (**Figure 3.**1) following PCR optimisation.

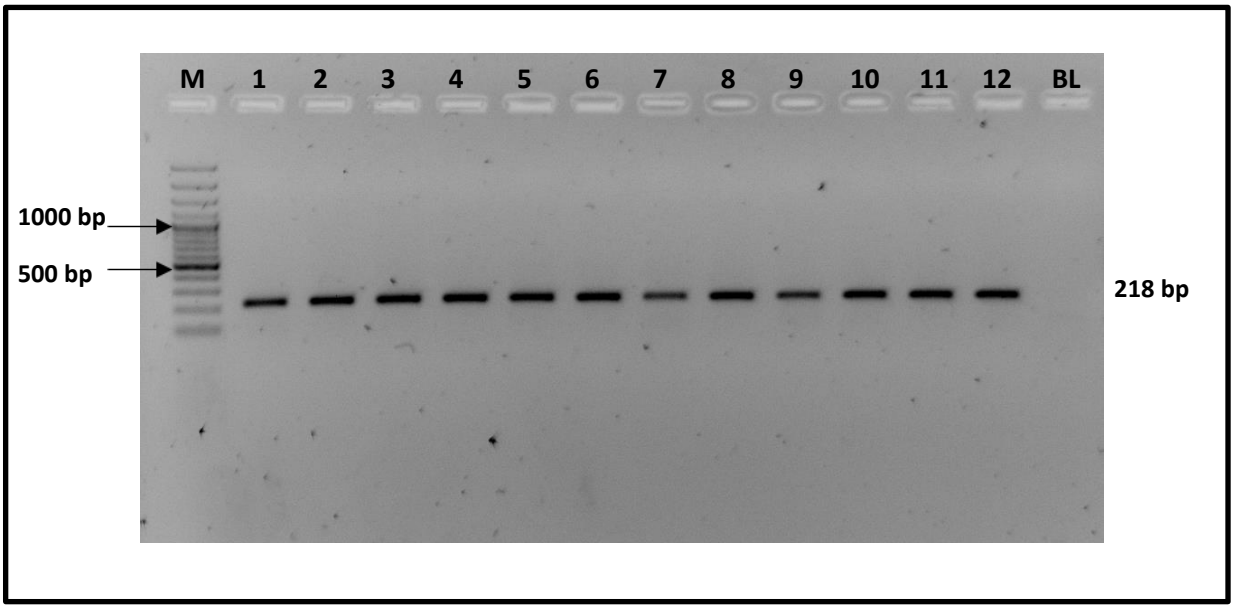


Figure 3.1: PCR amplification of the 218 bp *APOE* fragment electrophoresed at 100 V for 60 minutes on a 1.5 % (w/v) agarose gel. Lane M represents the GeneRuler® 100 bp Plus DNA Ladder Molecular Weight Marker (Fermentas, Ontario, Canada). Lanes 1-12 represent 12 case cohort DNA samples. Lane BL is a no template control.

Genotyping of the *APOE rs429358T>C* variant was performed by PCR-RFLP using the *AflIII* restriction enzyme. Digestion of the 218 bp PCR product resulted in 2 fragments sized

168 bp and 50 bp in the *APOE rs429358TT* homozygotes. If the C allele is present, the restriction site is abolished. *APOE rs429358CC* homozygotes present with one 218 bp fragment, while *APOE rs429358TC* heterozygotes present with three fragments sized 50 bp, 168 bp, and 218 bp. **Figure 3.2** shows successful RFLP digestion of the 218 bp *APOE* fragment with *AflIII* for the characterisation of *APOE rs429358T>C.*

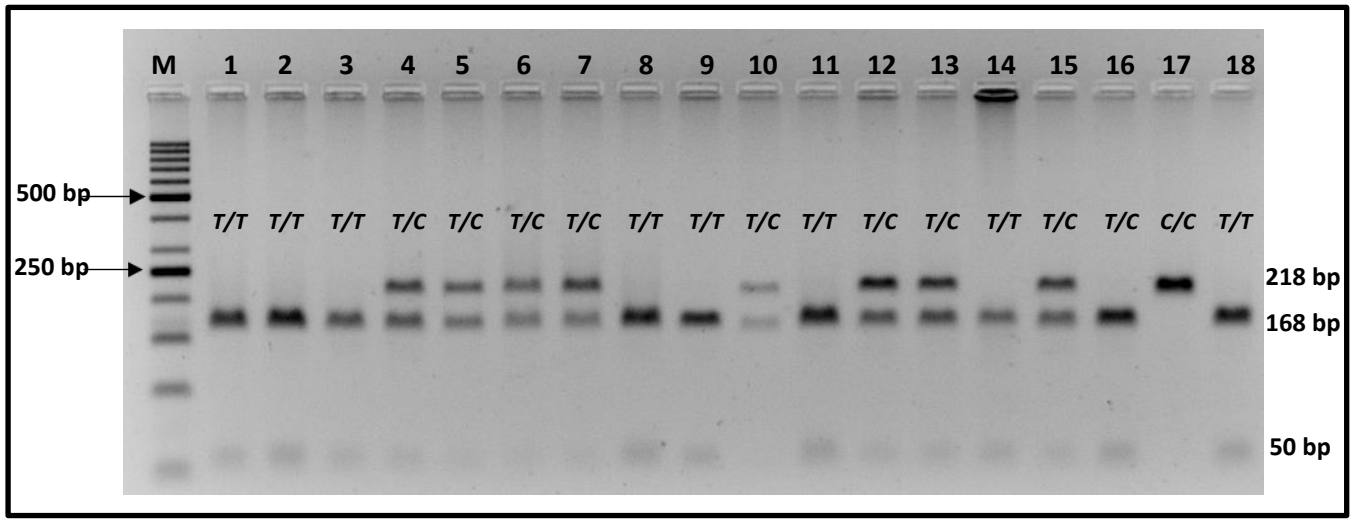


Figure 3.2: *AflIII* Digest of the 218 bp *APOE* fragment electrophoresed at 60 V for 6 hours on a 3.5 % (w/v) agarose gel. Lane M represents the GeneRuler® 50 bp Plus DNA Ladder Molecular Weight Marker (Fermentas, Ontario, Canada). Lanes 1, 2, 3, 8, 9, 14, 16, and 18 represent the *T/T* genotype. Lanes 4, 5, 6, 7, 10, 12, 13, and 15 represent the *T/C* genotype. Lane 17 represents the *C/C* genotype.

Genotyping of the *APOE rs7412C>T* variant was performed by PCR-RFLP using the *HaeII* restriction enzyme. Digestion of the 218 bp PCR product resulted in fragments sized 195 bp and 23 bp in *APOE rs7412CC* homozygotes. If the T allele is present, the restriction site is abolished. *APOE rs7412TT* homozygotes present with one 218 bp fragment, while *APOE rs7412CT* heterozygotes present with three fragments sized 23 bp, 195 bp, and 218 bp. **Figure 3.3** shows successful RFLP digestion of the 218 bp *APOE* fragment with *HaeII* for the characterisation of *APOE rs7412C>T.* The 23 bp band has run off the gel by the time the gel has run long enough to resolve the 195 bp and 218 bp bands. The molecular weight marker in lane M has run slightly faster than the digested PCR products in lanes 1, 2, and 3. This is a result of the GelRed® nucleic acid dye interaction (Biotium, Inc., Fremont, CA, USA).

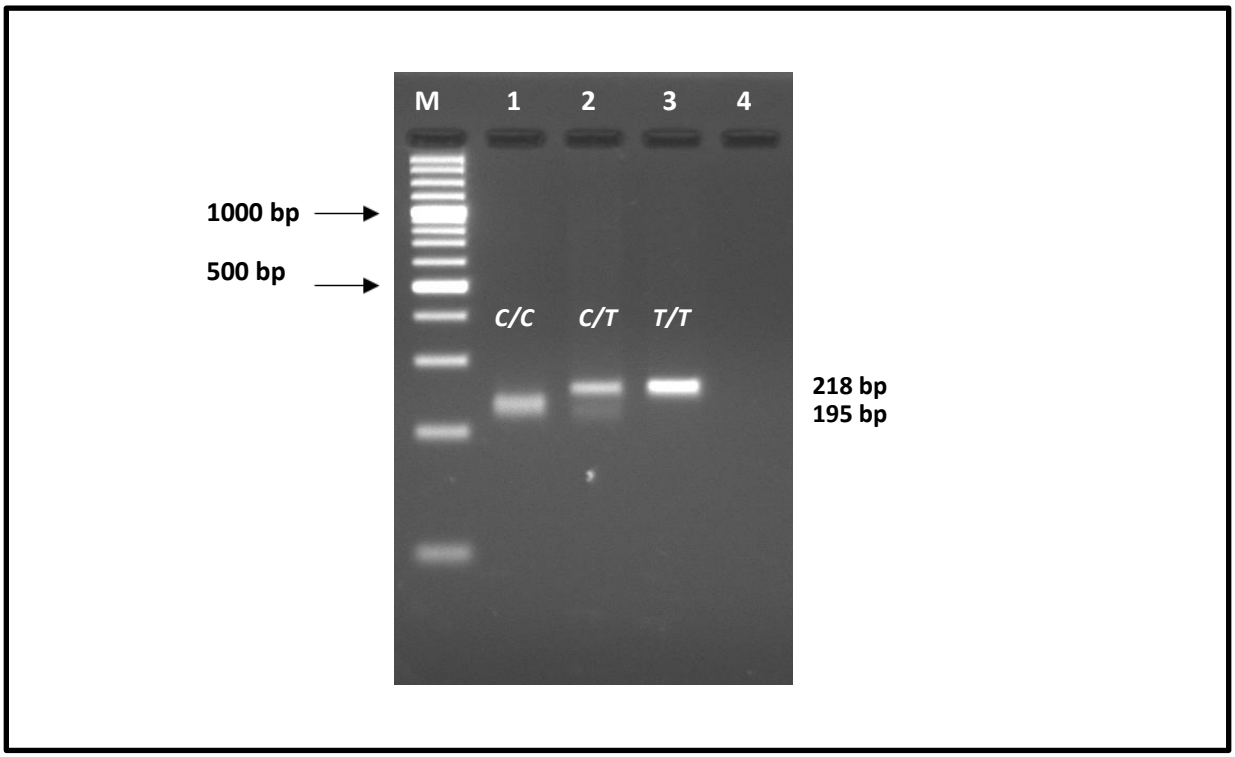


Figure 3.3: *HaeII* Digest of the 218 bp *APOE* fragment electrophoresed at 60 V for 6 hours on a 3.5 % (w/v) agarose gel. Lane M represents the GeneRuler® 100 bp Plus DNA Ladder Molecular Weight Marker (Fermentas, Ontario, Canada). Lane 1, represent the *C/C* genotype. Lanes 2 represent the *C/T* genotype. Lane 3 represents the *T/T* genotype. Lane 4 is empty. The 23 bp band has run off the gel.

The PCR-RFLP protocols for genotyping the *APOE rs429358T>C* and *APOE rs7412C>T* variants were quite challenging due to the very similar sizes of the digested bands, with *APOE rs7412C>T* being difficult to call in particular. As such, it was essential to validate genotyping calls by direct cycle dye-termination sequencing. **Figure 3.4** shows the electropherograms of three successfully aligned *APOE* PCR amplified fragments, which have been sequenced to verify the *rs7412C/C, rs7412C/T* and *rs7412T/T* genotypes, and the *APOE rs429358T/T, APOE rs429358T/C,* and *APOE rs429358C/C* genotypes. Following the successful genotyping and validation, thereof, of the *APOE rs429358T>C* and *APOE rs7412C>T* variants, the APOE isoform of each patient could be determined using a combination of the *APOE rs429358T>C* and *APOE rs7412C>T* genotypes.

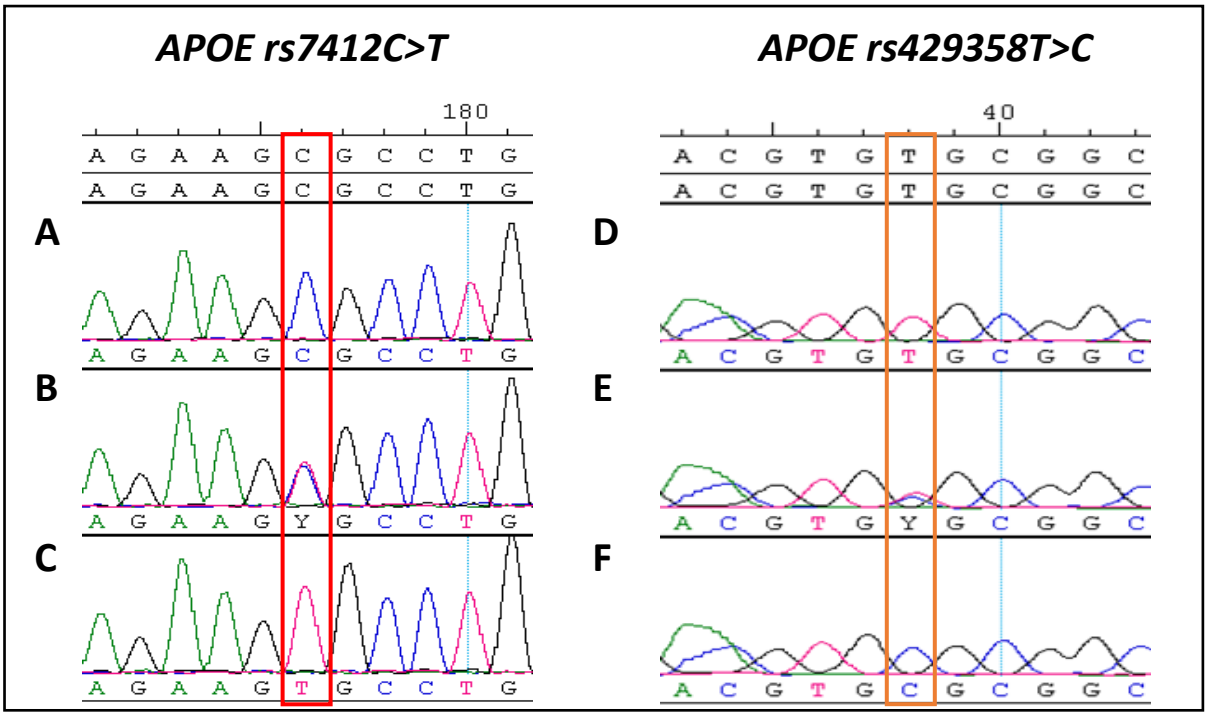


Figure 3.4: Electropherogram of sequenced *APOE* PCR amplified fragments. The *APOE rs7412C>T* variant is highlighted by the red box. The *APOE rs429358T>C* variant is highlighted by the orange box. **A** represents a sample with the *rs7412C/C* genotype. **B** represents a sample with the *rs7412C/T* genotype. **C** represents a sample with the *rs7412T/T* genotype. **D** represents a sample with the *rs429358T/T* genotype. **E** represents a sample with the *rs429358T/C* genotype. **F** represents a sample with the *rs429358C/C* genotype.

3.5.2) Characterisation of *PCSK9 rs505151A>G and PCSK9 rs28362286C>A*

The 168 bp *PCSK9* DNA fragment which flanks the *PCSK9 rs505151A>G and PCSK9 rs28362286C>A* SNPS was successfully amplified (**Figure 3.5**) following PCR optimisation.

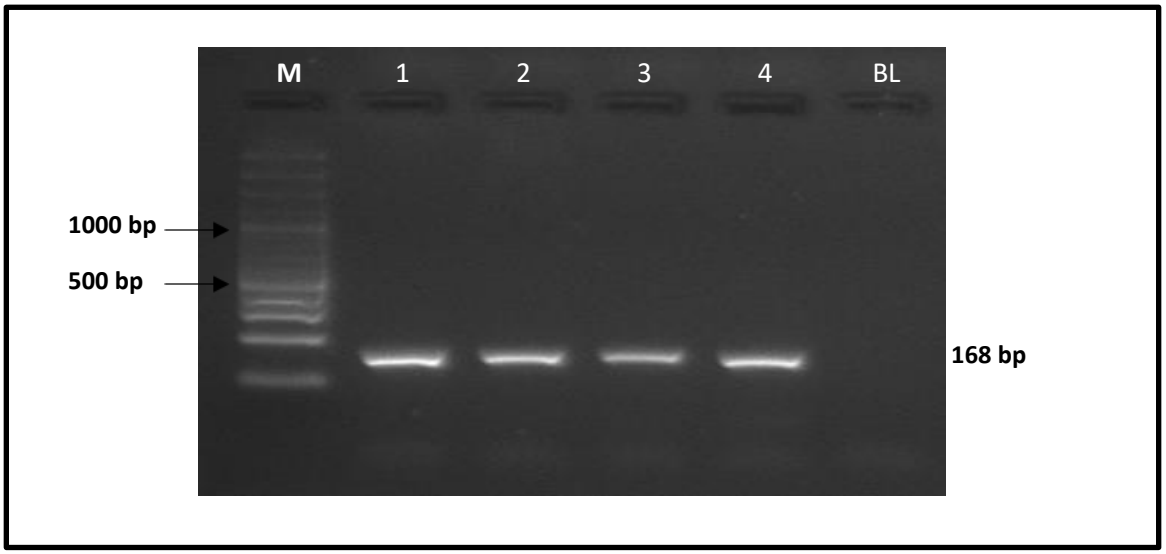


Figure 3.5: PCR amplification of the 168 bp *PCSK9* fragment electrophoresed at 100 V for 60 minutes on a 1.5 % (w/v) agarose gel**.** Lane M represents the GeneRuler® 100 bp Plus DNA Ladder Molecular Weight Marker (Fermentas, Ontario, Canada). Lanes 1-4 represent 4 case cohort DNA samples. Lane BL is a no template control.

The 168 bp *PCSK9* fragment flanking the *PCSK9 rs505151G>A* and *PCSK9 rs28362286C>A* variants was characterised by direct cycle dye-terminator sequencing. **Figure 3.6** shows electropherograms of the region encompassing both of *PCSK9* the variants, which was analysed on the DNASTAR® Lasergene Software (DNASTAR, Inc., Madison, USA) suite at the University of Cape Town.

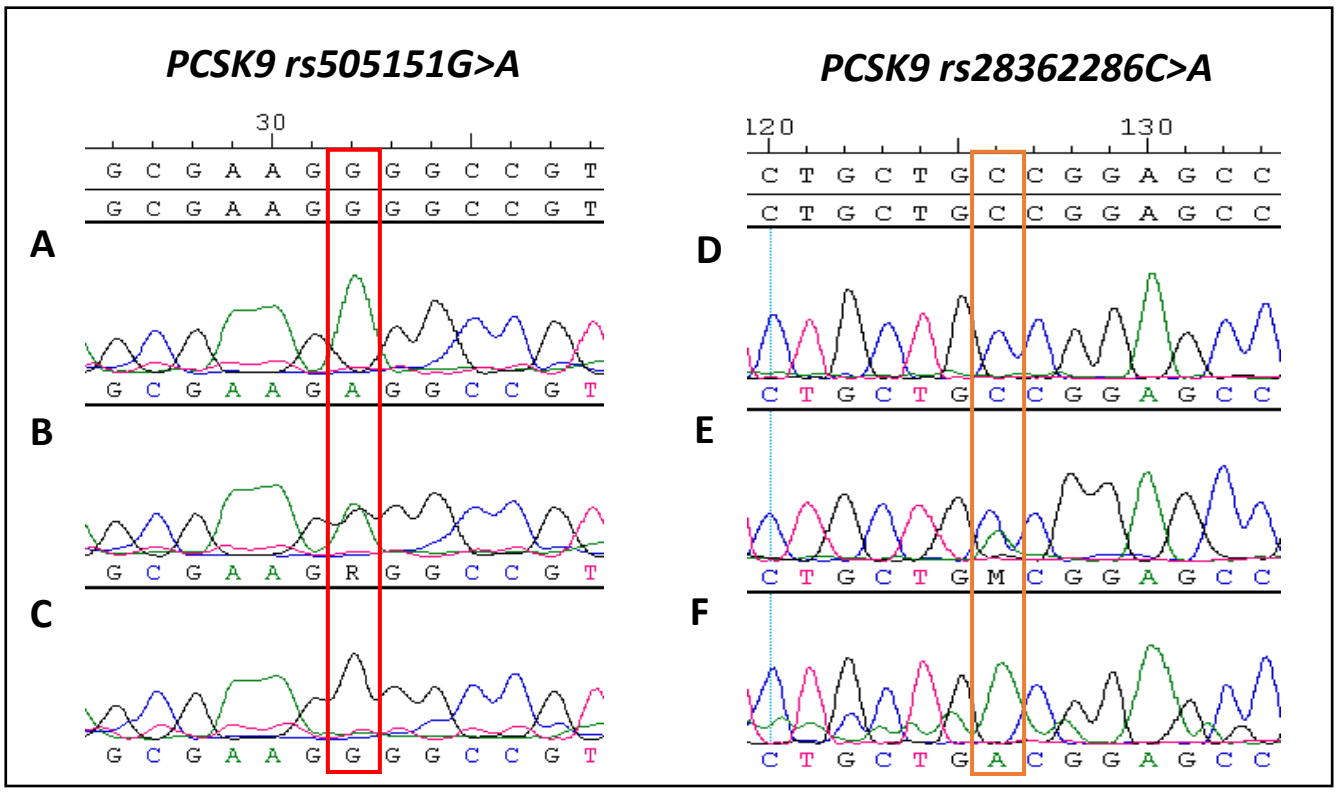


Figure 3.6: Electropherogram of sequenced *PCSK9* PCR amplified fragments**.** The *PCSK9 rs505151G>A* and *PCSK9 rs28362286C>A* variants are highlighted by the red and orange boxes, respectively. **A** represents a sample with the *PCSK9 rs505151A/A* genotype. **B** represents a sample with the *PCSK9 rs505151A/G* genotype. **C** represents a sample with the *PCSK9 rs505151G*/G genotype. **D** represents a sample with the *PCSK9 rs28362286C/C* genotype. **E** represents a sample with the *PCSK9 rs28362286C/A* genotype. **F** represents a sample with the *PCSK9 rs28362286A/A* genotype.

3.5.3) Characterisation of the *CETP* genetic variants

The 761 bp *CETP* DNA fragment which encompasses the 12 *CETP* SNPs was successfully amplified (**Figure 3.7**) following PCR optimisation.

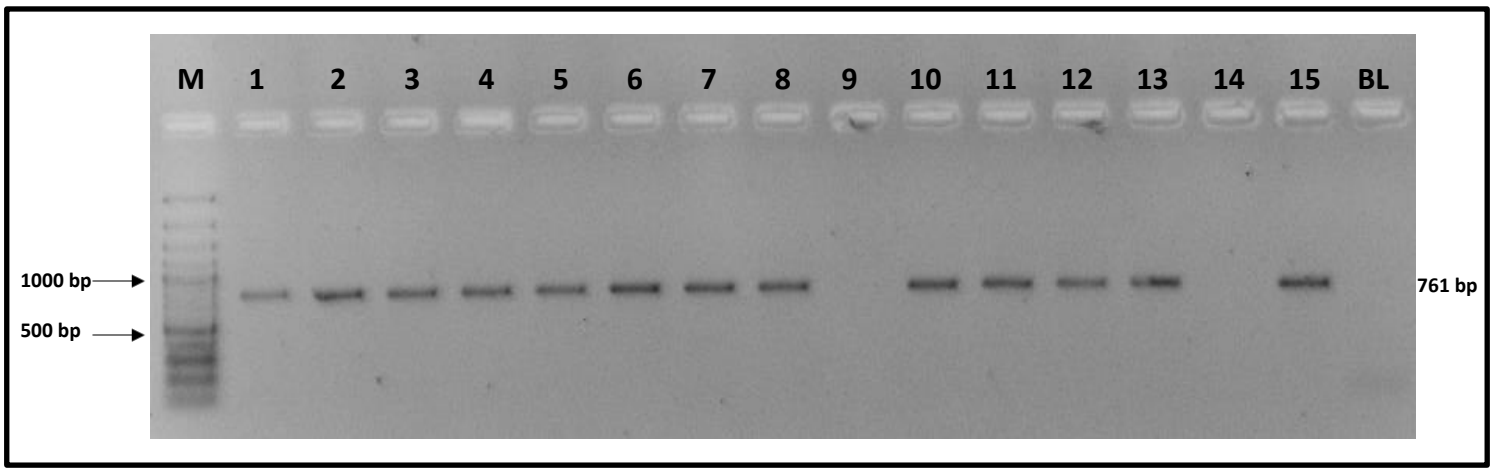


Figure 3.7: PCR amplification of the 761 bp *CETP* fragment electrophoresed at 100 V for 60 minutes on a 1.5% (w/v) agarose gel. Lane M represents the GeneRuler® 100 bp Plus DNA Ladder Molecular Weight Marker (Fermentas, Ontario, Canada). Lanes 1-15 represent 15 case cohort DNA samples. Lanes 9 and 14 failed PCR amplification. Lane BL is a no template control.

The 761 bp *CETP* fragment was characterised for the 12 *CETP* variants by direct cycle dye-terminator sequencing. **Figure 3.8** shows electropherograms of the regions encompassing the *CETP rs34065661C>G* and *rs708272G>A* variants, which was analysed on the DNASTAR® Lasergene Software (DNASTAR, Inc., Madison, USA) suite at the University of Cape Town. Examples of electropherograms for the characterisation of all additional *CETP* variants can be found in **Appendix IV.**

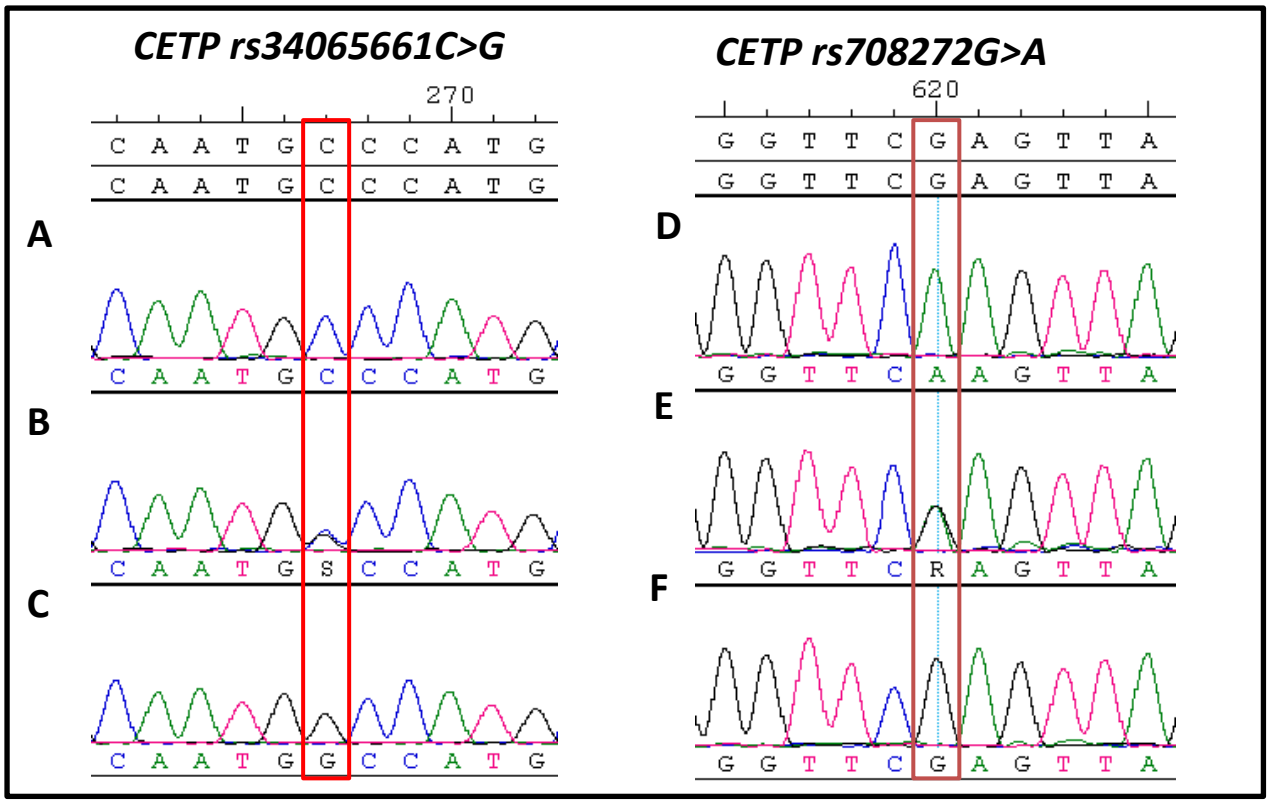


Figure 3.8: Electropherogram of sequenced *CETP* PCR amplified fragments**.** The *CETP rs34065661C>G* and *CETP rs708272G>A* variants are highlighted by the red and orange boxes, respectively. **A** represents a sample with the *CETP rs34065661C/C* genotype. **B** represents a sample with the *CETP rs34065661C/G* genotype. **C** represents a sample with the *CETP rs34065661G/G* genotype. **D** represents a sample with the *CETP rs708272A/A* genotype. **E** represents a sample with the *CETP rs708272A/G* genotype. **F** represents a sample with the *CETP rs708272G/G* genotype.

3.5.4) *APOE, PCSK9* and *CETP* genetic variant distribution

The genotype distribution, HWE analysis, and comparison of allele frequencies by stratification by statin use for the *APOE, PCSK9* and *CETP* genetic variants are shown in **Table 3.7** following successful genotyping and analysis. The *PCSK9 rs28362286C>A* and *CETP rs34680782C>A* variants do not follow HWE distribution in the combined cohort (p=0.028 and p=0.002, respectively). This is most likely due to the non-random selection of our study cohort. We report significant differences in the allele frequencies of the *APOE rs7412C>T* (p<0.001) variant when comparing the Sim+ and Sim- cohorts. The *APOE rs7412T* allele is found more frequently in the Sim- cohort.

Table 3.7: *APOE, PCSK9* and *CETP* genotype frequencies in the combined T2DM cohort with corresponding HWE analysis, and comparison of genotype frequencies in the statin stratified groups.

Gene	SNP ID	Genotype	Combined Cohort n (freq.)	HWE P Value	Sim+ Cohort n (freq.)	Sim- Cohort n (freq.)	Chi^2/FE P value (global)
APOE	***rs429358***	**T/T**	229 (55.05)	0.131	158 (54.48)	46 (53.49)	0.177
		T/C	167 (40.14)		122 (42.69)	33 (38.37)	
		C/C	20 (4.81)		10 (3.45)	7 (8.14)	
APOE	***rs7412***	**C/C**	287 (68.99)	0.386	214 (73.79)	45 (52.33)	**<0.001**
		C/T	120 (28.85)		72 (24.83)	37 (43.02)	
		T/T	9 (2.16)		4 (1.38)	4 (4.65)	
PCSK9	***rs505151***	**A/A**	227 (54.83)	0.833	156 (53.98)	51 (59.30)	0.669
		A/G	158 (38.16)		112 (38.75)	30 (34.88)	
		G/G	29 (7.01)		21 (7.27)	5 (5.81)	
PCSK9	***rs28362286***	**C/C**	399 (96.38)	**0.028**	278 (96.19)	83 (96.51)	0.151
		C/A	14 (3.38)		11 (3.18)	2 (2.33)	
		A/A	1 (0.24)		0 (0.00)	1 (1.16)	
CETP	***rs17231520***	**G/G**	352 (89.11)	0.892	247 (88.85)	71 (86.59)	0.655
		G/A	42 (10.64)		30 (10.79)	11 (13.41)	
		A/A	1 (0.25)		1 (0.36)	0 (0.00)	
CETP	***rs34065661***	**C/C**	351 (88.86)	0.827	247 (88.85)	71 (86.59)	0.655
		C/G	43 (10.89)		30 (10.79)	11 (13.41)	
		G/G	1 (0.25)		1 (0.36)	0 (0.00)	
CETP	***rs711752***	**G/G**	141 (35.70)	0.251	99 (35.61)	29 (35.37)	0.386
		G/A	199 (50.38)		136 (48.92)	45 (54.88)	
		A/A	55 (13.92)		43 (15.47)	8 (9.76)	
CETP	***rs708272***	**G/G**	142 (35.95)	0.231	100 (35.97)	29 (35.37)	0.373
		G/A	199 (50.38)		135 (48.56)	45 (54.88)	
		A/A	54 (13.67)		43 (15.47)	8 (9.76)	

CETP	***rs5884***	**C/C**	357 (90.38)	0.301	249 (89.57)	76 (92.68)	0.800
		C/A	36 (9.11)		27 (9.71)	6 (7.32)	
		A/A	2 (0.51)		2 (0.72)	0 (0.00)	
CETP	***rs34680782***	**C/C**	384 (97.22)	**0.002**	271 (97.48)	80 (97.56)	1.000
		C/A	10 (2.53)		6 (2.16)	2 (2.44)	
		A/A	1 (0.25)		1 (0.28)	0 (0.00)	
CETP	***rs17231534***	**C/C**	357 (90.38)	0.315	250 (89.93)	76 (92.68)	0.453
		C/A	38 (9.62)		28 (10.07)	6 (7.32)	
		A/A	0 (0.00)		0 (0.00)	0 (0.00)	
CETP	***rs3816117***	**C/C**	158 (40.00)	0.823	115 (41.37)	31 (37.80)	0.768
		C/T	182 (46.08)		123 (44.24)	40 (48.78)	
		T/T	55 (13.92)		40 (14.39)	11 (13.41)	
CETP	***rs561260717***	**C/C**	394 (99.75)	0.980	278 (100)	82 (100)	NA
		C/T	1 (0.25)		0 (0.00)	0 (0.00)	
		T/T	0 (0.00)		0 (0.00)	0 (0.00)	
CETP	***rs34119551***	**T/T**	393 (99.49)	0.960	277 (99.64)	81 (99.78)	0.401
		A/T	2 (0.51)		1 (0.36)	1 (1.22)	
		A/A	0 (0.00)		0 (0.00)	0 (0.00)	
CETP	***rs5030708***	**C/C**	394 (99.75)	0.980	277 (99.64)	82 (100)	1.000
		C/T	1 (0.25)		1 (0.36)	0 (0.00)	
		T/T	0 (0.00)		0 (0.00)	0 (0.00)	
CETP	***rs183782798***	**G/G**	393 (99.49)	0.960	277 (99.64)	81 (99.78)	0.404
		G/T	2 (0.51)		1 (0.36)	1 (1.22)	
		T/T	0 (0.00)		0 (0.00)	0 (0.00)	

FE = Fisher's exact test used when n<5; bold indicates p<0.05 is significant. Freq.= frequency. HWE = Hardy Weinberg Equilibrium

3.5.5) APOE Isoform distribution

The distribution of APOE protein isoforms in the combined T2DM cohort could then be determined by the haplotype of the *APOE rs429358T>C* and *APOE rs7412C>T* genetic variants. The results are summarised in **Table 3.8**. The Ɛ2/Ɛ4 genotype was not seen in our cohort. There was a total of 40 *APOE rs429358T>C* and *APOE rs7412C>T* double heterozygotes in our cohort, for whom the APOE isoform could not be determined using the genotyping methods available to us. These individuals have the possibility of being heterozygous with any combination of the Ɛ1/ Ɛ2/ Ɛ3/ Ɛ4 alleles. Double heterozygotes were excluded from univariate analysis of the APOE genotype but were accounted for when investigating the contribution of the *APOE rs429358C* and *APOE rs7412T* alleles to altered lipid profiles in multivariate analysis. The Ɛ3 isoform is the most common in the indigenous black South African population (65 %), with the Ɛ4 allele (22 %) and Ɛ2 (13 %) allele being less common.

Table 3.8: Distribution of APOE isoform genotypes in the combined South African T2DM cohort.

APOE Genotype	**Combined Cohort** **N (frequency)**
APOE Ɛ2/Ɛ2	9 (0.02)
APOE Ɛ2/Ɛ3	79 (0.19)
APOE Ɛ2/Ɛ4	0 (0.00)
APOE Ɛ3/Ɛ3	141 (0.34)
APOE Ɛ3/Ɛ4	127 (0.31)
APOE Ɛ4/Ɛ4	19 (0.05)
Double Ɛ Heterozygotes	40 (0.09)

3.5.6) Comparison of the global minor allele frequencies of *APOE, PCSK9,* and *CETP* genetic variants

Variant allele frequencies were compared to global population frequencies for additional insight into the population specific prevalence of the genetic variants**.** We report, to the best of our knowledge, on the frequencies and distribution of the *PCSK9 rs505151A>G, CETP rs17231520G>A, CETP rs34065661C>G, CETP rs5884C>A, CETP rs34680782C>A, CETP rs17231534C>A, CETP rs3816117C>T, CETP rs561260717C>T, CETP rs34119551A>T,* and *CETP rs5030708C>T* SNPs for the first time in a black South African population **(Table 3.9**).

APOE, PSCK9, and *CETP* genetic variants' minor allele frequencies (MAFs) from the South African T2DM cohort is compared to data from the 1000 genomes project, GRCh38 reference genome build. The population groups included are the Luyha from Kenya (LWK), the Yoruba from Nigeria (YRI), the Han Chinese from China (CHB) and the Utah residents with Northern and Western European ancestry (CEU). Several SNPs including *PCSK9 rs28362286C>A, CETP rs34065661C>G, CETP rs17231520C>A, CETP rs5884C>A, CETP rs34680782C>A, CETP rs561260717C>T, CETP rs34119551T>A,* and *CETP rs183782798G>T* are present in African populations only. The *CETP rs561260717C>T, CETP rs34119551T>A, CETP rs183782798G>T,* and *CETP rs5030708C>T* were rare in our population with MAF<0.01, thus they were excluded from association analysis.

Table 3.9: Comparison of *APOE, PCSK9,* and *CETP* variant allele frequencies from a South African T2DM population to variant allele frequencies obtained from the 1000 genomes project.

Gene	Variant Allele	South African*	LWK	YRI	CHB	CEU
APOE	*rs429358C*	0.249	0.379	0.236	0.102	0.177
APOE	*rs7412T*	0.166	0.045	0.106	0.107	0.066
PCSK9	*rs505151G*	0.261	0.323	0.306	0.075	0.025
PSCK9	***rs28362286A***	**0.019**	**0.005**	**0.023**	**0.000**	**0.000**
CETP	***rs17231520A***	**0.056**	**0.111**	**0.097**	**0.000**	**0.000**
CETP	***rs34065661G***	**0.057**	**0.116**	**0.097**	**0.000**	**0.000**
CETP	*rs711752A*	0.391	0.303	0.194	0.437	0.449
CETP	*rs708272A*	0.389	0.303	0.194	0.437	0.449
CETP	***rs5884A***	**0.051**	**0.086**	**0.019**	**0.000**	**0.000**
CETP	***rs34680782A***	**0.015**	**0.000**	**0.023**	**0.000**	**0.000**
CETP	*rs17231534A*	0.048	0.131	0.171	0.063	0.040
CETP	*rs3816117T*	0.370	0.409	0.426	0.451	0.490
CETP	***rs561260717T***	**0.001**	**0.005**	**0.000**	**0.000**	**0.000**
CETP	***rs34119551A***	**0.003**	**0.025**	**0.000**	**0.000**	**0.000**
CETP	*rs5030708T*	0.001	0.000	0.000	0.000	0.030
CETP	***rs183782798T***	**0.003**	**0.005**	**0.005**	**0.000**	**0.000**

*Minor allele frequency (MAF) from this study cohort; Bold indicates African specific SNP.

3.5.7) Linkage Disequilibrium (LD) analysis

The degree of LD between the *APOE rs429358T>C* and *rs7412C>T*, the *PCSK9 rs505151G>A and rs28362286C>A,* and the multi-allelic *CETP* SNPs were assessed using r^2 statistics calculated by SHESIS. *APOE, PCSK9,* and *CETP* are all present on different chromosomes so results represent independent analysis of the variants in each gene. The r^2 coefficients fall between 0 -1and are used to measure the degree of linkage disequilibrium between a pair of SNPs. r^2 =1 suggest no recombination has occurred between the two SNPs and that they are in perfect LD.

There is no evidence of LD between the *APOE rs429358T>C* and *APOE rs7412C>T* SNP pair (r^2=0.054) or *PCSK9 rs505151G>A* and *PCSK9 rs28362286C>A* SNP pair (r^2=0.000). The *CETP rs17231520G>A* and *CETP rs34065661C>G*, and *CETP rs711752A>G* and *CETP rs708272G>A* SNP pairs are almost in complete LD with an r^2 coefficient of 0.98 and 0.97, respectively (**Figure3.9**). The *CETP rs708272G>A* was used as a marker for the *CETP rs711752A>G*, while the *CETP rs34065661C>G* SNP was used as a marker for *CETP rs17231520G>A*. Further analysis was reported on *CETP rs708272G>A* and *CETP rs34065661C>G*.

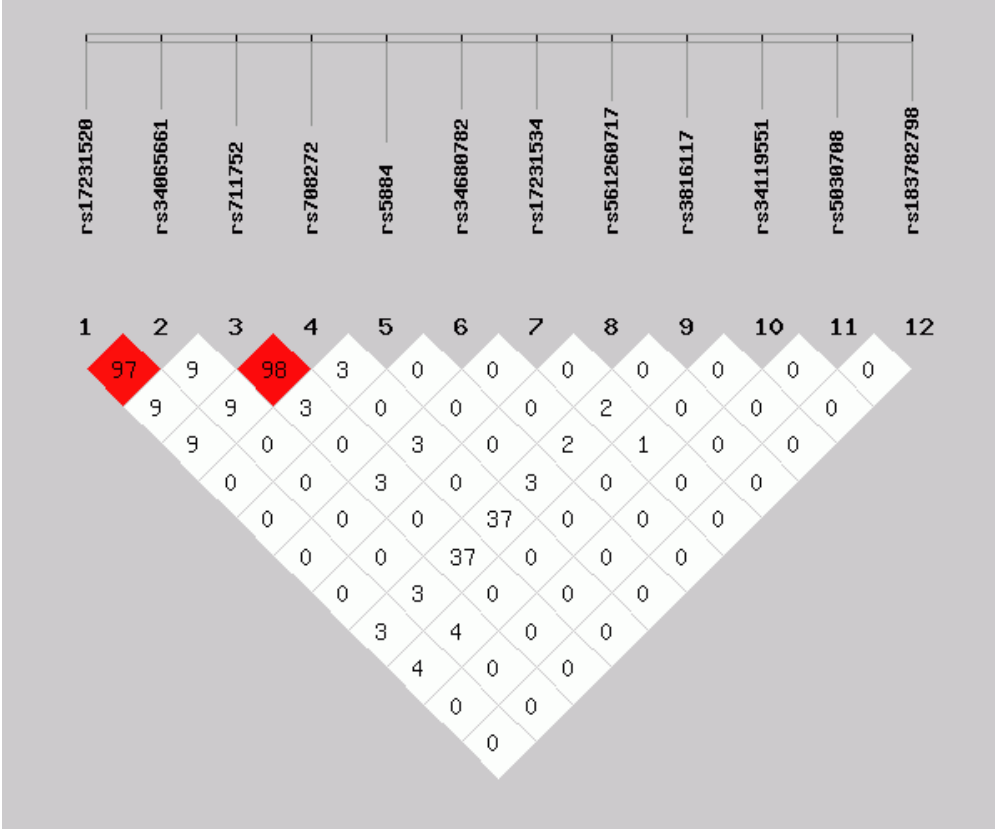


Figure 3.9: LD analysis of *CETP* genetic variant pairs. LD plot showing the r^2 percentage coefficient which indicates the degree of LD between variant pairs. Red blocks indicated SNP pairs with high degrees of LD, while white blocks indicate SNP pairs with low degrees of LD.

3.6) Association of *APOE, PSCK9* and *CETP* genetic variation with altered lipid profiles in the black South African T2DM cohort

We used two methods to investigate the relationship between *APOE, PCSK9,* and *CETP* genetic variations and aberrant lipid profiles of T2DM patients in our study cohort. The primary method was non-parametric univariate investigation using the Mann-Whitney U test and Kruskal-Wallis tests. The effect of each genetic variant on lipid profile, on its own, was investigated in this analysis, with statin usage was accounted for by stratification. However, this method does not take into account the combined effect of multiple genetic variants and relevant multiple cofounder variables. We, thus, set out to investigate the potential impacts of these variable through multivariate linear regression, using a generalised linear model. In

doing so we aimed to identify genetic biomarkers which could potentially predict aberrant lipid profile outcomes.

3.6.1) The impact of *APOE* genetic variation on aberrant lipid profiles

Analysis of *APOE* in our T2DM cohort reveal a significant association between the Ɛ4 allele and lower median HDL-c levels in the combined cohort, compared to the Ɛ3 allele (p=0.034) (**Table 3.10**). Patients with the Ɛ4/Ɛ4 genotype had on average 22.40 % lower median HDL-c levels than those with the more common Ɛ3/Ɛ3 genotype. This association was repeated in the Sim- group (p=0.004), but not in the Sim+ group (p=0.391) after stratification by statin use. However, a non-significant trend to lower HDL-c in the Ɛ4/Ɛ4 genotype was seen in the Sim+ group. Multivariate analysis (**Table 3.16**) showed that the *APOE rs429358C* (Ɛ4) allele confers 11.9 % risk of lower HDL-c levels when adjusting for gender and statin use (OR=0.881, p=1.64e^{-04}).

Results in **Table 3.11** show APOE Ɛ2/Ɛ2 patients having 25.09 % lower median TC and 17.26 % lower median LDL-c than Ɛ3/Ɛ3 patients in the combined cohort (p<0.0001 in both cases). These associations were repeated in the Sim+ group (p=0.027 and p=0.003, respectively) but not in the Sim- group (p=0.127 and p=0.485, respectively). Multivariate analysis (**Table 3.16**) showed that the *APOE rs7412T* (Ɛ2) allele confers a 25.30 % risk of lower TC levels (OR=0.747, p=0.028), and a 24.10 % risk of lower LDL-c levels (OR=0.759, p=0.012) when adjusting for statin use.

Thus, there is enough evidence to postulate that the *APOE* genetic variations do play a significant role in dyslipidaemia in T2DM black South Africans. *APOE rs429358T>C* and *APOE rs7412C>T* are biomarkers of altered HDL-c, and TC and LDL-c in the black South African T2DM population, respectively.

Table 3.10: Univariate analysis of association between APOE Ɛ3 and APOE Ɛ4 alleles, lipid profile and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
APOE Genotype – Combined cohort					
Ɛ3/Ɛ3 (n=141)	6.6 (2.2 – 28.9)	5.34 (1.64 – 10.99)	1.69 (0.33 – 13.67)	**1.25 (0.15 – 2.94)**	3.07 (0.48 – 5.59)
Ɛ3/Ɛ4 (n=127)	6.9 (3.3 – 17.6)	5.25 (2.01 – 13.28)	1.76 (0.40 – 14.53)	**1.14 (0.19 – 2.61)**	2.82 (0.9 – 6.08)
Ɛ4/Ɛ4 (n=19)	6.6 (4.0 – 16.2)	4.45 (2.78 – 7.66)	2.36 (0.62 – 5.81)	**0.97 (0.43 – 2.42)**	2.74 (0.6 – 5.61)
Global P value	0.705	0.229	0.913	**0.034**	0.236
APOE Genotype – Sim+					
Ɛ3/Ɛ3 (n=108)	6.9 (2.2 – 28.9)	5.57 (1.64 – 10.99)	1.86 (0.47 – 13.67)	1.25 (0.15 – 2.94)	3.20 (0.48 – 5.95)
Ɛ3/Ɛ4 (n=96)	7.1 (3.6 – 17.6)	5.74 (2.01 – 13.28)	1.98 (0.40 – 14.53)	1.23 (0.58 – 2.61)	3.34 (0.90 – 6.08)
Ɛ4/Ɛ4 (n=10)	8.7 (5.7 – 16.2)	5.55 (2.78 – 7.66)	2.54 (0.62 – 5.81)	1.13 (0.43 – 2.42)	3.21 (0.60 – 5.61)
Global P Value	0.233	0.610	0.718	0.391	0.656
APOE Genotype – Sim-					
Ɛ3/Ɛ3 (n=17)	5.65 (2.70 – 8.20)	3.75 (2.07-4.96)	1.13 (0.33 – 2.98)	**1.24 (0.67 – 1.60)**	2.01 (1.67 – 2.32)
Ɛ3/Ɛ4 (n=22)	5.40 (3.30 – 11.20)	3.34 (2.35-5.70)	1.34 (0.54 – 4.84)	**0.80 (0.19 – 1.36)**	1.85 (1.61 – 2.10)
Ɛ4/Ɛ4 (n=6)	5.50 (4.00 – 11.80)	3.72 (3.20-4.12)	1.33 (0.66 – 4.45)	**0.97 (0.83 – 1.45)**	1.83 (1.38 – 2.27)
Global P Value	0.802	0.140	0.603	**0.004**	0.674

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c; Sim+ = patients on statins; Sim- = patients not on any statins

Table 3.11: Univariate analysis of association between APOE Ɛ3 and APOE Ɛ2 alleles, lipid profile and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
APOE Genotype – Combined cohort					
Ɛ3/Ɛ3 (n=141)	6.6 (2.2 – 28.9)	**5.34 (1.64 – 10.99)**	1.69 (0.33 – 13.67)	1.25 (0.15 – 2.94)	**3.07 (0.48 – 5.95)**
Ɛ2/Ɛ3 (n=79)	6.3 (2.9 – 18.6)	**4.35 (1.87 – 7.26)**	1.48 (0.44 – 6.03)	1.26 (0.11 – 2.26)	**2.27 (0.76 – 5.08)**
Ɛ2/Ɛ2 (n=9)	6.1 (3.5 – 14.3)	**4.00 (2.65 – 6.96)**	1.23 (0.68 – 3.37)	1.39 (0.56 – 1.85)	**2.54 (1.15 – 4.09)**
Global P value	0.159	**<0.0001**	0.159	0.663	**<0.0001**
APOE Genotype – Sim+					
Ɛ3/Ɛ3 (n=108)	6.9 (2.2 – 28.9)	**5.57 (1.64 – 10.99)**	1.86 (0.47 – 13.67)	1.25 (0.15 – 2.94)	**3.20 (0.48 – 5.95)**
Ɛ2/Ɛ3 (n=47)	6.9 (4.00 – 18.6)	**5.09 (2.50 – 7.26)**	1.62 (0.70 – 6.03)	1.36 (0.76 – 2.26)	**2.45 (1.11 – 5.08)**
Ɛ2/Ɛ2 (n=4)	11.25 (6.10 – 14.3)	**5.61 (5.30 – 6.96)**	2.32 (1.98 – 3.37)	1.74 (0.95 – 1.85)	**3.00 (2.54 – 4.09)**
Global P Value	0.379	**0.027**	0.426	0.555	**0.003**
APOE Genotype – Sim-					
Ɛ3/Ɛ3 (n=17)	5.65 (2.70 – 8.20)	3.75 (2.07 – 4.96)	1.13 (0.33 – 2.98)	1.19 (1.05 – 1.34)	1.91 (0.71 – 3.29)
Ɛ2/Ɛ3 (n=25)	5.60 (2.90 – 17.30)	3.62 (1.87 – 4.88)	1.32 (0.44 – 3.69)	1.01 (0.87 – 1.16)	1.99 (0.76 – 3.09)
Ɛ2/Ɛ2 (n=4)	4.15 (3.50 – 6.10)	3.20 (2.65 – 3.65)	0.83 (0.68 – 1.23)	1.07 (0.43 – 1.71)	1.43 (1.15 – 2.78)
Global P Value	0.439	0.127	0.479	0.225	0.485

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c. Sim+ = patients on statins; Sim- = patients not on any statins

3.6.2) The impact of *PCSK9* genetic variation on aberrant lipid profiles

We reported no significant associations between the *PCSK9 rs505151A>G* variation and lipid profiles through univariate analysis (**Appendix V, Table 5.2**) and multivariate analysis (**Table 3.16)**. This SNP does not play a significant role in our population group and cannot be used as a biomarker for dyslipidaemia.

We report no significant associations between the *PCSK9 rs28362286C>A* variation, using a dominant genetic model, and lipid profiles through univariate analysis (**Table 3.12**) or multivariate analysis (**Table 3.16)**. We do observe a significant association between the *PCSK9 rs28362286C/A+A/A* genotypes, when applying a dominant genetic model, and increased HbA1c% (p=0.03). Significance was not repeated when stratifying for statin usage, but a trend to increased HbA1c% is seen in both groups. The *PCSK9 rs28362286C/A* and *A/A* combined genotypes trended towards significantly higher HDL-c in patients on statins (p=0.078), and higher TG in patients on no statins (p=0.095). The role of this SNP need to be further evaluated in a larger cohort.

Table 3.12: Univariate analysis of association of the *PCSK9 rs28362286C>A* with variant lipid profile and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***PCSK9 rs28362286C>A* Combined**					
C/C (n=399)	**6.50 (2.20 –28.9)**	4.93 (1.64 – 13.28)	1.61 (0.33 – 14.53)	1.23 (0.11 – 3.42)	2.71 (0.48 – 6.08)
C/A + A/A (n=14+1)	**9.0 (5.00 – 18.40)**	5.56 (2.72 – 7.53)	2.06 (1.08 – 3.80)	1.50 (0.80 – 2.32)	2.34 (1.40 – 4.57)
P Value	**0.030**	0.949	0.131	0.166	0.295
***PCSK9 rs28362286C>A* – Sim+**					
C/C (n=277)	6.90 (2.20 –28.9)	5.37 (1.64 – 13.28)	1.79 (0.40 - 14.53)	1.25 (0.15 – 3.42)	3.07 (0.48-6.08)
C/A + A/A (n=11+0)	9.0 (5.10-14.30)	5.78 (3.23 – 7.53)	2.06 (1.08 – 3.80)	1.51 (0.89 – 2.32)	3.19 (1.40-4.57)
P Value	0.507	0.752	0.547	0.078	0.587
***PCSK9 rs28362286C>A* – Sim-**					
C/C (n=83)	5.50 (2.60 – 17.30)	3.67 (1.78 – 7.64)	1.22 (0.33 – 4.84)	1.02 (0.11-2.09)	1.90 (0.71 – 5.17)
C/A + A/A (n=2+1)	7.80 (5.0 – 11.20)	3.49 (2.72 – 4.81)	2.61 (1.40 – 2.76)	0.80 (0.65-1.12)	1.49 (1.43 – 2.34)
P Value	0.223	0.832	0.095	0.365	0.505

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c. Sim+ = patients on statins; Sim- = patients not on any statins

3.6.3) The impact of *CETP* genetic variation on aberrant lipid profiles

The *CETP rs34065661C/G + G/G* genotypes, when applying a dominant genetic model, were associated with increased median HDL-c in our combined cohort and Sim+ cohort (p=0.017 and p=0.026, respectively), and trended towards significantly lower TG levels (**Table 3.13**). Significance was lost in multivariate analysis (**Table 3.16**) when adjusting for the effects of gender, HbA1c%, and statin use. The *CETP rs34065661G* allele approached a significant increase in risk of high LDL-c levels (OR=1.37, p=0.057), when adjusting for statin use.

Analysis of *CETP rs708272G>A* in our study cohort showed no significant associations with altered lipids (**Table 3.14**). The *CETP rs708272A* allele trended towards a significant (p=0.073) increase in HDL-c when compared to the *rs708272G* allele in the combined cohort, but significance was not repeated when stratifying by statin use. Multivariate analysis (**Table 3.17**) confirmed these results.

The *CETP rs3816117T* allele showed a trend towards significantly lower HDL-c levels in the combined cohort (p=0.083) (**Table 3.15**). However, this trend is not observed once stratified by statin use or in multivariate analysis (**Table 3.16**). The *CETP rs34680782A* allele was significantly associated with 85.2 % increased risk of higher LDL-c in multivariable analysis, adjusting for statin use (OR=1.85, p=0.022). There were no significant associations in the *CETP rs34680782C>A* univariate analysis, but a non-significant trend towards higher LDL-c was seen in the *CETP rs34680782C/A + A/A* genotypes when applying a dominant genetic model (**Appendix V**). No other significant associations were found for the other genotyped *CETP* SNPs (**Appendix V**).

Table 3.13: Univariate analysis of association of the *CETP rs34065661C>G* variant with lipid profile and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***CETP rs34065661C>G* Combined**					
C/C (n=351)	6.60 (2.20 –18.40)	4.90 (1.64 – 11.32)	**1.74 (0.33 – 14.53)**	**1.20 (0.11 – 3.42)**	2.71 (0.48 – 6.08)
C/G + G/G (n=43+1)	7.30 (3.10 – 17.30)	4.93 (2.97– 8.03)	**1.41 (0.54 – 4.96)**	**1.37 (0.22 – 2.64)**	2.64 (1.46 – 5.17)
Global P Value	0.294	0.787	**0.068**	**0.017**	0.831
***CETP rs34065661C>G* – Sim+**					
C/C (n=247)	6.90 (2.20 –17.60)	5.38 (1.64 – 11.32)	**1.95 (0.40 - 14.53)**	**1.24 (0.15 – 3.42)**	3.06 (0.48-6.08)
C/G + G/G (n=30+1)	8.20 (5.10-16.70)	5.43 (3.38 – 8.03)	**1.48 (0.59 – 4.96)**	**1.43 (0.22 – 2.64)**	3.02 (1.58-5.05)
Global P Value	0.222	0.507	**0.085**	**0.026**	0.593
***CETP rs34065661C>G* – Sim-**					
C/C (n=71)	5.50 (2.60 – 12.10)	3.67 (1.78 – 5.70)	1.32 (0.33 – 4.84)	0.99 (0.11-1.87)	1.91 (0.71 – 3.29)
C/G + G/G (n=11+0)	5.65 (3.10 – 17.30)	3.58 (2.97 – 7.64)	1.21 (0.54 – 2.81)	1.23 (0.27-1.71)	1.91 (1.46 – 5.17)
Global P Value	0.341	0.967	0.586	0.462	0.581

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c. Sim+ = patients on statins; Sim- = patients not on any statins

Table 3.14: Univariate analysis of association of the *CETP rs708272G>A* variant with lipid profile and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***CETP rs708272G>A* - Combined**					
G/G (n=142)	6.60 (3.6 – 18.40)	4.82 (1.87 – 7.76)	1.73 (0.40 – 13.67)	**1.17 (0.19 – 2.51)**	2.69 (0.60 – 5.54)
G/A (n=199)	6.80 (2.20 – 17.60)	4.94 (1.78 – 11.32)	1.62 (0.33 – 14.53)	**1.24 (0.11 – 3.42)**	2.72 (0.48 – 6.08)
A/A (n=54)	6.45 (2.70 – 16.70)	5.11 (1.64 – 7.72)	1.62 (0.59 – 5.81)	**1.24 (0.59 – 2.94)**	2.67 (0.68 – 5.46)
Global P value	0.948	0.465	0.864	**0.073**	0.849
***CETP rs708272G>A* – Sim+**					
G/G (n=100)	6.90 (3.60 – 16.70)	5.22 (2.11 – 7.76)	1.87 (0.40 – 13.67)	1.24 (0.19 – 2.51)	3.03 (0.60 – 5.54)
G/A (n=135)	7.10 (2.20 – 17.60)	5.56 (2.01 – 11.32)	1.79 (0.56 – 6.03)	1.29 (0.15 – 3.42)	3.18 (0.48 – 6.08)
A/A (n=43)	6.80 (4.40 – 16.70)	5.35 (1.64 – 7.72)	1.82 (0.59 – 5.81)	1.31 (0.59 – 2.94)	2.91 (0.68 – 5.46)
Global P Value	0.379	0.659	0.738	0.252	0.351
***CETP rs708272G>A* – Sim-**					
G/G (n=29)	5.40 (3.90 – 12.10)	3.66 (1.87 – 4.90)	1.24 (0.55 – 3.69)	0.93 (0.33 – 1.87)	1.89 (0.71 – 3.09)
G/A (n=45)	5.55 (2.60 – 17.30)	3.64 (1.78 – 5.70)	1.23 (0.33 – 4.84)	1.12 (0.11 – 1.57)	1.92 (0.97 – 3.29)
A/A (n=8)	5.40 (2.70 – 11.10)	3.97 (2.60 – 7.64)	1.65 (0.66 – 2.17)	1.03 (0.67 – 1.71)	2.32 (1.29 – 5.17)
Global P Value	0.986	0.675	0.844	0.694	0.306

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin Sim+ = patients on statins; Sim- = patients not on any statins

Table 3.15: Univariate analysis of association of the *CETP rs3816117C>T* variant with lipid profile and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***CETP rs3816117C>T* - Combined**					
C/C (n=158)	6.50 (2.20 – 16.70)	5.03 (1.64 – 11.32)	1.49 (0.47 – 14.53)	**1.27 (0.11 – 2.94)**	2.76 (0.68 – 5.95)
C/T (n=181)	6.90 (3.30 – 18.40)	4.88 (1.78 – 10.99)	1.78 (0.33 – 7.37)	**1.19 (0.11 – 3.24)**	2.68 (0.48 – 6.08)
T/T (n=55)	6.10 (3.60 – 16.70)	4.78 (2.07 – 7.53)	1.64 (0.55 – 13.67)	**1.17 (0.58 – 2.51)**	2.59 (0.71 – 5.54)
Global P value	0.162	0.498	0.418	**0.083**	0.685
***CETP rs3816117C>T* – Sim+**					
C/C (n=115)	7.00 (2.20 – 16.70)	5.54 (1.64 – 11.32)	1.63 (0.47 – 14.53)	1.33 (0.15 – 2.94)	3.08 (0.68 – 5.95)
C/T (n=123)	7.10 (4.00 – 17.60)	5.28 (2.01 – 10.99)	1.98 (0.40 – 7.37)	1.24 (0.18 – 3.42)	3.03 (0.48 – 6.08)
T/T (n=40)	6.60 (3.60 – 16.70)	5.19 (2.11 – 7.53)	1.83 (0.55 – 13.67)	1.25 (0.58 – 2.51)	3.04 (1.21 – 5.54)
Global P Value	0.596	0.505	0.435	0.145	0.773
***CETP rs3816117C>T* – Sim-**					
C/C (n=31)	5.42 (2.60 – 11.20)	3.69 (2.35 – 4.90)	1.39 (0.54 – 4.84)	1.00 (0.11 – 1.71)	1.91 (0.97 – 5.17)
C/T (n=40	5.56 (3.30 – 17.30)	3.70 (1.78 – 5.70)	1.21 (0.33 – 3.38)	1.03 (0.11 – 1.87)	1.92 (0.79 – 3.29)
T/T (n=11)	5.00 (3.90 – 11.10)	3.50 (2.07 – 4.64)	1.40 (1.02 – 3.69)	0.93 (0.65 – 1.43)	1.86 (0.71 – 3.03)
Global P Value	0.456	0.532	0.457	0.839	0.585

Values represent median (range); Bold indicates significance p<0.05; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin
Sim+ = patients on statins; Sim- = patients not on any statins

Table 3.16: Multivariate analysis of association between *APOE, PCSK9* and *CETP* genetic variations and lipid profiles in the combined South T2DM cohort, by multivariable linear regression, using a generalised linear model.

SNP/Allele	Total Cholesterol@			Triglycerides#			LDL-c@			HDL-c^		
	OR	**95%CI**	**P value**	**OR**	**95%CI**	**P value**	**OR**	**95%CI**	**P value**	**OR**	**95%CI**	**P value**
APOE rs429358C	0.962	0.794 - 1.166	0.742	1.080	0.858 -1.36	0.584	0.986	0.839 - 1.159	0.887	**0.881**	**0.815 - 0.952**	**1.64e^{-04*}**
APOE rs7412T	**0.747**	**0.601 - 0.928**	**0.028**	0.786	0.607 - 1.02	0.124	**0.759**	**0.634 - 0.909**	**0.012**	0.970	0.890 - 1.056	0.554
PCSK9 rs505151A	0.890	0.748 - 1.060	0.273	0.910	0.736 - 1.12	0.464	1.010	0.873 - 1.168	0.910	0.989	0.921 - 1.062	0.794
PCSK9 rs28362286A	1.306	0.818 - 2.085	0.349	1.131	0.647 - 1.98	0.717	1.082	0.737 - 1.588	0.735	1.009	0.835 - 1.219	0.939
CETP rs34065661G	1.328	0.955 - 1.847	0.158	0.696	0.467 - 1.04	0.136	**1.373**	**1.045 - 1.804**	**0.057**	1.063	0.929 - 1.216	0.459
CETP rs708272A	1.171	0.956 - 1.434	0.202	1.122	0.878 - 1.43	0.441	1.047	0.885 - 1.239	0.654	1.049	0.966 - 1.139	0.342
CETP rs5884A	1.385	1.006 - 1.907	0.095	1.044	0.709 - 1.54	0.854	1.228	0.936 - 1.612	0.214	1.102	0.961 - 1.263	0.244
CETP rs34680782A	**1.860**	**1.085 - 3.188**	**0.059**	0.737	0.388 - 1.40	0.436	**1.852**	**1.191 - 2.882**	**0.022**	0.964	0.776 - 1.198	0.783
CETP rs17231534A	1.298	0.916 - 1.839	0.219	1.514	0.994 - 2.31	0.106	0.979	0.733 - 1.307	0.903	1.032	0.896 - 1.190	0.712
CETP rs3816117T	1.081	0.885 - 1.320	0.523	1.269	0.996 - 1.62	0.107	1.040	0.880 - 1.230	0.700	0.964	0.889 -1.045	0.451

N=369 due to missing data; *= random shuffling of p value 10000x; Bold indicated statistically significant p values (<0.05). Total cholesterol = TC; Triglycerides = TG; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; BMI = Body mass index; HbA1c = Haemoglobin A1C. OR = odds ratio. CI = Confidence Interval. @=adjusted for statin usage; #=adjusted for Age, statin use, BMI, and HbA1c%; ^= adjusted for gender, statin use, and HbA1c%.

3.6.3.1) Association of *CETP* haplotype with HDL-c levels

A haplotype analysis was performed using data from three *CETP* SNPs, namely, *CETP rs34065661C>G, CETP rs3816117C>T,* and *CETP rs708272G>A,* which showed significant associations and/or trends with altered HDL-c. This was done in order to determine if having a combination of previously associated potentially pathogenic *CETP* alleles could have a greater contribution to altered HDL-c. For this analysis individuals were spilt into two haplogroups using a dominant genetic model (at least one allele present): those with a combination of the HDL-c raising genotypes *rs34065661C/G + C/C, rs708272G/A + A/A,* and *rs3816117C/C* vs those with a combination of the HDL-c lowering genotypes *rs34065661G/G, rs708272G/G,* and *rs3816117C/T + T/T.*

Patients with the *rs34065661C/G + C/C, rs708272G/A + A/A,* and *rs3816117C/C* genotypes had significantly higher levels of HDL-c (p=0.001) than those with the *rs34065661G/G, rs708272G/G,* and *rs3816117C/T + T/T* genotypes in the combined cohort (**Figure 3.10**). The association was repeated in patients on statin therapy (p=0.011), but not in patients not on statin therapy (p=0.1402). However, there was an observable trend towards lower HDL-c in this group. The *CETP* haplotype groups had no significant associations with other altered lipids (p >0.10 in all cases). Thus, we can see that a combination of the *CETP rs34065661C>G, CETP rs3816117C>T,* and *CETP rs708272G>A* SNPs may be biomarkers ofaltered HDL-c in the black South African population.

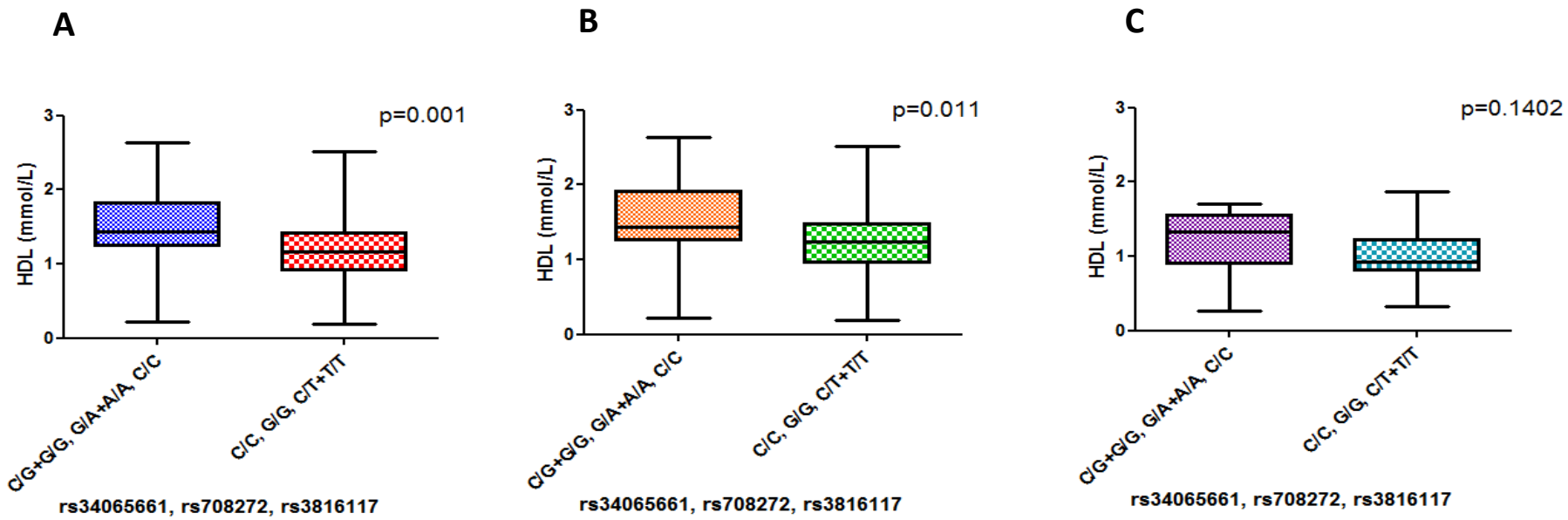


Figure 3.10: Haplotype analysis of *CETP* HDL-c associated variants. Box and whisker plot showing the comparison of individuals with a combination of the HDL-c raising genotypes *rs34065661C/G + C/C*, *rs708272G/A + A/A* and *rs3816117C/C* vs those with a combination of the HDL-c lowering genotypes *rs34065661G/G*, *rs708272G/G*, and *rs3816117C/T + T/T* in: **A**) the combined T2DM study cohort, **B**) individuals on statins, and **C**) individuals not on any statin therapy.

Chapter 4: Discussion and Conclusion

In this study we are primarily interested in identifying genetic factors associated with abnormal lipid profiles of black South African patients with T2DM. Thus, we carried out genetic research to identify population-specific genetic markers that are associated with altered lipid metabolism in indigenous black South African patients with T2DM. As the lifestyles and diets of the South African population change, it is crucial to understand the role of genetic predisposition to dyslipidaemia.

We aimed to determine the prevalence of known genetic variations and to identify novel population-specific variations, which affect lipid metabolism, to better understand the impact of genetics on dyslipidaemia in a South African T2DM population. Ultimately, we hope to identify a panel of lipid metabolism-relevant genetic variations which can be used to better inform clinicians in the diagnosis of dyslipidaemia and aid in the application of appropriate management, and preventative strategies. In this Chapter, the results of our study are further interrogated in the context of previously published literature. We discuss the clinical and demographic features of our study cohort and the qualitative and quantitative differences in allele distribution of the studied genes. Most importantly, we discuss the role of the studied genetic variants on altered lipid metabolism and determine if they are useful as biomarkers.

4.1) Clinical and demographic characteristics of study participants

All of the patients included in this study were black South Africans of Bantu origin. This was important as there is little data on genetic biomarkers of dyslipidaemia in black South Africans. Also, equally important, is the fact that genetic markers are very often population specific due to evolutionary and geographical history. Thus, inclusion of other South African population groups would complicate analysis, requiring population stratification.

The clinical and demographical information of the combined cohort (n= 417) is summarised in **Table 3.1.** A number of patients had missing clinical information which was taken into account in further analysis. The median age of the combined cohort was 59 years old, with ages ranging from 19 to 91 years. This is expected, as the risk of developing T2DM has been shown to increase with age, possibly explaining the larger proportion of patients over 40

years old. A meta-analysis of 77 studies on the epidemiology of T2DM in China found that the prevalence rate of T2DM increased with age[132]. The prevalence rate was six to sevenfold higher in those aged 55-74 years compared to those aged 20-34 years[132]. The same study found an increased prevalence of T2DM in women at 11.6 % (95% CI = 10.0–13.1%), than in men at 9.9 % (95 % CI = 8.8–11. 0%)[132]. In comparison, the reported incidence of diabetes in South African is 7.7 % in men and 11.8% in females[115]. This may explain the higher proportion of women (65.04 %) than men (34.96 %) in our cohort. It must be noted that these reported incidence rates may be due to differences in lifestyle factors, such as physical activity, which are also T2DM risk factors.

The median BMI of our cohort was 29.5 kg/m^2. Of the study cohort 136 (32.61 %) patients were classified as overweight (BMI 25- 29.9 kg/m^2). Even more worryingly, 183 (43.88 %) patients presented with a BMI of 30.0 kg/m^2 or greater, which the WHO classifies as obese. This proportion of obese patients (43.88 %) is higher than the reported 25 % of the general South African population by the WHO, but was similar to the incidence of obesity in the Heart of Soweto study (43 %)[114,115]. The number of obese T2DM patients in this South African Cohort is also similar to the proportion reported Centres for Disease Control and Prevention (CDC) National Diabetes Statistics Report (45.3 %)[133]. Thus, T2DM patients tend to have higher BMIs, or a high BMI increases your risk of T2DM. This may also be due to lifestyle factors such as lack of physical activity and poor diet. Only 82 (21.11 %) of the cohort self-reported as smoking regularly. This number may be underestimated as patients often omit this information on self-reporting questionnaires. No information was available on alcohol usage as the clinical recruiters thought the information given by the patients was unreliable, thus it was not recorded in patient files.

The median HbA1c% of the study cohort was 6.60, with a wide range of between 2.20 and 28.90. Also, 212 (50.84 %) patients had HbA1c% $\geq$ 6.5, which is a strong indicator of T2DM. Stable HbA1c formation is limited to approximately 100-120 days, the lifespan of erythrocytes. Thus, HbA1c% is a reflection of the average blood glucose level of the previous 2 to 3 months and is a useful parameter for monitoring long-term blood glucose control in T2DM patients[122]. Rapid changes to blood glucose levels can be detected as the preceding 30 days contribute more to HbA1c% than the earlier 90 – 120 days[105].

All of the patients in this cohort were sent to the clinic with a confirmed T2DM diagnosis, with some recruited patients being regular visitors over a long period of time and others on

their first visit. Thus, in this study HbA1c% is an indicator of how well a patient is controlling their blood glucose levels, which is affected by T2DM treatment adherence and response. Patients with very high HbA1c% (e.g. 28.9 %) may not be adhering or responding to treatment or may be first time visitors needing further treatment. However, Insulin and Metformin treatment status, as well as adherence data was not made available in this study, so this could not be determined.

4.2) Lipid profile analysis of the study cohort

Four groups of distinct lipid profiles: Group 1 (high TC), group 2 (high TG), group 3 (mixed dyslipidaemia) and group 4 (normal lipid levels) were included to ensure that a there was a wide range of lipid traits present in the cohort, to tease out any genetic associations. Thus, the overall combined lipid profile of the study cohort is a reflection of this non-random selection.

The lipid profiles of this study cohort were significantly different to the results published from the Heart of Soweto study, which examined the lipid profiles of 1832 individuals of African descent, among others. For the sake of this comparison, the definition of dyslipidaemic traits was taken from the Heart of Soweto study[114]. The proportion of dyslipidaemic TC, TG, and LDL-c patients is higher in our study, while the number of patients with HDL-c <1.0 is approximately 2 times lower in our study. However, it must be noted that the patients in the Heart of Soweto study were included based on suspected CVD diagnosis, while our study consisted of T2DM patients. Only 9 % of the Heart of Soweto study patients had T2DM. Thus, there are several co-morbidities or influencing factors that differ between the two studies, which could account for the vast difference in dyslipidaemia proportions. Also, the cohort in our study was specifically selected based on lipid traits.

This comparison provides useful insights into the prevalence of dyslipidaemia in black South Africans. While African populations were traditionally found to have favourable lipid profiles, there is clear evidence that this is changing and the burden of dyslipidaemia in these population groups is growing. This has dire implications for CVD risk profiles, especially in patients with co-morbidities such as T2DM. The effect of these factors on lipid profiles will be discussed further.

4.3) Statin therapy in the study cohort and its implications

There are a number of issues that come with investigating dyslipidaemia in a T2DM cohort. The most apparent is the confounding effect that can be seen with the patients being treated with statins. Dyslipidaemia in diabetics is very common. It has been estimated that 60-70 % of T2DM patients present with some form of aberrant lipid profile[109]. In our study cohort 291 patients (77 %) were on statin (mostly simvastatin) treatment regimes. There is a close relationship between glucose and lipid metabolism, as both processes are regulated in the liver. The molecular mechanisms are not fully understood, but the association between T2DM and dyslipidaemia is well documented. T2DM are already at greater risk for developing more severe CVD (see **Chapter 1.5**). CVD accounts for 70 % of deaths in T2DM patients[105]. Dyslipidaemia is in itself a major CVD risk factor (See **Chapter 1.4**), thus T2DM patients are regularly prescribed lipid lowering statin therapies as CVD prophylaxis.

In South Africa, statin therapy is recommended for high risk T2DM patients, regardless of their baseline lipid profiles, who are over the age of 40 years or who have had diabetes for more than 10 years with one or more additional CVD risk factors (Smoking, hypertension, etc.)[105]. In lower risk T2DM patients, i.e. those younger than 40 years and who have had diabetes for less than 10 years with no history of CVD or chronic kidney disease, statin therapy is recommended in those with LDL-c levels > 1.8 mmol/l[105]. In our cohort, 321 patients of 397 with LDL-c levels available (81 %) have LDL-c > 1.8 mmol/l. However, not all of these patients are on statin treatment regimes. The patients, who are at high CVD risk due to their T2DM status, may not be on statin therapy regiments at the time of recruitment as they may be new comers to the Diabetes clinic. An advantage of a study of this nature is the possibility to compare the effects of additional environmental and genetic factors on lipid profiles in individuals on statins and on those not on statins.

The use of lipid lowering therapies was an immediately identifiable confounding factor. Statins, or 3-hydroxy-3-methylglutarylcoenzyme A (HMG-CoA) reductase inhibitors, have a well-documented effect on lipid profiles, especially with prolonged use[46]. The lowering of LDL-c by statins is dose dependent and is summarised in a comprehensive review and meta-analysis by Collins *et al.* An average reduction of LDL-c of 23 % can be achieved with 5mg of simvastatin a day (low intensity regimen)[46]. As dosage is doubled, an approximate reduction of 6 % in average LDL-c can be achieved, resulting in a 37 % reduction of LDL-c with 40 mg a day[46,134].

Thus, one would expect to find lower cholesterol levels in patients who are on statin treatment, especially those on long term or high intensity regimens. However, this is not what was observed when stratifying the current study cohort by statin use and comparing median lipid profiles. A statistically significant increased median TC of 32.47 % (p = 0.0001), TG of 33.87 % (p = 0.0001), LDL-c of 38.31 % (p = 0.0001), and HDL-c of 21.09 % (p = 0.0001) was observed in those on statins than compared to those not on statins. This was confirmed by multivariate analysis (**Table 3.6**). This result is seemingly counterintuitive. A possible explanation is that the individuals who had, on average, the least favourable lipid profiles at the time of the recruitment were most likely put on statin treatment regimens (Sim+). The individuals who were not on statins (Sim-) may not have needed to be, due to more favourable lipid profiles or the fact that they were new to the clinic. Thus, the more dyslipidaemic patients were on statins, which accounts for the higher median cholesterol levels for patients on statin treatment compared to those not on statins. The efficacy of statins in reducing the lipid levels for those on treatment would be affected by response to the drug, as well as drug adherence. Therapy response can be influenced by a multitude of clinical factors, but also by genetic variation in drug metabolism pathways. This is supported by the large range of lipid values seen in this group.

The Sim+ group had higher median HbA1c% levels compared to the Sim- group (6.9 % vs.5.5 %, respectively) and were, thus, poorer controllers of blood glucose (p=0.0001). Additionally, the Sim+ had a significantly higher median age (59 years and 50.5 years; p=0.0002), and higher median BMI approaching significance (30.1 kg/m^2 and 29.4 kg/m^2; p=0.058) compared to the Sim- group. Again, this could be explained by the statin therapy recommendation for all T2DM patients over the age of 40[105].

Treatment status has a significant effect on observed lipid profiles and was needed to be accounted for in the further analysis of the potential genetic contribution to dyslipidaemia. One possibility would be to artificially raise the LDL-c values in the Sim+ group by a conservative 23 % based on the reported lowering effects of statins in the literature[46]. However, statin dosage, response to treatment, and adherence data affects response as well. Another option would be to convert the continuous lipid profile values into the categorical groups of either “Dyslipidaemic” or “Non-dyslipidaemic” based on defined cut-offs, with those on statins automatically grouped into the “Dyslipidaemic” regardless of actual lipid levels. You would then assess each of the independent variables for risk of developing a dyslipidaemic trait by logistic regression. This has the added benefit of resolving the

nonnormality issues for the dependent lipid variables. However, this method was not performed in this study as using the cut-offs for ideal lipid profiles in T2DM patients as defined by the Society for Endocrinology, Metabolism and Diabetes of South Africa guidelines would result in the "Non-dyslipidaemic" groups having too small sample sizes in comparison with the "Dyslipidaemic" groups.

Most importantly, in complex diseases, the effect size of single SNPs on phenotypic traits is usually small as these disorders are highly polygenic. The potential effect of the SNPs on modifying lipid traits within these groups may be undetectable if the effect size is not large enough, especially with small sample sizes like in this study. Thus, in the further analysis of genetic variations on altered lipid profiles we stratified by statin use to help account for the observed effects. Statin therapy status is an indicator of TC, TG, and LDL-c dyslipidaemia.

4.4) The associations between clinical characteristics, demographic features, and aberrant lipid profiles

We used two methods to determine if there was any association between the clinical or demographic features of our cohort and altered lipid profiles. We used the non-parametric Spearman's rank test of correlation for univariate analysis and then carried out multiple linear regression, using the generalised linear model method described in **Chapter 2.16.7** to confirm previous analysis. The combination of analysis methods was used to tease out the true effects of these variables on the lipid profiles in our cohort, which can be compared to the published literature to better inform subsequent genetic analysis.

The female gender was significantly positively correlated with older age (Rho=0.102, p=0.048), which may reflect the greater proportion of elderly and female participants in our study. Women had increased HDL-c levels in both univariate (Rho=0.110, p=0.034) and multivariate analysis (OR=1.12, p=0.044). This agrees with the published literature. One randomized, double blind dietary trial found that women had greater elevations of HDL-c in response to dietary supplementation than men do[135]. This may be due to women channelling dietary cholesterol in HDL-c more efficiently than men do, or the fact that they have increased lipoprotein lipase activities[135].

Older age was correlated with increased HDL-c levels (Rho=0.136, p=0.009), but this association was not repeated in multivariate analysis when adjusting for multiple variables

(p=0.328). However, older age was associated with very slight risk of decreased TG (OR=0.98, p=0.003). TG levels are inversely correlated with HDL-c levels. Thus, the observed correlation of old age with increased HDL-c may be due to there being more females of older age in the cohort. There is not enough evidence to say that age plays a major role in dyslipidaemia in this cohort. Older age was additionally correlated with increased HbA1c% (Rho=0.117, p=0.025) and statin use (Rho=0.206, p=0.001). This is not surprising as age is associated with increased incidence of T2DM, and statins are recommended in all T2DM over the age of 40 years[105,132].

Smoking status was not significantly correlated with any other clinical variables or altered lipid profiles. However, current smokers were found to be less likely to have high TC (OR=0.72; p<0.001), high LDL-c (OR=0.68; p<0.001) and more likely to have low HDL-c (OR=1.47; p<0.001) in the Heart of Soweto study[114]. It is possible that the smoking status of each patient might not be accurate as it was self-reported. The role of smoking in dyslipidaemia in our cohort needs to be further investigated and should still be accounted for in future studies due to reported associations in the literature and biological significance to CVD and T2DM[12].

Patients with greater BMIs were significantly correlated with increased TG levels in both univariate and multivariate analysis (Rho=0.159, p=0.002; OR=1.03, p=0.011). BMI was not associated with any other lipids in our study. However, BMI was moderately associated with increased TG, TC and LDL-c in the Heart of Soweto study[114]. Thus, our observation of the role of BMI only partly agree with previously published results, but BMI is important in dyslipidaemia and should be still accounted for in future studies.

Poorly controlled T2DM (measured by increased HbA1c%) was significantly correlated with increased TG (Rho=0.378, p<0.001; OR=1.11, p=3.26e^{-05}) and very slightly decreased HDL-c (OR=0.99, p=3.28e^{-05}). It is reported that the most frequent aberrant lipid profiles encountered among T2DM patients are moderately increased TG, and LDL-c and decreased HDL-c[105]. Elevated TG and TC are proxies for elevated remnant VLDL and CM particles, which are denser and highly atherogenic. LDL-c particles are smaller and denser so there are higher concentrations in TC than suggested by the Friedewald equation derived LDL-c values[105]. Thus, poor metabolic controllers need to be closely monitored in our population.

4.5) *APOE, PSCK9* and *CETP* genetic variation distribution

The distribution of selected *APOE, PSCK9* and *CETP* genetic variations was investigated in the study cohort, following successful genotyping. Genotype frequencies were evaluated for deviation from HWE and compared using the between the Sim+ and Sim- groups in order to determine if either of the groups had a higher frequency of potentially lipid modulating variations.

Two identified variants, *PCSK9 rs28362286C>A* (p=0.028) and the *CETP rs34680782C>A* (p=0.002), deviated from HWE. Usually this is an indication of genotyping errors, however, these two variants were genotypes for by direct cycle Sanger sequencing, so this is unlikely. A more likely explanation is that the assumptions of HWE were not met. The study cohort was not completely randomly selected. Samples were included based on the four lipid profile groups as explained in **Chapter 2.6**. This could enrich the frequency of variant SNPs which may be associated with altered lipids. The other explanation is that these SNPs are under selection forces in this population group.

Genetic variant minor allele frequencies (MAF) were compared to the global frequencies of the Luyha from Kenya (LWK), the Yoruba from Nigeria (YRI), the Han Chinese from China (CHB) and the Utah residents with Northern and Western European ancestry (CEU) (**Table 3.9**). Seven variants are common in all of the observed population groups. These variants may play a common role in dyslipidaemia across populations. Conversely, five variants with MAF>0.01 were found in only African populations, with similar frequencies between African ethnic groups. The Heart of Soweto study reported significantly different lower ($p < 0.01$) TC, LDL-c and TG levels in individuals of African descent compared to white, mixed ancestry and Indian ancestry South Africans[114]. These differences are possibly due to unique lipid altering population specific genetic variants. Thus, we are especially interested in the African specific variants as they may explain this variance in lipid profiles between the population groups.

APOE rs7412C>T, had a statistically significant difference in genotype frequency between the Sim+ and Sim- groups (p<0.001). A higher proportion of *APOE rs7412C/T* and *APOE rs7412T/T* individuals were observed in the Sim- group. The *APOE rs7412T* variant, which is a proxy for the APOE Ɛ2 isoform, has been consistently associated with altered lipid profiles[69]. Thus, the *APOE rs7412C>T* polymorphism may contribute to the observed

differences between the lipid profiles of Sim+ and Sim- and will be discussed in detail further.

4.6) *APOE, PSCK9* and *CETP* linkage disequilibrium analysis

LD analysis was performed on *APOE, PSCK9* and *CETP* variants separately as each of these genes are on different chromosomes. LD analysis gives us genome level insight into a populations evolutionary history, accounting for natural selection and geographical subdivision events[130]. There was no evidence of LD between the *APOE rs429358T>C* and *APOE rs7412C>T* SNP pair (r^2=0.054). This agrees with data from the 1000 genomes project which reports no evidence of LD between these SNPs in the YRI, CEU, or CHB population groups (r^2=0.000 in all)[136]. The *PCSK9 rs505151A>G* and *PCSK9 rs28362286C>A* SNP pair had no evidence of LD in our cohort (r^2=0.000), or in the YRI population group (r^2=0.000)[136]. The *PCSK9 rs28362286C>A* SNP is only found in African populations.

The *CETP rs17231520G>A* and *rs34065661C>G*, and the *CETP rs711752A>G* and *rs708272G>A* SNP pairs were almost in complete LD (r^2=0.98 and r^2=0.97, respectively (**Figure 3.9**). There was no evidence of significant LD between any of the other *CETP* SNPs. Our results are consistent with a study done by Pirim *et al.*, which sequenced the *CETP* gene in Caucasian and African populations. The two *CETP* SNP pairs were in complete LD in both population groups. They found a much higher degree of LD between more *CETP* SNPs in the Caucasian group than in the African group[97]. It is well established from studies on mitochondrial, genomic and sex-chromosome DNA that African populations are the more genetically variable due to maintaining larger populations over longer periods time[137]. This allows the effects of genetic drift, recombination and mutation to be fixed in these populations and, thus, African populations tend to have smaller LD blocks, while bottleneck and founder effects in contemporary European and Asian populations has led to larger LD blocks in these groups[137].

The confirmation of complete LD of the two *CETP* SNP pairs in the South African population is useful as it means that analysis of the effect on lipid profiles only needs to be done on one SNP in each group. If two alleles are in complete LD, then one of the alleles can be used as a genetic marker of the other. Similarly, an allele in complete LD with a defining haplotype can be used as a marker of that haplotype. Thus, you can infer the allele frequency of the other alleles by genotyping only one marker allele. This has implications for gene

mapping of causative alleles. A SNP may be consistently linked to a disease phenotype but may have no predicted structural or regulatory effects. The SNP may be in LD with the real causative SNP, so it is still useful to genotype for it as a biomarker. This is important in the case of a gene with a hotspot of reportedly associated SNPs in a small genomic region such as *CETP*[97]. Thus, in further analysis *CETP rs708272G>A* was used as a marker for the *CETP rs711752A>G*, while the *CETP rs34065661C>G* SNP was used as a marker for *CETP rs17231520G>A*.

4.7) The association between *APOE*, *PSCK9* and *CETP* genetic variations and aberrant lipid profiles

We used two methods to investigate the relationship between genotyped *APOE, PCSK9,* and *CETP* genetic variations and aberrant lipid profiles of T2DM patients in our study cohort. The primary method was non-parametric univariate investigation using the Mann-Whitney U test and Kruskal-Wallis tests, while adjusting for statin usage by stratification. We then carried out multiple linear regression, using the generalised linear model method described in **Chapter 2.16.7** for reasons discussed in this chapter. This was done to confirm previous analysis and to potentially reveal significant associations hidden by confounding factors. We discuss the role of each variant and its implications further here.

4.7.1) The impact of *APOE* genetic variation on aberrant lipid profiles

The *APOE rs429358T>C* change encodes a destabilising Cys122Arg amino acid change (ε4) in the helix of the short hinge region that connects the C and N-terminal domains of APOE. It is thought that this changes the interaction of the domains which affects lipid binding affinity[69]. The ε4 isoform binds with better affinity to VLDL, decreasing its clearance and reducing the number of VLDL components available for incorporation into HDL-c[69]. This was shown in APOE knockout mice, which had lower HDL-c in those with the ε4 isoform compared to ε3[74]. Additionally comprehensive meta-analysis showed a weak inverse relationship between HDL-c and the *APOE* (ε2/ε2, ε2/ε3, ε2/ε4, ε3/ε3, ε3/ε4, ε4/ε4) genotypes, as well as a strong linear association between the genotypes in this order and LDL-c[77].

Analysis of APOE isoforms in our T2DM cohort agree with the literature and reveal a significant association between the ε4 isoform and lower median HDL-c levels in the combined cohort, compared to the ε3 isoform (p=0.034) (**Table 3.10**). Patients with the ε4/ ε4 genotype had lower median HDL-c levels than those with the more common ε3/ε3 genotype. This association was repeated in the Sim- group (p=0.004), but not in the Sim+ group (p=0.391) after stratification by statin use. However, a non-significant trend to lower HDL-c in the ε4/ ε4 genotype was seen in the Sim+ group. Multivariate analysis (**Table 3.16**) showed that the *APOE rs429358C* allele confers risk of lower HDL-c levels when adjusting for gender and statin use, which is in agreement with published literature.

The *APOE rs7412C>T* SNP encodes the arginine158cysteine amino acid change (ε2) which drastically reduces the proteins ability to bind to LDLR[69]. It is thought that the ε2 isoform impairs hydrolysis of TG rich VLDL, which is LPL mediated, leading to lowered LDL-c levels in transgenic mice[69,76]. However, the mechanism is not fully understood. Additionally, children with the ε2 allele were found to have lower LDL-c and higher HDL-c levels[75]. Results from our study (**Table 3.11**) agree with the literature, with ε2/ε2 patients having lower median TC and median LDL-c than ε3/ε3 patients in the combined cohort (p<0.0001 in both cases). These associations were repeated in the Sim+ group (p=0.027 and p=0.003, respectively) but not in the Sim- group (p=0.127 and p=0.485, respectively). Multivariate analysis (**Table 3.16**) showed that the *APOE rs7412T* allele confers a risk of lower TC levels (OR=0.747, p=0.028), and lower LDL-c levels (OR=0.759, p=0.012) when adjusting for statin use.

While we did not see the same linear and inverse relationships between the *APOE* (ε2/ε2, ε2/ε3, ε2/ε4, ε3/ε3, ε3/ε4, ε4/ε4) genotypes and LDL-c and HDL-c, respectively, we did repeat the associations when comparing ε2 to ε3 and ε3 to ε4. Significance that was not repeated in Sim+ or Sim- groups could be influenced by reducing sample sizes during stratification. Additionally, inter-individual response and adherence to statin therapy is not accounted for which may play a significant role. Lipid levels in the Sim+ group are differentially affected by statin action, possibly masking the effects caused by genetic variation. The Sim- group may have additional genetic variations which mitigate and mask the effects of *APOE* variation on LDL-c. With the multivariate analysis, however, there is enough evidence to postulate that the common *APOE* variations do play a significant role in dyslipidaemia in T2DM black South Africans and should be accounted for in future studies.

4.7.2) The impact of *PCSK9* genetic variation on aberrant lipid profiles

Genetic variation in *PCSK9* gene has been an area of great interest in recent studies of dyslipidaemia. It plays a vital role in the degradation and recycling of the LDLR and, thus, genetic variations, which can either enhance or abolish its function, have been keenly studied for their effects on altered LDL-c. Therefore, we decided to investigate the effects of a common reportedly gain-of-function variation *PCSK9 rs505151A>G,* and a rare African specific loss-of-function variation *PCSK9 rs28362286C>A* in our cohort.

PCSK9 rs505151A>G encodes a glutamic acid to glycine (E670G) change in the cysteine-rich C- terminal domain of PCSK9. It is thought that this variation may increase the affinity for PCSK9 to bind to LDLR, thus enhancing LDLR degradation[88] as described in **Chapter1.4.6.** The *PCSK9 rs505151G* allele has been associated, in a meta-analysis, with increased LDL-c levels (OR=1.546, p <0.001) in all studied ethnicities[88]. We reported no such significant associations between the *PCSK9 rs505151G>A* variation and lipid profiles through univariate analysis (**Appendix V, Table 5.2**) and multivariate analysis (**Table 3.16)**. It appears that this SNP does not play a significant role in our population.

The *PCSK9 rs28362286C>A* encodes a C679X truncation that disrupts the folding of the C-terminal domain, thought to be important for the successful secretion of the protein[80]. PCSK9 with the C679X truncation is efficiently expressed and processed in human liver cells, retaining catalytic activity, but is retained in the ER and is not secreted into the cell medium[92]. The C679X variant was associated with a decrease in LDL-c levels by 27% in Zimbabwean women[91]. We report no such significant associations between the *PCSK9 rs28362286C>A* variation, using a dominant genetic model, and lipid profiles through univariate analysis (**Table 3.12**) or multivariate analysis (**Table 3.16)**. Another study found no significant difference in LDL-c levels between 679X carriers and C679 homozygotes in black South African women[138]. It may be possible that enough functional protein is excreted in *PCSK9 rs28362286C/A* heterozygotes and that associations with lower LDL-c can only be seen in a recessive genetic model which is difficult to analyse with such a rare mutation. The cohort we studied only had one *PCSK9 rs28362286A/A* homozygote, which was genotyped through sequencing, so significant differences in median lipid levels could not be determined. This patient should not have any functional PCSK9 secreted into the cell medium, and should have lower LDL-c. The 48year old female, who was not on statin therapy, had a lipid profile

as follows: TC (2.72 mmol/L), TG (1.40 mmol/L), HDL-c (0.65 mmol/L), and LDL-c (1.43 mmol/L). Her lipid profile was low across the board, and she was a poor controller T2DM (HbA1c% =7.8). Thus, it is plausible that her *PCSK9 rs28362286A/A* variation plays a protective role against LDL-c dyslipidaemia.

While the analysis of *PCSK9 rs505151A>G* and *PCSK9 rs28362286C>A* did not yield any significant results in our study cohort, the role of *PCSK9* variation in South Africans still need further elucidation. There are many additional variations which could have an impact which were not studied in this population due to time constraints.

4.7.3) The impact of *CETP* genetic variation on aberrant lipid profiles

The role of *CETP* in the dyslipidaemia of South Africans populations is relatively under investigated. The common *CETP rs708272G>A* (*TaqIB*) variant has been reported in a population of black South African women, where the *TaqIB* B2 allele was associated with lower LDL-c levels[138]. *CETP rs708272G>A* is a synonymous mutation with no known effect on CETP function and no predicted effects on regulation or splicing site alteration. It has been hypothesised that reported association is caused by another SNP in strong linkage disequilibrium (LD) with *CETP rs708272G>A.* There is evidence that *CETP rs708272G>A* is in strong LD ($r^2 = 0.74$) with the 5' upstream *CETP rs183130C>T* SNP, which was strongly associated with increased HDL-c in individuals of African descent[97]. The *CETP rs183130C>T* variant is predicted to alter a transcription binding site, therefore, it is a good candidate for the causative allele in the haplogroup[139]. However, analysis of *CETP rs708272G>A* in our study cohort showed no significant associations with altered lipids (**Table 3.14**). The *CETP rs708272A* allele showed a trend towards a significant increase in HDL-c (p=0.073) when compared to the *rs708272G* allele in the combined cohort, but significance was not repeated when stratifying by statin use. Multivariate analysis (**Table 3.16**) confirmed that there was no significant associations with lipid levels. As *CETP rs708272G>A* is not causative, we would need to investigate the *CETP rs183130C>T* SNP in the future to determine its distribution and role in our population. This would be a further fine mapping of the *CETP rs708272G>A* haplogroup and may provide useful insight going further.

The *CETP rs34065661C>G* encodes the A15G amino acid change. However, the amino acid is cleaved upon secretion of protein, so it has no effect on CETP activity but may affect

CETP secretion efficiency[140]. The *CETP rs34065661C/G + G/G* genotypes, when applying a dominant genetic model, were associated with increased median HDL-c in our combined cohort and Sim+ cohort (p=0.017 and p=0.026, respectively) (**Table 3.13**). This agrees with results published by Pirim *et al.*[97]. Significance was lost in multivariate analysis (**Table 3.16**) when adjusting for the effects of gender, HbA1c%, and statin use. The *CETP rs34065661G* allele approached a significant increase in risk of high LDL-c levels (OR=1.37, p=0.057), when adjusting for statin use. This may be a false positive result caused by small sample size.

The *CETP rs3816117T* allele showed a trend towards significantly lower HDL-c levels in the combined cohort (p=0.083) (**Table 3.15**). However, this trend is not observed once stratified by statin use or in multivariate analysis (**Table 3.16**). The *CETP rs3816117C>T* variant is intronic with no reported functional significance. Interestingly, the *CETP rs34680782A* allele was significantly associated with 85.2% increased risk of higher LDL-c in multivariable analysis, adjusting for statin use (OR=1.85, p=0.022). There were no significant associations in the *CETP rs34680782C>A* univariate analysis, but a non-significant trend towards higher LDL-c was seen in the *CETP rs34680782C/A + A/A* genotypes when applying a dominant genetic model. The SNP is not predicted to be functional and has not been associated with LDL-c in the literature, but further investigation is warranted. No other significant associations with the other *CETP* variations and altered lipid profiles was observed (**Appendix V, Table 3.16**).

One problem with investigating the role of genetics in a complex disorder with multiple confounding effects, is that the contribution of single genetic variants may be lost if their effect sizes are small. Thus, we decided to investigate the association of a combination of three *CETP* SNPs, namely, *CETP rs34065661C>G, CETP rs3816117C>T,* and *CETP rs708272G>A,* which showed significant associations and/or trends with altered HDL-c and which have been individually associated with altered HDL-c in the literature[97]. The process involved is described in **Chapter 3.6.3**. Patients with the *rs34065661C/G + C/C, rs708272G/A + A/A,* and *rs3816117C/C* genotypes had significantly higher levels of HDL-c (p=0.001) than those with the *rs34065661G/G, rs708272G/G,* and *rs3816117C/T + T/T* genotypes in the combined cohort (**Figure 3.10**). The association was repeated in patients on statin therapy (p=0.011), but not in patients not on statin therapy (p=0.1402). We can see that a combination of *CETP* SNPs can have a significant effect on altered HDL-c, highlighting the polygenic nature of dyslipidaemia. Therefore, it may be important in the future to genotype as

many *CETP* SNPs as possible in a large cohort to determine the combined effect as *CETP* plays a role in dyslipidaemia in our population.

4.8) Study limitations

The study cohort was selected from a larger cohort of individuals recruited at the diabetes clinic at the Chris Hani Baragwaneth hospital in Johannesburg as part of a main protocol on the genetic susceptibility to T2DM carried out by Dr. Donald Tanyanyiwa. A large proportion of these patients were on lipid lowering medications (mostly simvastatin) as this is a common pharmacological therapy prescribed by clinicians to patients visiting the clinic. This was major confounding factor when investigating the statistical association of genetic variants with altered lipid levels and had to be accounted for in analysis.

None of the continuous variables assessed in this study had normal distributions according to the Shapiro-Wilk test. Only age approached a normal distribution ($p=0.033$). This has implications for analysis as many tests assume Gaussian distributions for accuracy. Several transformations (e.g. log10 transformation) were attempted to improve normality, with no success. Outliers were checked for but not removed from analysis as there was already a small sample size, and extreme values were of interest to us. Thus, non-parametric tests were used to analyse the data further.

The sample size was small for a genetic association study and, due to resource constraints, the number of genetic variants that could be analysed was limited. This meant that previously associated genetic variations in other lipid metabolism relevant genes such as *LDLR, LPA*, *LCAT, LPL, APOB, ANGPTL3* and *ANGPTL4* could not be studied in our cohort. This is reflected in the fact that few of our investigated SNPs were associated with altered lipid profiles, while no SNPs were reported with altered TG levels. There is great deal of variation in lipid profiles which is unaccounted for. Additionally, number of potentially significant SNPs in *PCSK9* and *CETP* were not studied in our cohort, so we do not have the full picture of the contribution of these genes to dyslipidaemia in our cohort. It also has implications for statistical analysis. Power is reduced, HWE deviation may be affected, normality of the distribution of variables is affected, and type I error rates increase.

A number of patients had missing clinical data which meant that they had to be excluded from certain statistical analysis. This further reduced the sample sizes available for analysis.

There was no information on alcohol usage, so its effect on dyslipidaemia in our cohort could not be investigated. The LDL-c values of patients were determined by an indirect method, the Friedewald equation, which has intrinsic limitations. In in cases of dysbetalipoproteinemia (Type Ill hyperlipoproteinemia), VLDL-c concentrations are underestimated, and LDL-c concentrations are overestimated. Additionally, in cases where TG ≥ 4.5mmol/l the equation overestimates VLDL-c concentrations and underestimates LDL-c concentrations[121]. LDL-c cannot be accurately estimated where TG ≥ 4.5mmol/l, thus, these patients had to be excluded from analysis of LDL-c. Our study did not have suitable matched controls for comparison. The nature of the study cohort, with associated T2DM co-morbidities, made this a hard task to accomplish.

4.9) Conclusion and future directions

Our study set out to characterize the impact of genetic variations in *APOE, PCSK9,* and *CETP* on aberrant lipid metabolism in an indigenous black South African population with T2DM, with the aim of identifying a set of population specific biomarkers which could be used to better predict dyslipidaemia in the black South African population.

In doing so we report on the frequencies and distribution of the *PCSK9 rs505151A>G, CETP rs17231520G>A, CETP rs34065661C>G, CETP rs5884C>A, CETP rs34680782C>A, CETP rs17231534C>A, CETP rs3816117C>T, CETP rs561260717C>T, CETP rs34119551A>T,* and *CETP rs5030708C>T* SNPs for the first time in a black South African population. We investigated the association between clinical factors and altered lipid profiles in a black South African T2DM cohort, and found that gender is associated with HDL-c, age is associated with TG, BMI is associated with TG, and HbA1c% is associated with both TG and HDL-c. Importantly, we show that statin use is a biomarker of dyslipidaemia for TC, TG, and LDL-c.

Lastly, we identify several genetic biomarkers of dyslipidaemia in the black South African T2DM cohort. We show that the *APOE rs7412T* (ε2) allele is associated with decreased TC and LDL-c, the *APOE rs429358C* (ε4) allele is associated with decreased HDL-c, and that a combination of *CETP rs34065661C>G, CETP rs3816117C>T,* and *CETP rs708272G>A* alleles are associated with altered HDL-c. We also identify two SNPs *PCSK9 rs28362286C>A,* and *CETP rs34680782C>A* which need to be further investigated in our population for a conclusive picture of their effect on LDL-c.

While our study design had limitations, it assisted us in achieving the aim we set out to achieve. However, there are a number of improvements which could be made and future work to be done. The number one issue with association studies is sample size. This is a study design factor that is greatly influenced by available resources. This can be overcome by clever study design and, most importantly, collaboration. The pooling of resources in studies with similar end points.

Ideally, a panel of multiple lipid metabolism relevant variations in genes such as *PSCK9, CETP, LDLR, LPA, LCAT, LPL, APOB, ANGPTL3* and *ANGPTL4*, identified as being significant in other African populations, would be set up and genotyped for in a larger longitudinal study cohort of black South African T2DM patients. Including matched controls. The associations could then be investigated and the variance in lipid profiles better explained. Next generation sequencing could then be employed to identify novel biomarkers in patients with extreme unexplained lipid profiles.

The end goal would be determining risk of developing dyslipidaemia, early on in life, in a method similar to that published by Nuotio *et al*.[141]. The panel of associated genetic variations to are used to create a weighted genetic risk score. Childhood lipid levels, which they show track to adulthood, could be taken along with established clinical risk factors and used in conjunction with the weighted genetic risk score and the risk for developing dyslipidaemia established[141]. We could identify patients who are at high risk for developing dyslipidaemia early on and employ the correct prophylaxis management strategies to prevent future CVD events and possible preventable fatalities.

In conclusion, it is evident that there is still a lot of work to be done in order to fully characterise the role of genetic variation in dyslipidaemia in the South African population. This MSc project is a step in the right direction and achieves its aim in improving the understanding of potential biomarkers of dyslipidaemia in a South African T2DM population.

References

1. WHO. Global status report on noncommunicable diseases. (2014). doi:ISBN 9789241564854
2. WHO. Global Health Estimates 2015: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2015. (2015).
3. Mathers, C. *et al.* WHO methods and data sources for country-level causes of death 2000-2015. *Evid. Res.* (2017).
4. Statistics South Africa. Mortality and causes of death in South Africa, 2015: Findings from death notification. *Stat. release P0309.3* 1–127 (2016). doi:Statistical release P0309.3
5. Ross, R. Inflammation or Atherogenesis. *N. Engl. J. Med.* **340,** 115–126 (1999).
6. WHO. Global atlas on cardiovascular disease prevention and control. (2011).
7. Napoli, C., Armiento, F. P. D., Mancini, F. P., Postiglione, A. & Witztum, J. L. Fatty Streak Formation Occurs in Human Fetal Aortas and is Greatly Enhanced by Maternal Hypercholesterolemia. *J. Clin. Invest.* **100,** 2680–2690 (1997).
8. Navab, M. *et al.* The Yin and Yang of Oxidation in the Development of the Fatty Streak. *Arter. Thromb Vasc Biol* **16,** 831–42 (1996).
9. Joseph, P. *et al.* Reducing the Global Burden of Cardiovascular Disease, Part 1: The Epidemiology and Risk Factors. *Circ. Res.* **121,** 677–694 (2017).
10. WHO. Global Health Risks: Mortality and burden of disease attributable to selected major risks. *Bull. World Health Organ.* **87,** 646–646 (2009).
11. Ng, M. *et al.* Smoking Prevalence and Cigarette Consumption in 187 Countries, 1980-2012. *Jama* **311,** 183 (2014).
12. Teo, K.K. *et al.* Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: a case-control study. *Lancet* **368,** 647–658 (2006).
13. Barnoya, J. & Glantz, S. A. Cardiovascular effects of secondhand smoke: Nearly as large as smoking. *Circulation* **111,** 2684–2698 (2005).
14. Smyth, A. *et al.* Alcohol consumption and cardiovascular disease, cancer, injury, admission to hospital, and mortality: A prospective cohort study. *Lancet* **386,** 1945–1954 (2015).
15. Department of Health. Alcohol Guidelines Review – Report from the Guidelines development group to the UK Chief Medical Officers. 1–44 (2016).
16. Di Cesare, M. *et al.* Trends in adult body-mass index in 200 countries from 1975 to 2014: A pooled analysis of 1698 population-based measurement studies with 19.2 million participants. *Lancet* **387,** 1377–1396 (2016).
17. WHO. NCD Profile South Africa. (2014). doi:10.1057/9781137269201.0014
18. Yusuf, S. *et al.* Obesity and the risk of myocardial infarction in 27,000 participants

from 52 countries: a case-control study. *Lancet* **366,** 1640–1649 (2005).

19. Gelber, R. P. *et al.* Measures of Obesity and Cardiovascular Risk Among Men and Women. *J. Am. Coll. Cardiol.* **52,** 605–615 (2008).

20. Rush, E. C. *et al.* BMI, fat and muscle differences in urban women of five ethnicities from two countries. *Int. J. Obes.* **31,** 1232–1239 (2007).

21. Punyadeera, C. *et al.* Ethnic differences in lipid metabolism in two groups of obese South African women. *J. Lipid Res.* **42,** 760–767 (2001).

22. Cholesterol Treatment Trialists' Collaboration. Protocol for a prospective collaborative overview of all current and planned randomized trials of cholesterol treatment regimens. *Am. J. Cardiol.* **75,** 1130–1134 (1995).

23. MacMahon, S. *et al.* Blood cholesterol and vascular mortality by age, sex, and blood pressure: A meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths. *Lancet* **370,** 1829–1839 (2007).

24. Cholesterol Treatment Trialists' Collaboration. Efficacy and safety of cholesterol-lowering treatment: Prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins. *Lancet* **366,** 1267–1278 (2005).

25. The Emerging Risk Factors Collaboration. Major Lipids, Apolipoproteins, and Risk of Vascular Disease. *Jama* **302,** 1993 (2009).

26. Russell, D. W. Cholesterol biosynthesis and metabolism. *Cardiovasc. Drugs Ther.* **6,** 103–10 (1992).

27. Dietschy, J. Regulation of cholesterol metabolism in man and in other species. *Klin. Wochenschr.* 338–345 (1984).

28. Ramasamy, I. Recent advances in physiological lipoprotein metabolism. *Clin. Chem. Lab. Med.* **52,** 1695–1727 (2014).

29. Black, D. D. Development and physiological regulation of intestinal lipid absorption. I. Development of intestinal lipid absorption: cellular events in chylomicron assembly and secretion. *Am. J. Physiol. Gastrointest. Liver Physiol.* **293,** G519-24 (2007).

30. Havel, R. J. Triglyceride-rich lipoproteins and plasma lipid transport. *Arterioscler. Thromb. Vasc. Biol.* **30,** 9–19 (2010).

31. Tiwari, S. & Siddiqi, S. A. Intracellular trafficking and secretion of VLDL. *Arterioscler. Thromb. Vasc. Biol.* **32,** 1079–1086 (2012).

32. Helkin, A. *et al.* Dyslipidemia Part 1—Review of Lipid Metabolism and Vascular Cell Physiology. *Vasc. Endovascular Surg.* **50,** 107–118 (2016).

33. Goldstein, J. L. & Brown, M. S. The LDL receptor. *Arterioscler. Thromb. Vasc. Biol.* **29,** 431–438 (2009).

34. Horton, J. D., Goldstein, J. L. & Brown, M. S. SREBPs: activators of the complete program of cholesterol and fatty acid synthesis. *J. Clin. Invest.* **109,** 1125–1131 (2002).

35. Galeano, N. F., Al-Haideri, M., Keyserman, F., Rumsey, S. C. & Deckelbaum, R. J. Small dense low density lipoprotein has increased affinity for LDL receptor-independent cell surface binding sites: a potential mechanism for increased

atherogenicity. *J. Lipid Res.* **39,** 1263–1273 (1998).

36. Glomset, J. A. The plasma lecithin:cholesterol acyltransferase reaction. *J. Lipid Res.* **9,** 155–167 (1968).

37. O'Connell, B.J & Genest, Jr, J. High-Density Lipoproteins and Endothelial Function. *Circulation* **104,** 1978–1983 (2001).

38. Yu, X., Fu, Y., Zhang, D., Yin, K. & Tang, C. Foam cells in atherosclerosis. *Clin. Chim. Acta* **424,** 245–252 (2013).

39. Vedhachalam, C. *et al.* Mechanism of ATP-binding cassette transporter A1-mediated cellular lipid efflux to apolipoprotein A-I and formation of high density lipoprotein particles. *J. Biol. Chem.* **282,** 25123–25130 (2007).

40. Tall, A. R. Plasma High Density Lipoproteins. *J. Clin. Investig. Clin. Investig.* **86,** 379–384 (1990).

41. Francone, O. L. & Gurakar, A. Distribution and Functions of Lecithin : Cholesterol Acyltransferase and Cholesteryl Ester Transfer Protein in Plasma Lipoproteins. *J Biol Chem* **264,** 7066–7072 (1989).

42. Settasatian, N. *et al.* The Mechanism of the Remodeling of High Density Lipoproteins by Phospholipid Transfer Protein. *J. Biol. Chem.* **276,** 26898–26905 (2001).

43. Lund-Katz, S. & Phillips, M. C. *High Density Lipoprotein Structure–Function and Role in Reverse Cholesterol Transport*. **51,** (2010).

44. Stancu, C. & Sima, A. Statins: mechanism of action and effects. *J. Cell. Mol. Med.* **5,** 378–387 (2001).

45. Bellosta, S., Ferri, N., Bernini, F., Paoletti, R. & Corsini, A. Non-lipid-related effects of statins. *Ann. Med.* **32,** 164–176 (2000).

46. Collins, R. *et al.* Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet* **388,** 2532–2561 (2016).

47. Rosenson, R. S. *et al.* Dysfunctional HDL and atherosclerotic cardiovascular disease. *Nat. Rev. Cardiol.* **13,** 48–60 (2015).

48. Sanin, V., Pfetsch, V. & Koenig, W. Dyslipidemias and Cardiovascular Prevention: Tailoring Treatment According to Lipid Phenotype. *Curr. Cardiol. Rep.* **19,** (2017).

49. Barter, P. J., Caulfield, M. J., Mats, E. & Scott, G. Effects of Torcetrapib in Patients at High Risk for Coronary Events. *N. Engl. J. Med.* **357,** (2007).

50. Schwartz, G. G. *et al.* Effects of Dalcetrapib in Patients with a Recent Acute Coronary Syndrome. *N. Engl. J. Med.* **367,** 2089–2099 (2012).

51. Nicholls, S. J. *et al.* Assessment of the clinical effects of cholesteryl ester transfer protein inhibition with evacetrapib in patients at high-risk for vascular outcomes: Rationale and design of the ACCELERATE trial. *Am. Heart J.* **170,** 1061–1069 (2015).

52. Catapano, A. L. *et al.* 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias. *Eur. Heart J.* **37,** 2999–3058l (2016).

53. Langsted, A. *et al.* Nonfasting cholesterol and triglycerides and association with risk of

myocardial infarction and total mortality: The Copenhagen City Heart Study with 31years of follow-up. *J. Intern. Med.* **270,** 65–75 (2011).

54. Jørgensen, A. B. *et al.* Genetically elevated non-fasting triglycerides and calculated remnant cholesterol as causal risk factors for myocardial infarction. *Eur. Heart J.* **34,** 1826–1833 (2013).

55. Do, R., Willer, C. J., Schmidt, E. M. & Sengupta, S. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45,** 1345–1352 (2013).

56. Labreuche, J., Touboul, P.-J. & Amarenco, P. Plasma triglyceride levels and risk of stroke and carotid atherosclerosis: A systematic review of the epidemiological studies. *Atherosclerosis* **203,** 331–345 (2009).

57. Nordmann, A. J. *et al.* Effects of low-carbohydrate vs low-fat diets on weight loss and cardiovascular risk factors: a meta-analysis of randomized controlled trials. *Arch Intern Med* **166,** 285–293 (2006).

58. Dattilo, A. M. & Kris-Etherton, P. M. Effects of Weight Reduction on Blood Lipids and Lipoproteins a Meta-Analysis. *Am. Clin.* **56,** 320–328 (1992).

59. Chowdhury, I. Alcohol and Dyslipidemia. in *Alcohol, Nutrition, and Health Consequences* (eds. Watson, R. R., Preedy, V. R. & Zibadi, S.) 329–339 (Humana Press, 2013). doi:10.1007/978-1-62703-047-2_26

60. Maeda, K., Noguchi, Y. & Fukui, T. The effects of cessation from cigarette smoking on the lipid and lipoprotein profiles: A meta-analysis. *Prev. Med. (Baltim).* **37,** 283–290 (2003).

61. Heller, D., De Faire, U., Pedersen, N., Dahlen, G. & McClearn, G. Genetic and environmental influences on serum lipid levels in twins. *N. Engl. J. Med.* **328,** 1150–1156 (1993).

62. Musunuru, K. & Kathiresan, S. Surprises from Genetic Analyses of Lipid Risk Factors for Atherosclerosis. *Circ. Res.* **118,** 579–585 (2016).

63. Rader, D. J., Cohen, J. & Hobbs, H. H. Monogenic hypercholesterolemia : new insights in pathogenesis and treatment. **111,** 1795–1803 (2003).

64. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).

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

66. Peloso, G. M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94,** 223–232 (2014).

67. Lange, L. A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94,** 233–245 (2014).

68. Voight, B. F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380,** 572–80 (2012).

69. Ramasamy, I. Update on the molecular biology of dyslipidemias. *Clin. Chim. Acta*

454, 143–185 (2016).

70. Weisgraber, K. H. Apolipoprotein E: Structure-Function Relationships. in *Lipoproteins, Apolipoproteins, and Lipases* (eds. Anfinsen, C. B., Edsall, J. T., Richards, F. M. & Eisenberg, D. S. B. T.-A. in P. C.) **45,** 249–302 (Academic Press, 1994).

71. Beisiegel, U., Weber, W., Ihrke, G., Herz, J. & Stanley, K. K. The LDL-receptor-related protein, LRP, is an apolipoprotein E-binding protein. *Nature* **341,** 162–164 (1989).

72. Mahley, R. W., Huang, Y. & Rall Jr., S. C. Pathogenesis of type III hyperlipoproteinemia (dysbetalipoproteinemia). Questions, quandaries, and paradoxes. *J. Lipid Res.* **40,** 1933–1949 (1999).

73. Weisgraber, K. H. Apolipoprotein E distribution among human plasma lipoproteins: role of the cysteine-arginine interchange at residue 112. *J. Lipid Res.* **31,** 1503–1511 (1990).

74. Hopkins, P. C. R., Huang, Y., McGuire, J. G. & Pitas, R. E. Evidence for differential effects of apoE3 and apoE4 on HDL metabolism. *J. Lipid Res.* **43,** 1881–9 (2002).

75. CR, I. *et al.* Apolipoprotein E2 allele is associated with an anti-atherogenic lipoprotein profile in children: the Columbia University BioMarkers Study. *Pediatrics* **106,** 568–575 (2000).

76. Huang, Y., Liu, X. Q., Rall Jr., S. C. & Mahley, R. W. Hypo-low density lipoprotein cholesterol effect in apolipoprotein E2 transgenic mice is caused by impaired lipoprotein lipase-mediated lipolysis of triglyceride-rich lipoproteins. *Circulation* **96,** I-39 (1997).

77. Bennet, A. M. *et al.* Association of Apolipoprotein E Genotypes With Lipid Levels and Coronary Risk. *Jama* **298,** 1300–1311 (2007).

78. Abifadel, M. *et al.* Mutations and polymorphisms in the proprotein convertase subtilisin kexin 9(PCSK9) gene in cholesterol metabolism and disease. *Hum. Mutat.* **30,** 520–529 (2009).

79. Seidah, N. G. & Prat, A. Precursor convertases in the secretory pathway, cytosol and extracellular milieu. *Essays Biochem.* **38,** 79 LP-94 (2002).

80. Cunningham, D. *et al.* Structural and biophysical studies of PCSK9 and its mutants linked to familial hypercholesterolemia. *Nat. Struct. Mol. Biol.* **14,** 413–419 (2007).

81. Seidah, N. G. *et al.* The secretory proprotein convertase neural apoptosis-regulated convertase 1 (NARC-1): Liver regeneration and neuronal differentiation. *Proc. Natl. Acad. Sci.* **100,** 928–933 (2003).

82. Peterson, A. S., Fong, L. G. & Young, S. G. PCSK9 function and physiology. *J. Lipid Res.* **49,** 1152–1156 (2008).

83. Lagace, T. A. *et al.* Secreted PCSK9 decreases the number of LDL receptors in hepatocytes and inlivers of parabiotic mice. *J. Clin. Invest.* **116,** 2995–3005 (2006).

84. Maxwell, K. N., Fisher, E. A. & Breslow, J. L. Overexpression of PCSK9 accelerates the degradation of the LDLR in a post-endoplasmic reticulum compartment. *Proc.*

Natl. Acad. Sci. U. S. A. **102,** 2069–74 (2005).

85. Chen, S. N. *et al.* A Common PCSK9Haplotype, Encompassing the E670G Coding Single Nucleotide Polymorphism, Is a Novel Genetic Marker for Plasma Low-Density Lipoprotein Cholesterol Levels and Severity of Coronary Atherosclerosis. *J. Am. Coll. Cardiol.* **45,** 1611–1619 (2005).

86. Evans, D. & Beil, F. U. The E670G SNP in the PCSK9 gene is associated with polygenic hypercholesterolemia in men but not in women. *BMC Med. Genet.* **7,** 66 (2006).

87. Hsu, L.A. *et al.* The PCSK9 gene E670G polymorphism affects low-density lipoprotein cholesterol levels but is not a risk factor for coronary artery disease in ethnic Chinese in Taiwan . *Clin Chem Lab Med.* **47,** 154 (2009).

88. Cai, G. *et al.* The associations between proprotein convertase subtilisin/kexin type 9 E670G polymorphism and the risk of coronary artery disease and serum lipid levels: a meta-analysis. *Lipids Health Dis.* **14,** 149 (2015).

89. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbs, H. H. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *N. Engl. J. Med.* **354,** 1264–1272 (2006).

90. Kotowski, I. K. *et al.* A spectrum of PCSK9 alleles contributes to plasma levels of low density lipoprotein cholesterol. *Am J Hum Genet* **78,** (2006).

91. Hooper, A. J., Marais, A. D., Tanyanyiwa, D. M. & Burnett, J. R. The C679X mutation in PCSK9 is present and lowers blood cholesterol in a Southern African population. *Atherosclerosis* **193,** 445–448 (2007).

92. Zhao, Z. *et al.* Molecular Characterization of Loss-of-Function Mutations in PCSK9 and Identification of a Compound Heterozygote. *Am. J. Hum. Genet.* **79,** 514–523 (2006).

93. Bruce, C., Chouinard Jr, R. A. & Tall, A. R. Plasma lipid transfer proteins, high-density lipoproteins, and reverse cholesterol transport. **18,** 297–330 (1998).

94. Tall, A. R. & Rader, D. J. The Trials and Tribulations of CETP Inhibitors. *Circ. Res.* **122,** 106-112 (2018).

95. Jiang, X. C. *et al.* Down-regulation of mRNA for the low density lipoprotein receptor in transgenic mice containing the gene for human cholesteryl ester transfer protein. Mechanism to explain accumulation of lipoprotein B particles. *J. Biol. Chem.* **268,** 27406–27412 (1993).

96. Boekholdt, S. M. & Thompson, J. F. Natural genetic variation as a tool in understanding the role of CETP in lipid levels and disease. *J Lipid Res.* **44,** 1080–1093 (2003).

97. Pirim, D. *et al.* Resequencing of the CETP gene in American whites and African blacks: Association of rare and common variants with HDL-cholesterol levels. *Metabolism.* **65,** 36–47 (2016).

98. Gotoda, T. *et al.* Cholesteryl Ester Transfer Protein Deficiency Caused by a Nonsense Mutation Detected in the Patient′s Macrophage mRNA. *Biochem. Biophys. Res. Commun.* **194,** 519–524 (1993).

99. Yamashita, S. *et al.* Total deficiency of plasma cholesteryl ester transfer protein in subjects homozygous and heterozygous for the intron 14 splicing defect. *Biochem. Biophys. Res. Commun.* **170,** 1346–1351 (1990).

100. Nagano, M. *et al.* Molecular mechanisms of cholesteryl ester transfer protein deficiency in Japanese. *J. Atheroscler. Thromb.* **11,** 110–121 (2004).

101. Corella, D. *et al.* Association of TaqIB polymorphism in the cholesteryl ester transfer protein gene with plasma lipid levels in a healthy Spanish population. *Atherosclerosis* **152,** 367–376 (2000).

102. Ruan, X. *et al.* Association of two CETP polymorphisms with HDL levels in the Chinese obese population. *Obesity* **17,** 2196–2201 (2009).

103. Wu, Z. *et al.* Association of cholesteryl ester transfer protein (CETP) gene polymorphism, high density lipoprotein cholesterol and risk of coronary artery disease: a meta-analysis using a Mendelian randomization approach. *BMC Med. Genet.* **15,** 118 (2014).

104. Dachet, C., Poirier, O., Cambien, F., Chapman, J. & Rouis, M. New functional promoter polymorphism, CETP/-629, in cholesteryl ester transfer protein (CETP) gene related to CETP mass and high density lipoprotein cholesterol levels: role of Sp1/Sp3 in transcriptional regulation. *Arterioscler. Thromb. Vasc. Biol.* **20,** 507–515 (2000).

105. Society for Endocrinology, Metabolism and Diabetes of South Africa. SEMDSA 2017 Guidelines for the Management of Type 2 diabetes mellitus. *J. Endocrinol. Metab. Diabetes South Africa* **22,** 1–182 (2017).

106. Low Wang, C. C., Hess, C. N., Hiatt, W. R. & Goldfine, A. B. Clinical update: Cardiovascular disease in diabetes mellitus. *Circulation* **133,** 2459–2502 (2016).

107. Booth, G. L., Kapral, M. K., Fung, K. & Tu, J. V. Relation between age and cardiovascular disease in men and women with diabetes compared with non-diabetic people: a population-based retrospective cohort study. *Lancet* **368,** 29–36 (2006).

108. Tancredi, M. *et al.* Excess Mortality among Persons with Type 2 Diabetes. *N. Engl. J. Med.* **373,** 1720–1732 (2015).

109. Parhofer, K. G. Interaction between Glucose and Lipid Metabolism: More than Diabetic Dyslipidemia. *Diabetes Metab J* **39,** 353–362 (2015).

110. Wu, L. & Parhofer, K. G. Diabetic dyslipidemia. *Metabolism.* **63,** 1469–1479 (2014).

111. Carmena, R., Duriez, P. & Fruchart, J.C. Atherogenic lipoprotein particles in atherosclerosis. *Circulation* **109,** III2-7 (2004).

112. Akinboboye, O., Idris, O. & Akinkugbe, O. Trends in coronary artery disease and associated risk factors in sub-Saharan Africans. *J. Hum. Hypertens.* **17,** 381–7 (2003).

113. Zoratti, R. A review on ethnic differences in plasma triglycerides and high-density-lipoprotein cholesterol: Is the lipid pattern the key factor for the low coronary heart disease rate in people of African origin? *Eur. J. Epidemiol.* **14,** 9–21 (1998).

114. Sliwa, K. *et al.* Different lipid profiles according to ethnicity in the Heart of Soweto study cohort of de novo presentations of heart disease. *Cardiovasc. J. Afr.* **23,** 389–395 (2012).

115. WHO. Diabetes country profiles, South Africa, 2016. (2016).

116. International Diabetes Federation. IDF Diabetes Atlas. (2015).

117. Hird, T. R. *et al.* Burden of Diabetes and First Evidence for the Utility of HbA1c for Diagnosis and Detection of Diabetes in Urban Black South Africans: The Durban Diabetes Study. *PLoS One* **11,** (2016).

118. Naing, N. N. Determination of Sample Size. *Malaysian J. Med. Sci.* **10,** 84–86 (2003).

119. Sugiuchi, H. *et al.* Direct measurement of high-density lipoprotein cholesterol in serum with polyethylene glycol-modified enzymes and sulfated alpha-cyclodextrin. *Clin. Chem.* **41,** 717–723 (1995).

120. Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **18,** 499–502 (1972).

121. Warnick, G. R., Knopp, R. H., Fitzpatrick, V. & Branson, L. Estimating low-density lipoprotein cholesterol by the Friedewald equation is adequate for classifying patients on the basis of nationally recommended cutpoints. *Clin. Chem.* **36,** 15–19 (1990).

122. Goldstein, D. E. *et al.* Tests of Glycemia in Diabetes. *Diabetes Care* **27,** 1761 LP-1773 (2004).

123. Bunn, H., Gabbay, K. & Gallop, P. The Glycosylation of Hemoglobin : Relevance to Diabetes Mellitus. *Science* **200,** 21–27 (1978).

124. Koenig, R. J., Peterson, C. M., Kilo, C., Cerami, A. & Williamson, J. R. Hemoglobin AIc as an indicator of the degree of glucose intolerance in diabetes. *Diabetes* **25,** 230 LP-232 (1976).

125. UK Prospective Diabetes Study (UKPDS) Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* **352,** 837–853 (1998).

126. Diabetes Control and Complications Trial Research Group. Effect of intensive diabetes treatment on the development and progression of long-term complications in adolescents with insulin-dependent diabetes mellitus: Diabetes Control and Complications Trial. *J. Pediatr.* **125,** 177–188 (1994).

127. Gustafson, S., Proper, J. A., Bowie, E. J. & Sommer, S. S. Parameters affecting the yield of DNA from human blood. *Anal. Biochem.* **165,** 294–299 (1987).

128. Marrzoq, L. F. A., Sharif, F. A. & Abed, A. A. Relationship between ApoE gene polymorphism and coronary heart disease in Gaza Strip. *J. Cardiovasc. Dis. Res.* **2,** 29–35 (2011).

129. Mayo, O. A Century of Hardy – Weinberg Equilibrium. *Twin Res. Hum. Genet.* **11,** 249–256 (2008).

130. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9,** 477–485 (2008).

131. Bishara, A. J. & Hittner, J. B. Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches.

Psychol. Methods **17,** 399–417 (2012).

132. Yang, L. *et al.* Prevalence of type 2 diabetes mellitus among inland residents in China (2000–2014): A meta-analysis. *J. Diabetes Investig.* **7,** 845–852 (2016).

133. Centers for Disease Control and Prevention. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. *US Dep. Heal. Hum. Serv.* (2014).

134. Law, M. R., Wald, N. J. & Rudnicka, A. R. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *BMJ* **326,** 1423 (2003).

135. Clifton, P. M. & Nestel, P. J. Influence of gender, body mass index, and age on response of plasma lipids to dietary fat plus cholesterol. *Arterioscler. Thromb. Vasc. Biol.* **12,** 955 LP-962 (1992).

136. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68 (2015).

137. Tishkoff, S. A. & Williams, S. M. Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* **3,** 611 (2002).

138. Ellman, N., Keswell, D., Collins, M., Tootla, M. & Goedecke, J. H. Ethnic differences in the association between lipid metabolism genes and lipid levels in black and white South African women. *Atherosclerosis* **240,** 311–317 (2018).

139. Thompson, J. F., Wood, L. S., Pickering, E. H., Dechairo, B. & Hyde, C. L. High-density genotyping and functional SNP localization in the CETP gene. *J. Lipid Res.* **48,** 434–443 (2007).

140. Thompson, J. F. *et al.* Frequency and function of CETP variants among individuals of Asian ancestry. *Atherosclerosis* **202,** 241–247 (2018).

141. Nuotio, J. *et al.* Prediction of Adult Dyslipidemia Using Genetic and Childhood Clinical Risk Factors: The Cardiovascular Risk in Young Finns Study. *Circ. Cardiovasc. Genet.* **10,** (2017).

Appendices:

Appendix I: List of Reagents

1) **DNA Extraction:**

PBS (1 mM KH_2PO_4, 154 mM NaCl, 5.6 mM Na_2HPO_4, pH 7.4):
0.136 g KH_2PO_4
9.0 g NaCl
0.795 g Na_2HPO_4
Dissolved in 1 L dH_2O. Check pH. Autoclaved.

1 M Tris-HCl (pH 8):
121.1 g Tris base dissolved in 1 L dH_2O.
Check pH 8 and adjust accordingly. Autoclaved.

Saturated NaCl:
40 g of NaCl added to 100 ml of dH_2O and mixed until completely saturated

Sucrose Triton-X Lysing Buffer:
10 ml 1 M Tris-HCl (pH 8)
5 ml 1 M $MgCl_2$
10 ml Triton X-100
Made up to 1 L with dH_2O. Autoclave and store at 4 °C
Decant as needed and add 11 g/100 ml Sucrose (D (+) saccharose) before use.

T20E05:
20 ml 1 M Tris-HCl
10 ml 0.5 M EDTA
Made up to 1 L with dH_2O. Autoclaved

10 % SDS
10 g SDS dissolved in 100 mL dH_2O.

Proteinase K

2) **Agarose Gel Electrophoresis:**

Seakem® LE Agarose (Lonza, Rockland, USA)

Nusieve™ 3:1 agarose gel (Lonza, Rockland, USA)

10 X Tris-Borate EDTA (TBE) buffer:
108 g Tris Base

55 g of boric acid
7.5 g EDTA
Made up to 1 L with dH_2O. Autoclaved.
Diluted to 1 X TBE in dH_2O.

EZ-VISION™ DNA Dye (VWR Life Science, AMRESCO, Philadelphia, USA)

GelRed® nucleic acid dye (Biotium, Inc., Fremont, CA, USA)

3) PCR amplification:

5 X Green GoTaq® Reaction Buffer (Promega, Madison, USA)

10 mM dNTPs (BioLine, London, UK)

25 mM $MgCl_2$ (Promega, Madison, USA)

GoTaq® DNA polymerase (Promega, Madison, USA)

Sabax water for injections (sdH_2O) (Adcock Ingram, Johannesburg, South Africa)

4) PCR-RFLP:

AflIII RE (New England BioLabs, Inc., Ipswich, USA)

NEB3.1 buffer (New England BioLabs, Inc., Ipswich, USA)

HaeII RE (New England BioLabs, Inc., Ipswich, USA)

CutSmart® buffer (New England BioLabs, Inc., Ipswich, USA)

5) Post PCR Clean-up:

Fast alkaline phosphatase (FastAP) (Thermo Fisher Scientific, Wyham, USA)

Exonuclease I (*Exo*I) (Fermentas, Ontario, Canada)

6) Direct Cycle Sanger Seqeuncing

ABI Prism® BigDye® Terminator Cycle Sequencing v3.1 Kit (Applied Biosystems, Carlsbad, CA, USA):

BigDye® Terminator v 3.1 mix containing fluorescently-labelled ddNTPs and *Taq* DNA polymerase

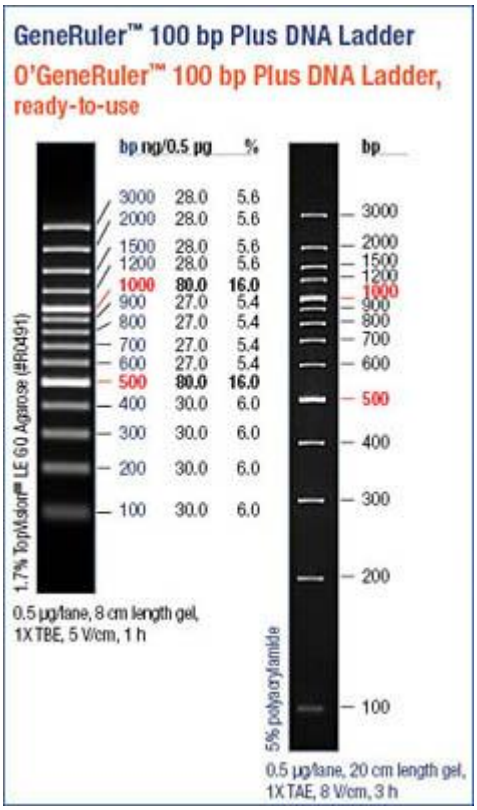
Sequencing buffer

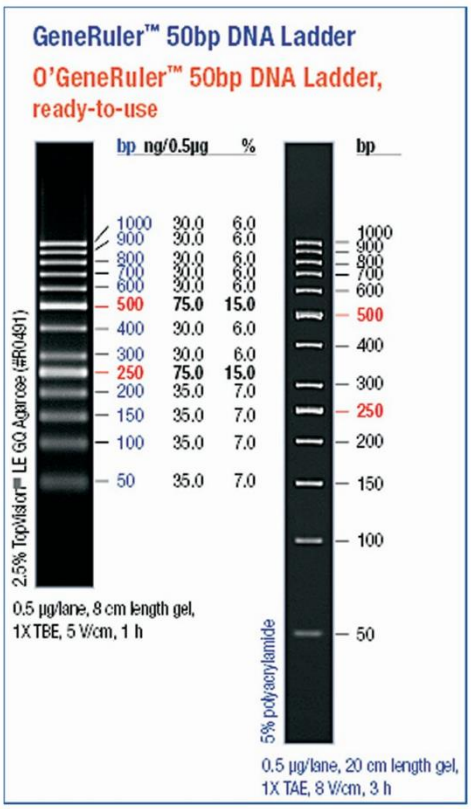
3 M NaOAc:
246.1 Sodium Acetate in 1 L dH_2O. Autoclaved.

70 % (v/v) ice-cold ethanol

Hi-DI (highly de-ionised) formamide

GeneRuler™ 100 bp Plus DNA Ladder and 50bp DNA Ladder:





Appendix II: DNA Quality Control

DNA Integrity Gel

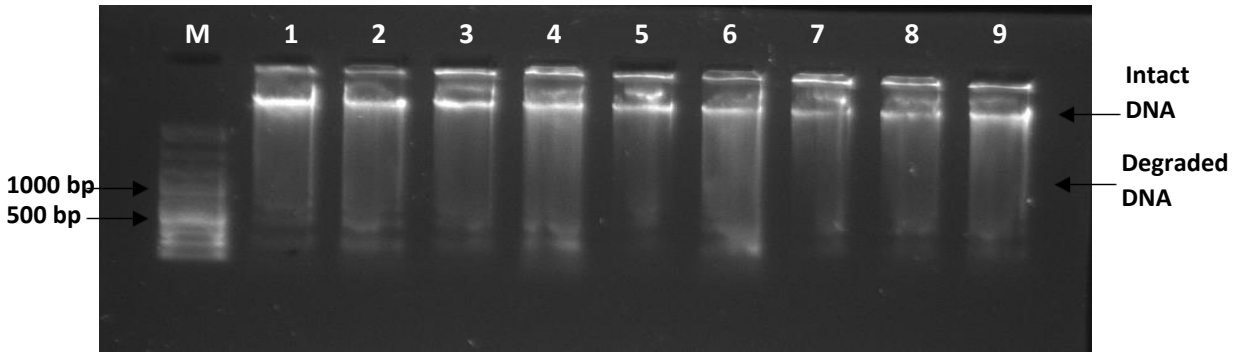


Figure 5.1: DNA integrity by 1% (w/v) agarose gel electrophoresis at 100V for 1hour. Lane M represents the GeneRuler® 100 bp Plus DNA Ladder Molecular Weight Marker (Fermentas, Ontario, Canada). Lanes 1-9 represent genomic DNA samples extracted from whole blood. A band of intact DNA can be seen, as well as smearing indicating various extents of degradation.

NanoDrop™ 1000 spectrophotometer example:

Table5.1: Extracted DNA concentration and purity determination by NanoDrop™ spectrophotometry

Sample ID	**Concentration (ng/ul)**	**A_{260}nm/A_{280}nm**	**A_{260}nm/A_{230}nm**
4001	747.26	1.82	1.59
4002	380.31	1.83	1.93
4003	13.67	1.73	0.72
4004	394.55	1.76	1.43
4005	403.40	1.73	1.19
4006	689.78	1.78	1.34
4008	1192.68	1.81	1.53

Appendix III: PCR amplification optimisation - Temperature Gradient PCRs

APOE:

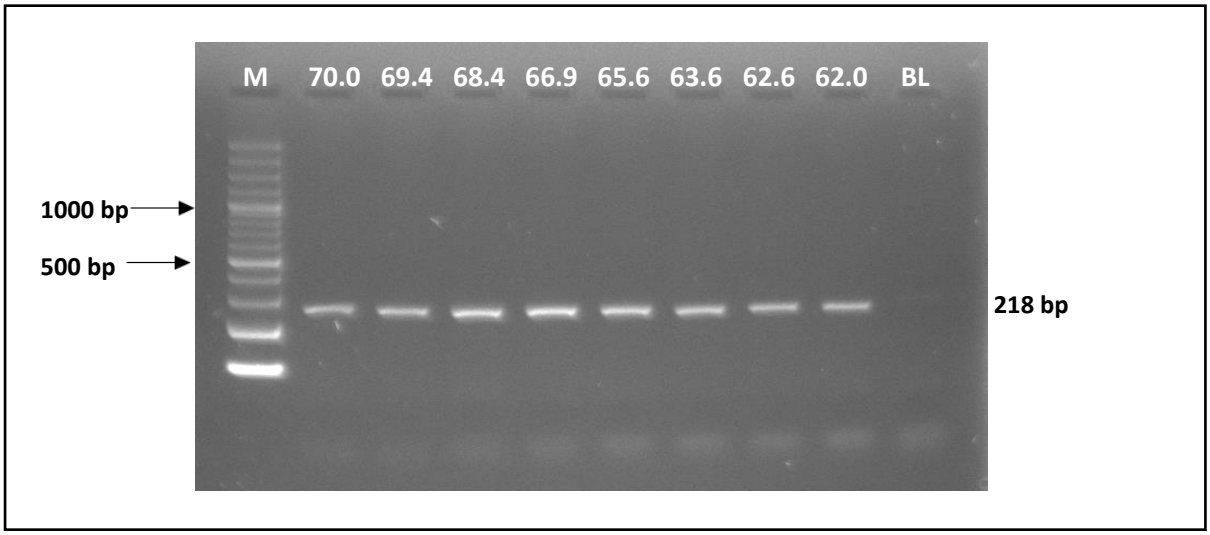


Figure 5.2: Temperature gradient PCR amplification of the 218 bp *APOE* fragment for T_a optimisation electrophoresed at 100 V for 60 minutes on a 1.5 % (w/v) agarose gel. Lane M represents the GeneRuler® 100 bp Plus DNA Ladder Molecular Weight Marker (Fermentas, Ontario, Canada). Lanes 2-9 represent amplification of control DNA at specified temperatures (°C). Lane BL is a no template control.

PSCK9:

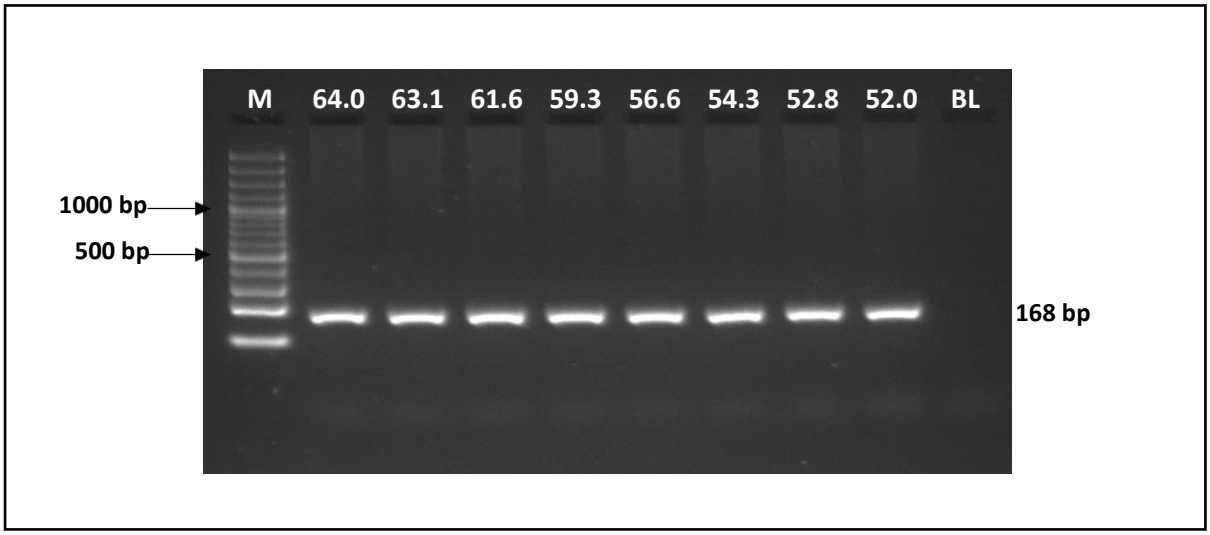


Figure 5.3: Temperature gradient PCR amplification of the 168 bp *PCSK9* fragment for T_a optimisation electrophoresed at 100 V for 60 minutes on a 1.5 % (w/v) agarose gel. Lane M represents the GeneRuler® 100 bp Plus DNA Ladder Molecular Weight Marker (Fermentas, Ontario, Canada). Lanes 2-9 represent amplification of control DNA at specified temperatures (°C). Lane BL is a no template control.

CETP:

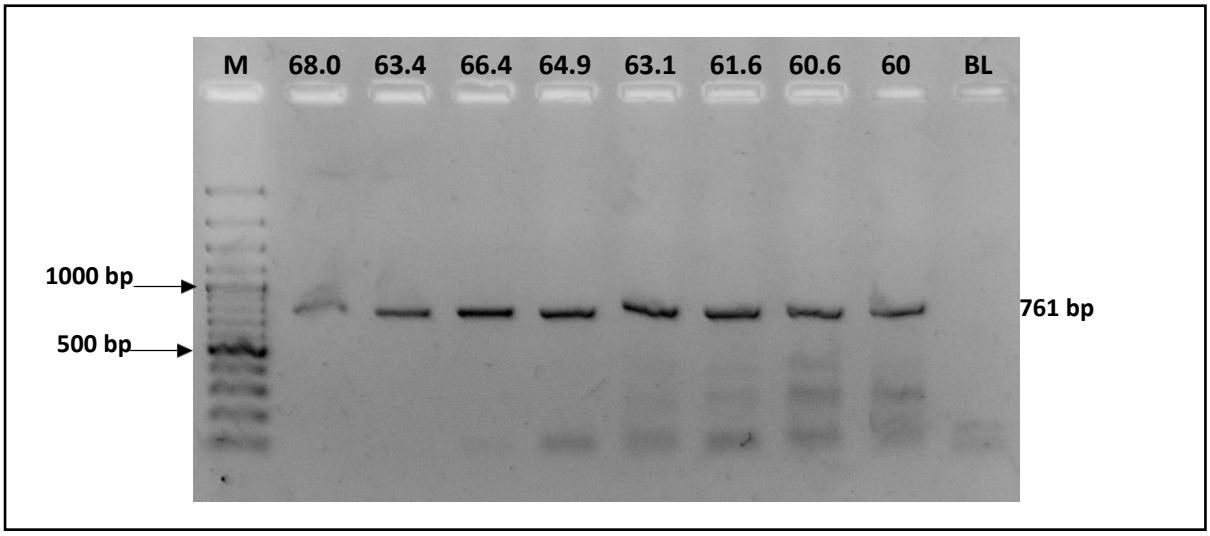


Figure 5.4: Temperature gradient PCR amplification of the 761 bp *CETP* fragment for T_a optimisation electrophoresed at 100 V for 60 minutes on a 1.5 % (w/v) agarose gel. Lane M represents the GeneRuler® 100 bp Plus DNA Ladder Molecular Weight Marker (Fermentas, Ontario, Canada). Lanes 2-9 represent amplification of control DNA at specified temperatures (°C). Lane BL is a no template control.

Appendix IV: *CETP* Electropherograms

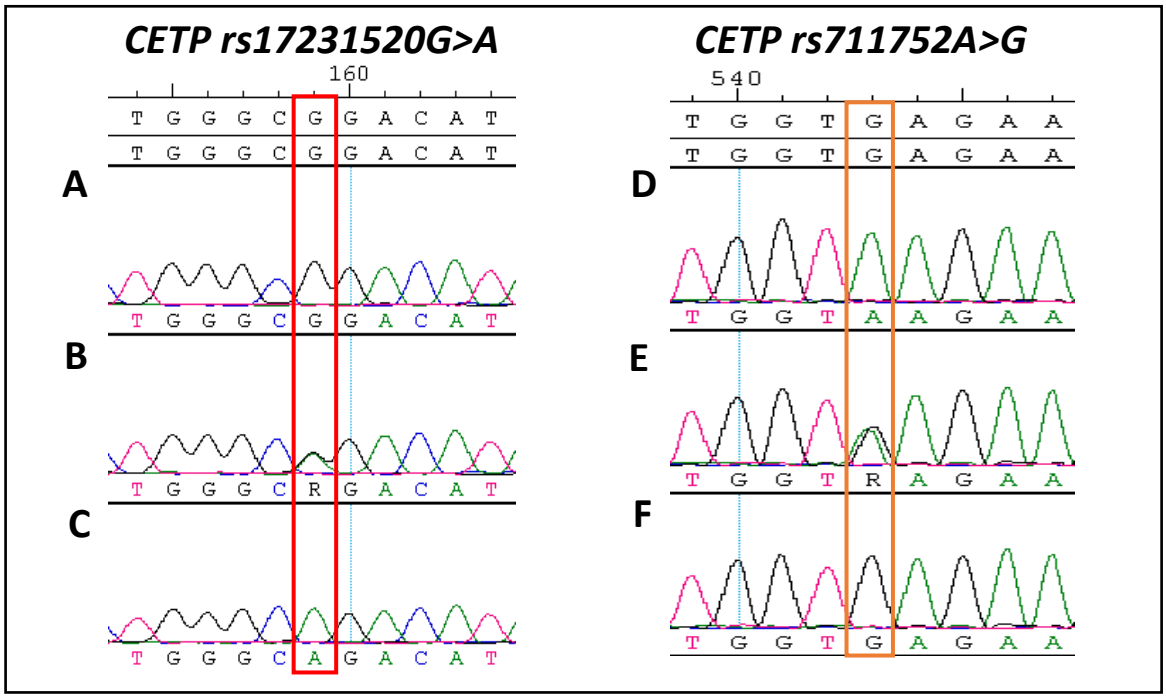


Figure 5.5: Electropherogram 1 of sequenced *CETP* PCR amplified fragments. The *CETP rs17231520G>A* and *CETP rs711752A>G* variants are highlighted by the red and orange boxes, respectively. **A** represents a sample with the *CETP rs17231520G/G* genotype. **B** represents a sample with the *CETP rs17231520G/A* genotype. **C** represents a sample with the *CETP rs17231520A/A* genotype. **D** represents a sample with the *CETP rs711752A/A* genotype. **E** represents a sample with the *CETP rs711752A/G* genotype. **F** represents a sample with the *CETP rs711752G/G* genotype.

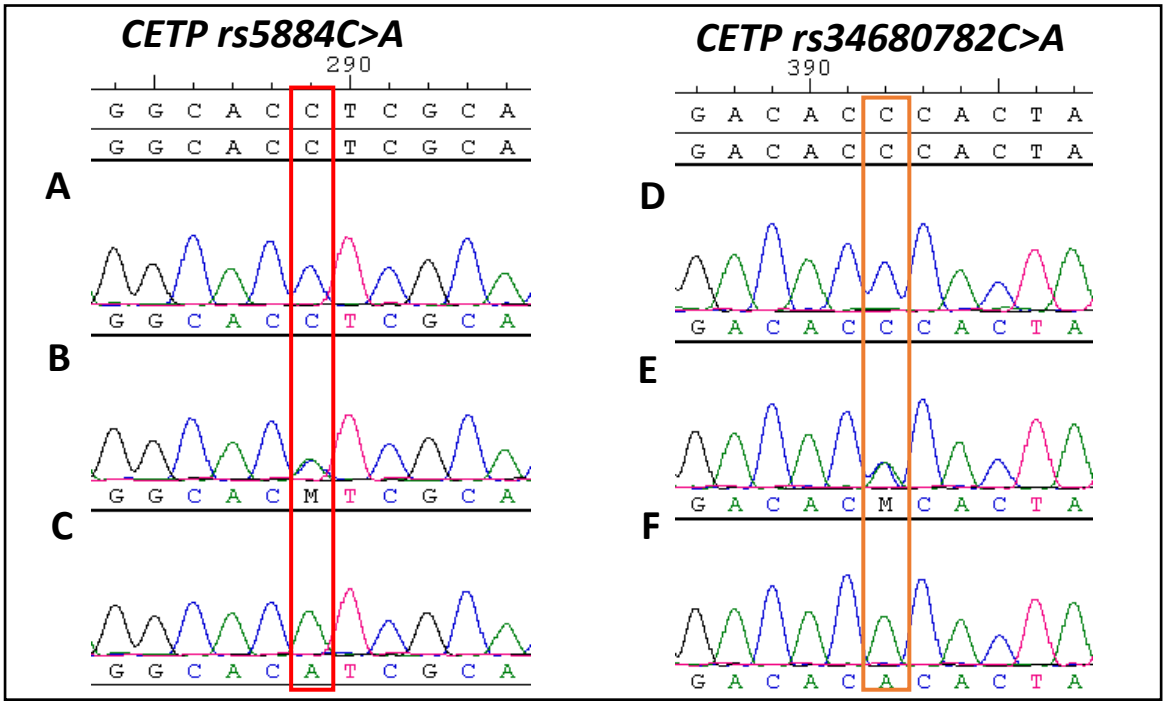


Figure 5.6: Electropherogram 2 of sequenced *CETP* PCR amplified fragments. The *CETP rs5884C>A* and *CETP rs34680782C>A* variants are highlighted by the red and orange boxes, respectively. **A** represents a sample with the *CETP rs5884C/C* genotype. **B** represents a sample with the *CETP rs5884C/A* genotype. **C** represents a sample with the *CETP rs5884A/A* genotype. **D** represents a sample with the *CETP rs34680782C/C* genotype. **E** represents a sample with the *CETP rs34680782C/A* genotype. **F** represents a sample with the *CETP rs34680782A/A* genotype.

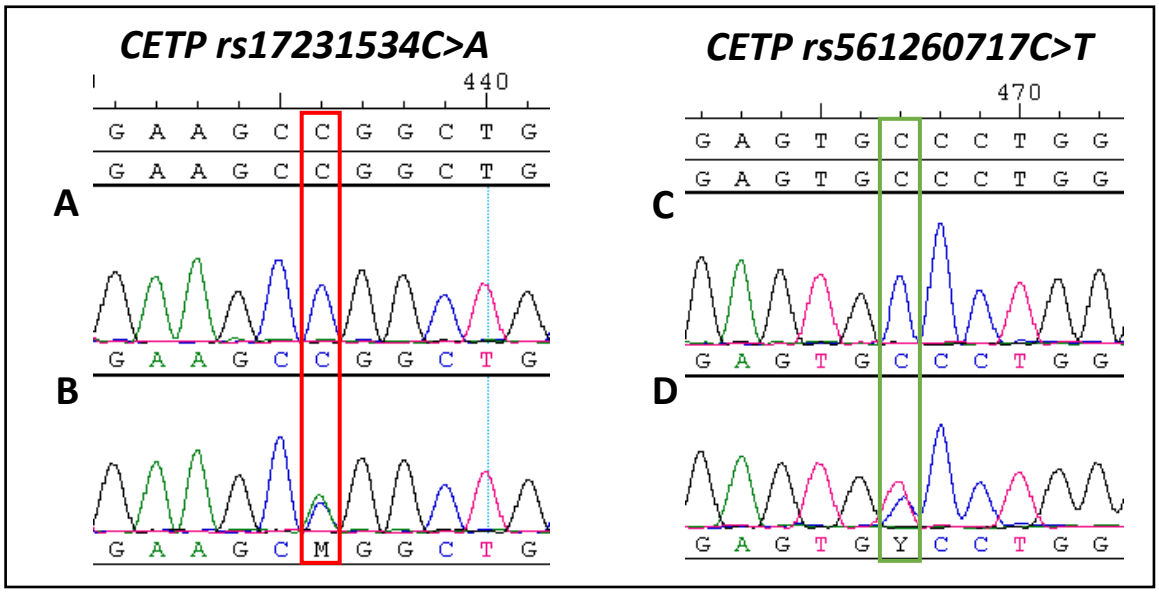


Figure 5.7: Electropherogram 3 of sequenced *CETP* PCR amplified fragments. The *CETP rs17231534C>A* and *CETP rs561260717C>T* variants are highlighted by the red and orange boxes, respectively. **A** represents a sample with the *CETP rs17231534C/C* genotype. **B** represents a sample with the *CETP rs17231534C/A* genotype. **C** represents a sample with the *CETP rs561260717C/C* genotype. **D** represents a sample with the *CETP rs561260717C/T* genotype.

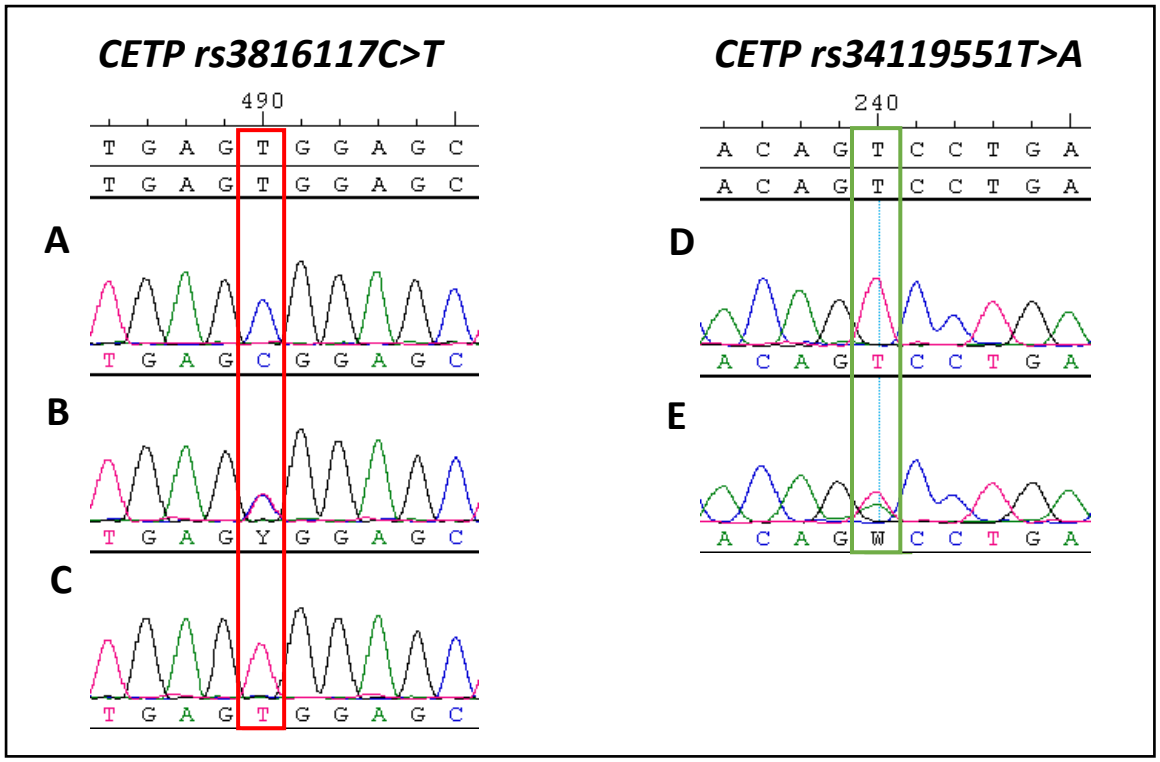


Figure 5.8: Electropherogram 4 of sequenced *CETP* PCR amplified fragments. The *CETP rs3816117C>T* and *CETP rs34119551T>A* variants are highlighted by the red and orange boxes, respectively. **A** represents a sample with the *CETP rs3816117C/C* genotype. **B** represents a sample with the *CETP rs3816117C/T* genotype. **C** represents a sample with the *CETP rs3816117T/T* genotype. **D** represents a sample with the *CETP rs34119551T/T* genotype. **E** represents a sample with the *CETP rs34119551T/A* genotype.

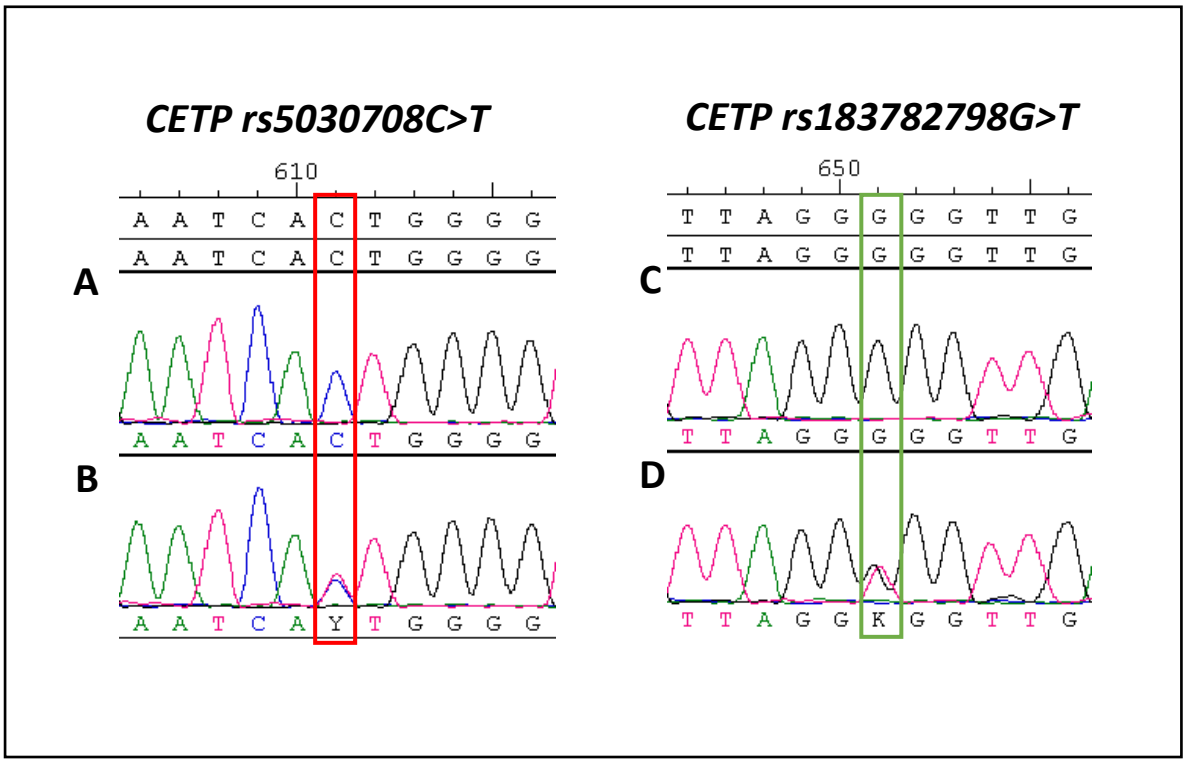


Figure 5.9: Electropherogram 5 of sequenced *CETP* PCR amplified fragments. The *CETP rs5030708C>T* and *CETP rs183782798G>T* variants are highlighted by the red and orange boxes, respectively. **A** represents a sample with the *CETP rs5030708C/C* genotype. **B** represents a sample with the *CETP rs5030708C/T* genotype. **C** represents a sample with the *CETP rs183782798G/G* genotype. **D** represents a sample with the *CETP rs183782798GG/T* genotype.

Appendix V: *PCSK9* and *CETP* univariate non-significant associations with lipid profiles

Table 5.2: Univariate analysis of association between *PCSK9 rs505151G>A*, lipid profile, and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***PCSK9 rs505151G>A* Combined**					
A/A (n=227)	6.60 (2.70 – 28.90)	4.93 (1.78 – 13.28)	1.66 (0.47 – 14.53)	1.25 (0.15 – 2.94)	2.68 (0.48 – 6.08)
A/G (n=158)	6.70 (2.20 – 17.60)	4.93 (1.64 – 7.66)	1.70 (0.55 – 5.34)	1.26 (0.11 – 2.26)	2.69 (0.60 – 5.61)
G/G (n=29)	6.00 (4.00 – 16.70)	4.87 (2.87 – 7.44)	1.29 (0.33 – 13.67)	1.39 (0.56 – 1.85)	2.74 (1.24 – 4.71)
Global P value	0.569	0.185	0.159	0.918	0.826
***PCSK9 rs505151G>A* – Sim+**					
A/A (n=156)	7.10 (3.60 – 28.90)	5.54 (2.04 – 13.28)	1.98 (0.47 – 14.53)	1.25 (0.15 – 3.42)	3.15 (0.48 – 6.08)
A/G (n=112)	6.90 (2.20 – 17.60)	5.36 (1.64 – 7.66)	1.86 (0.56 – 5.34)	1.29 (0.19 – 2.94)	2.94 (0.60 – 5.61)
G/G (n=21)	6.00 (4.00 – 16.70)	5.24 (3.40 – 7.44)	1.41 (0.40 – 13.67)	1.36 (0.79 – 2.42)	2.98 (1.86 – 4.14)
Global P Value	0.245	0.291	0.255	0.972	0.743
***PCSK9 rs505151G>A* – Sim-**					
A/A (n=51)	5.40 (2.70 – 17.30)	3.64 (1.78 – 7.64)	1.34 (0.54 – 3.69)	1.01 (0.11 – 2.09)	1.89 (0.76 – 5.17)
A/G (n=30)	5.70 (2.60 – 11.80)	3.67 (2.07 – 4.96)	1.13 (0.55 – 4.84)	0.96 (0.19 – 1.87)	1.88 (0.71 – 3.29)
G/G (n=5)	5.25 (4.70 – 7.00)	4.16 (2.87 – 4.90)	1.12 (0.33 – 2.49)	1.17 (0.67 – 1.37)	2.73 (1.24 – 3.09)
Global P Value	0.911	0.596	0.469	0.775	0.339

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c; Sim+ = patients on statins; Sim- = patients not on any statins

Table 5.3: Univariate analysis of association between *CETP rs17231520G>A*, lipid profile, and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***CETP rs17231520G>A* - Combined**					
G/G (n=352)	6.6 (2.20 – 18.40)	4.90 (1.64 – 11.32)	1.73 (0.33 – 14.53)	**1.20 (0.11 – 3.42)**	2.71 (0.48 – 6.08)
G/A+ A/A (n=42+1)	7.35 (3.10 – 17.30)	4.94 (2.97 – 8.03)	1.41 (0.54 – 4.96)	**1.36 (0.22 – 2.64)**	2.65 (1.46 – 5.17)
Global P value	0.249	0.687	0.078	**0.023**	0.709
***CETP rs17231520G>A* – Sim+**					
G/G (n=247)	6.9 (2.2 – 17.6)	5.38 (1.64 – 11.32)	1.95 (0.40 – 14.53)	**1.24 (0.15 – 3.42)**	3.06 (0.48-6.08)
G/A+ A/A (n=30+1)	8.2 (5.1 – 16.7)	5.43 (3.38 – 5.43)	1.48 (0.59 – 4.96)	**1.43 (0.22 – 2.64)**	3.02 (1.58-5.05)
Global P value	0.222	0.509	0.085	**0.026**	0.593
***CETP rs17231520G>A* – Sim-**					
G/G (n=71)	5.5 (2.60 – 12.10)	3.67 (1.78 – 5.70)	1.32 (0.33 – 4.84)	0.99 (0.11-1.87)	1.91 (0.71 – 3.29)
G/A+ A/A (n=11+0)	5.65 (3.10 – 17.30)	3.58 (2.97 – 7.64)	1.21 (0.54 – 2.81)	1.23 (0.27-1.71)	1.91 (1.46 – 5.17)
Global P value	0.341	0.967	0.586	0.462	0.581

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c; Sim+ = patients on statins; Sim- = patients not on any statins

Table 5.4: Univariate analysis of association between *CETP rs711752G>A*, lipid profile, and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***CETP rs711752G>A* – Combined**					
G/G (n=141)	6.55 (3.6 – 18.40)	4.81 (1.87 – 7.76)	1.74 (0.40 – 13.67)	1.17 (0.19 – 2.51)	2.69 (0.60 – 5.54)
G/A (n=199)	6.80 (2.20 – 17.60)	4.95 (1.78 – 11.32)	1.62 (0.33 – 14.53)	1.24 (0.11 – 3.42)	2.74 (0.48 – 6.08)
A/A (n=54)	6.45 (2.70 – 16.70)	5.03 (1.64 – 7.72)	1.58 (0.59 – 5.81)	1.26 (0.59 – 2.94)	2.66 (0.68 – 5.46)
Global P value	0.929	0.448	0.823	0.066	0.729
***CETP rs711752G>A* - On Statins**					
G/G (n=99)	6.90 (3.60 – 16.70)	5.24 (2.11 – 7.76)	1.88 (0.40 – 13.67)	1.24 (0.19 – 2.51)	3.02 (0.60 – 5.54)
G/A (n=136)	7.10 (2.20 – 17.60)	5.55 (2.01 – 11.32)	1.78 (0.56 – 14.53)	1.29 (0.15 – 3.42)	3.19 (0.48 – 6.08)
A/A (n=43)	6.80 (4.40 – 16.70)	5.35 (1.64 – 7.72)	1.82 (0.59 – 5.81)	1.31 (0.59 – 2.94)	2.91 (0.68 – 5.46)
Global P Value	0.845	0.677	0.708	0.271	0.339
***CETP rs711752G>A* – Not on Statins**					
G/G (n=29)	5.40 (3.90 – 12.10)	3.66 (1.87 – 4.90)	1.24 (0.55 – 3.69)	0.93 (0.33 – 1.87)	1.89 (0.71 – 3.09)
G/A (n=45)	5.55 (2.60 – 17.30)	3.64 (1.78 – 5.70)	1.23 (0.33 – 4.84)	1.12 (0.11 – 1.57)	1.92 (0.97 – 3.29)
A/A (n=8)	5.40 (2.70 – 11.10)	3.97 (2.60 – 7.64)	1.65 (0.66 – 2.17)	1.03 (0.67 – 1.71)	2.32 (1.29 – 5.17)
Global P Value	0.986	0.675	0.844	0.694	0.306

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c; Sim+ = patients on statins; Sim- = patients not on any statins

Table 5.5: Univariate analysis of association between *CETP rs5884C>A*, lipid profile, and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***CETP rs5884C>A* - Combined**					
C/C (n=357)	6.70 (2.20 –18.40)	4.90 (1.64 – 11.32)	1.65 (0.33 – 14.53)	1.23 (0.11 – 3.42)	2.67 (0.48 – 6.08)
C/A + A/A (n=36+2)	6.50 (2.90 – 15.20)	5.03 (3.03 – 8.03)	1.62 (0.47 – 6.03)	1.27 (0.11 – 2.64)	3.00 (1.27 – 5.61)
Global P Value	0.267	0.375	0.546	0.288	0.399
***CETP rs5884C>A* – Sim+**					
C/C (n=249)	7.10 (2.20 –17.60)	5.41 (1.64 – 11.32)	1.96 (0.40 - 14.53)	1.27 (0.15 – 3.42)	3.06 (0.48- 6.08)
C/A + A/A (n=27+2)	6.60 (4.40 - 15.20)	5.38 (3.03 – 8.03)	1.49 (0.47 – 6.03)	1.36 (0.66 – 2.64)	3.16 (1.55- 5.61)
Global P Value	0.199	0.709	0.081	0.347	0.497
***CETP rs5884C>A* – Sim-**					
C/C (n=76)	5.50 (2.60 – 17.30)	3.67 (1.78 – 7.64)	1.24 (0.33 – 4.84)	0.99 (0.11 – 1.71)	1.93 (0.71 – 5.17)
C/A + A/A (n=6+0)	4.65 (2.90 – 8.20)	3.59 (3.23 – 4.70)	1.76 (0.77 – 3.60)	1.21 (0.11 – 1.87)	1.77 (1.27 – 2.09)
Global P Value	0.236	0.776	0.359	0.398	0.227

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c; Sim+ = patients on statins; Sim- = patients not on any statins

Table 5.6: Univariate analysis of association between *CETP rs34680782C>A*, lipid profile, and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***CETP rs34680782C>A* - Combined**					
C/C (n=384)	6.70 (2.20 –18.40)	4.90 (1.64 – 11.32)	1.64 (0.33 – 14.53)	1.23 (0.11 – 3.42)	2.69 (0.48 – 6.08)
C/A + A/A (n=11+1)	6.00 (4.50 – 15.90)	5.85 (2.96 – 6.68)	1.56 (0.59 – 4.52)	0.92 (0.58 – 2.51)	3.53 (1.32 – 4.51)
P Value	0.791	0.304	0.857	0.173	0.117
***CETP rs34680782C>A* – On Statins**					
C/C (n=271)	7.00 (2.20 –17.60)	5.36 (1.64 – 11.32)	1.84 (0.40 - 14.53)	1.27 (0.15 – 3.42)	3.05 (0.48- 6.08)
C/A + A/A (n=6+1)	6.50 (4.90 - 15.90)	6.13 (4.01 – 6.68)	1.88 (0.78 – 3.56)	1.31 (0.83 – 2.51)	3.68 (2.82- 4.15)
P Value	0.931	0.183	0.819	0.869	0.145
***CETP rs34680782C>A* – Not On Statins**					
C/C (n=80)	5.50 (2.60 – 17.30)	3.67 (1.78 – 7.64)	1.26 (0.33 – 4.84)	1.02 (0.11 – 1.71)	1.91 (0.71 – 5.17)
C/A + A/A (n=2+0)	5.60 (5.40 – 5.80)	3.36 (2.96 – 3.75)	1.39 (0.22 – 1.56)	1.21 (0.11 – 1.87)	1.94 (1.32 – 2.56)
P Value	NA (n= C/A too small)	NA (n= C/A too small)	NA (n= C/A too small)	NA (n= C/A too small)	NA (n= C/A too small)

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c; Sim+ = patients on statins; Sim- = patients not on any statins

Table 5.7: Univariate analysis of association between *CETP rs17231534C>A*, lipid profile, and HbA1c% in the combined T2DM cohort, and after stratifying by statin usage

Variant	Median HbA1c (%)	Median Total Cholesterol (mmol/L)	Median Triglycerides (mmol/L)	Median HDL-C (mmol/L)	Median LDL-C (mmol/L)
***CETP rs17231534C>A* - Combined**					
C/C (n=357)	6.60 (2.20 –18.40)	4.42 (1.64 – 10.99)	1.67 (0.33 – 13.67)	1.23 (0.11 – 3.42)	2.70 (0.48 – 6.08)
C/A (n=38)	7.00 (4.00 – 16.70)	4.78 (3.03 – 11.32)	1.48 (0.68 – 14.53)	1.29 (0.41 – 2.80)	2.51 (0.90 – 5.25)
Global P Value	0.966	0.512	0.460	0.237	0.324
***CETP rs17231534C>A* – Sim+**					
C/C (n=250)	6.90 (2.20 –17.60)	5.42 (1.64 – 10.99)	1.86 (0.40 - 13.67)	1.26 (0.15 – 3.42)	3.15 (0.48- 6.08)
C/A (n=28)	7.20 (4.00 – 16.70)	5.14 (3.03 – 11.32)	1.71 (0.74 – 14.53)	1.40 (0.41 – 2.80)	2.75 (0.90 – 5.25)
Global P Value	0.655	0.308	0.799	0.181	0.086
***CETP rs17231534C>A* – Sim-**					
C/C (n=76)	5.50 (2.60 – 17.30)	3.65 (1.78 – 7.64)	1.26 (0.33 – 4.84)	1.02 (0.11 – 1.87)	1.90 (0.71 – 5.17)
C/A (n=6)	5.80 (5.00 – 6.10)	3.92 (3.11 – 4.90)	1.27 (0.68 – 2.49)	0.93 (0.56 – 1.39)	2.11 (0.82 – 3.09)
Global P Value	0.793	0.332	0.993	0.482	0.154

Values represent median (range); Bold indicates significance $p<0.05$; High-density lipoprotein cholesterol = HDL-c; Low-density lipoprotein cholesterol = LDL-c; HbA1c = Haemoglobin A1c; Sim+ = patients on statins; Sim- = patients not on any statins

Appendix VI: Consent Forms and Information Packet:

Study title: GENETICS AND DIABETIC DYSLIPIDAEMIA IN THE BLACK SOUTH AFRICANS

Greeting: Good day, my name is Donald Tanyanyiwa. I work in Chemical Pathology here at Chris Hani Baragwanath Academic Hospital

Introduction:

We are doing some work to help us understand causes of increased fat /lipids in some African people with diabetes. It is known that some people with diabetes develop abnormal levels of fats/lipids. People are born with fat carrying protein, which may be different because of changes (mutations/polymorphism) in the seed (genes) that produce them and that causes them to do their job differently; others do it faster than others. The change may occur in one seed or many seeds. This causes some people to have increased amounts of different types of fat because their fat carrying proteins will not be working properly or may be reduced. Fats/Lipids are required by the body, but when in excess it becomes harmful. The body normally manages to keep the fat within the required levels, but during illness or the presence of inherited disease, the body may not be able to clear the fat.

High amounts of these fats have been associated with disruption of the normal heart and blood transport system. A lot of research to identify the different causes of increased amount has been done but very little if any has been done to identify the common causes of fat accumulation in African people with diabetes. The study will examine to see if mutations/polymorphism in the genes of some (four proteins) associated with lipid metabolism are responsible for the abnormal fat/lipid (cholesterol and triglyceride) content in people with diabetes. The study will require only a small amount of blood (1.5 teaspoons) of the blood already sent to the laboratory for diagnostics purposes. There will be no immediate benefits to you but depending on the results benefits may be realised later. Patients who do not participate or withdraw from the study will continue to receive the same recommended standard intervention as the participants. It is hoped that the findings and advice given to the patients will go a long way in improving the quality of life. Venesection is a very safe method of obtaining blood for the repeat lipogram if required and the only risk that can be encountered is prolonged bleeding in those with some coagulopathy, a condition that will be excluded by the medical history on previous venesection.

Invitation to participate: We saw from your results that you are one of those diabetic people with increased amounts of these fats. Usually a small amount of blood (one and half (1½) teaspoons) is sent to test fat content in blood. We are asking your permission to keep that blood and find out the reason why the fat is increased

What is involved in the study – We will not need any blood from you and we will not examine you. However we will require your permission to go and look at your hospital/clinic file to get your age, and how you are being treated. If you are taking tablets we will record their names. You will not be expected to make any special visit to Chris Hani Baragwanath Academic Hospital for this study.

Risks: **There is no risk at all in this study because the fat carrying protein is already in the blood you sent to the laboratory and that is the one we will be using.**

Benefits - **The benefits of this study will only be realized if the cause of increased fat is identified and because it is the first such study, I will present it for as a PhD thesis.**

Participation is voluntary – It is your right to accept or refuses to take part in this study and whatever decision you make nothing will change regarding your treatment. It is also your right to tell us to destroy the blood that is already in the laboratory.

CONSENT FOR STYDY ON DIABETIC DYSLIPIDAEMIA IN BLACK SOUTH AFRICANS

Name----------

Address ----------

Analysis and storage of Biologic Samples: Plasma and DNA

I hereby consent to the removal, storage waiting processing, and analysis of the above material from my own body for the purpose of diagnosis and research into disorders of lipid and lipoprotein metabolism associated with diabetes. After due explanation I understand that

1) Conventional procedures and techniques are employed and that the health risk is minimal.

2) The material and results of the investigations remain strictly confidential according to medical practice and such ethical guidelines that govern research at the universities and country at large. To preserve anonymity the samples are coded by numbers and my written consent is required for release of identifiable information to another party

3) Precise diagnosis may not always be possible because the defect(s) may not yet be known or there is inadequate information to derive the defect owing to modulatory roles that other genes or the environment may play.

4) The stored material may be used anonymously in future to derive information or for research purposes. Such future use may be of no direct benefit to the subject.

5) Permission to participate in the study may be withdrawn at any time and any stored biological material will also be destroyed. The withdrawal will not affect the subject's future medical care.

6) I confirm that the purpose of this study was explained in the language that I preferred and I fully understood it. I have not received any financial benefit and I do not expect to receive any in future and undertake not to make any future financial claims following the outcome and publication of the finding.

Singed at ----------day ----------month---------2013

Print Name ---------- Signature----------

Witness (Subject's Choice)

Print Name ---------- Signature----------

Witness (Health Personnel)

Print Name ---------- Signature----------

Appendix VII: Ethics Clearance

UNIVERSITY OF CAPE TOWN
Faculty of Health Sciences
Human Research Ethics Committee

Room E52-24 Old Main Building
Groote Schuur Hospital
Observatory 7925
Telephone [021] 404 7682 • **Facsimile** [021] 406 6411
Email: nosi.tsame@uct.ac.za
Website: www.health.uct.ac.za/fhs/research/humanethics/forms

01 February 2017

HREC REF: 750/2016

Prof C Dandara
Human Genetics
N3.14.4
Werner Belt South

Dear Prof Dandara

PROJECT TITLE: GENETIC SUSCEPTIBILITY TO ABNORMAL LIPID LEVELS IN SOUTH AFRICANS (MSc-candidate J Evans) Sub-study linked to 089/2013

Thank you for submitting your response letter to the Faculty of Health Sciences Human Research Ethics Committee dated 23 January 2017.

It is a pleasure to inform you that the HREC has **formally approved** the above-mentioned study.

Approval is granted for one year until the 28th February 2018.

Please submit a progress form, using the standardised Annual Report Form if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.
(Forms can be found on our website: www.health.uct.ac.za/fhs/research/humanethics/forms)

We acknowledge that the student Jonathan Evans will be involved in this study.

Please note that for all studies approved by the HREC, the principal investigator **must** obtain appropriate institutional approval before the research may occur.

Please quote the HREC REF in all your correspondence.

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Yours sincerely

PROFESSOR M BLOCKMAN
CHAIRPERSON, FHS HUMAN RESEARCH ETHICS COMMITTEE
Federal Wide Assurance Number: FWA00001637.
Institutional Review Board (IRB) number: IRB00001938

Research Council (MRC-SA), Food and Drug Administration (FDA-USA), International Convention on Harmonisation Good Clinical Practice (ICH GCP), South African Good Clinical Practice Guidelines (DoH 2006), based on the Association of the British Pharmaceutical Industry Guidelines (ABPI),and Declaration of Helsinki (2013) guidelines.
The Human Research Ethics Committee granting this approval is in compliance with the ICH Harmonised Tripartite Guidelines E6: Note for Guidance on Good Clinical Practice (CPMP/ICH/135/95) and FDA Code Federal Regulation Part 50, 56 and 312.