Screening black South African females with Type 2 Diabetes Mellitus for mutations in the peroxisome proliferator-activated receptor gamma gene.

By

Elzette Nienaber

January 2016

Submitted in fulfillment of the requirements for the M.Med.Sc (Human Molecular Biology) degree

Faculty of Health Sciences Department of Haematology and Cell Biology University of the Free State, Bloemfontein

Supervisor: Dr. G.M. Marx

Co-Supervisor: Dr. D. Goedhals

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA

UFS UV

DECLARATION

I certify that the dissertation hereby submitted by me for the M.Med.Sc (Human Molecular Biology) degree at the University of the Free State is my independent effort and had not previously been submitted for a degree at another university or faculty. I furthermore waive copyright of the dissertation in favour of the University of the Free State.

_____

Elzette Nienaber

ACKNOWLEDGEMENTS

I would like to thank the following people who made this study possible:

- Dr Marx for her guidance and motivation and providing me with expert advice to perform this study.
- Dr Goedhals for her assistance and support.
- Mr A. Bester for his insight and guidance and teaching me everything I had to know about next generation sequencing analysis.
- National Research Foundation (NRF) Thuthuka grant for the financial support enabling me to complete the study.
- The Department of Haematology and Cell Biology for providing the necessary resources and facilities.
- The individuals participating in this study, the patients and controls, this study would not have been possible without you.
- My colleagues and friends at the Department of Haematology and Cell Biology for their support, motivation and friendship during this study.
- My parents, family and friends for their encouragement, prayer and for always believing in me.
- Lastly to my heavenly Father for always being the firm rock beneath my feet.

"Whatever you do, work at it with all your heart, as working for the Lord, not for men, since you know that you will receive an inheritance from the Lord as a reward. It is the Lord you are serving."

Colossians 3:23

CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS

%	Percentage
°C	Degree Celsius
3’	3 prime
5’	5 prime
A	Adenine
AA	Amino acid
ADA	American Diabetes Association
ADIPOQ	Adiponectin
AF	Activation function
AFR	African population
Ala	Alanine
ALL	General population
AMR	American population
BBQ	BlackBerry Quencher
BMI	Body Mass Index
bp	Base pairs
BWA	Burrows Wheeler Aligner
C	Cytosine
CDK5	Cyclin-dependent kinase 5
CDKN2A	Cyclin-dependent kinase inhibitor 2A
CDS	Coding Sequence
CEO	Chief Executive Officer
Chr3	Chromosome 3

CSV	Comma Separated Value
Ct	Cycle threshold
CTRB1/2	Chymotrypsinogen B1/2
DBD	DNA-binding domain
DCCT	Diabetes Control and Complications Trial
DM	Diabetes Mellitus
DNA	Deoxyribonucleic acid
EDTA	Ethylenediamine tetra acetic acid
et al.	*et alia* (and others)
EAS	Asian population
ECUFS	Ethics Committee of the Faculty of Health Sciences of the Free State
EUR	European population
6FAM	6 Carboxyfluorescein
Fig.	Figure
FPG	Fasting Plasma Glucose
FRET	Fluorescent resonance energy transfer
FTO	Fat mass and obesity-associated gene
g	Gram
G	Guanine
GB	Gigabyte
GDM	Gestational Diabetes Mellitus
GLIS3	GLIS family zinc finger 3
GLM1	Glioma susceptibility 1
Gln	Glutamine
GRC	Genome Reference Consortium

GRCh37	Genome Reference Consortium Human genome build 37
GWAS	Genome Wide Association Studies
HbA1c	Haemoglobin A1c
HEX	Hexachlorofluorescein
Hg19	Human genome number 19
HLA	Human Leukocyte antigen A
HPV	High-Performance cluster
IDF	International Diabetes Federation
IDT	Integrated DNA Technologies
IFG	Impaired fasting glucose
IFIH1	Interferon induced with helicase C domain 1
IGT	Impaired glucose tolerance
IGV	Integrative Genomics Viewer
IRS	Insulin receptor substrate
Kbp	Kilo base pairs
KCJ11	Potassium channel subfamily J member 11
Kg	Kilogram
LBD	Ligand-binding domain
Leu	Leucine
M	Molar
Mb	Mega base pair
Met	Methionine
min	Minutes
ml	Millilitre
mmol/l	Millimole per litre

MODY	Maturity-Onset-Diabetes of the Young
mRNA	Messenger ribonucleic acid
NCBI	National Centre for Biotechnology Information
NF-kB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NGS	Next generation sequencing
NGSP	National Glycohemoglobin Standardization Program
Nr	Number
NTC	No template control
OGTT	Oral glucose tolerance test
ORF	Open reading frame
p	Short arm of chromosome
PCR	Polymerase chain reaction
pH	Concentration of hydrogen ions in solution
PPARG (γ)	Peroxisome proliferator-activated receptor gamma
PPARs	Peroxisome proliferator-activated receptors
PPARα	Peroxisome proliferator-activated receptor alpha
PPARβ	Peroxisome proliferator-activated receptor beta
PPRE	Proliferator response element
PRINSEQ	PReprocessing and INformation of SEQuence data
Pro	Proline
PTPN2	Protein tyrosine phosphatase non-receptor type 2
QC	Quality Control
RNA	Ribonucleic acid
RPG	Random plasma glucose
rs	Reference single nucleotide polymorphism number

RXR	Retinoid X receptors
Sec	Seconds
®	Registered trademark
SAS	South Asian population
SEMDSA	The Society for Endocrinology, Metabolism and Diabetes of South Africa
SNP	Single nucleotide polymorphism
SUMO	Small ubiquitin-like modifier
T	Thymine
T1DM	Type 1 diabetes mellitus
T2DM	Type 2 diabetes mellitus
Taq	*Thermus aquaticus*
TCF7L2	Transcription factor 7 like 2
TE	Tris-EDTA
Tm	Melting temperature
TNFα	Tumor necrosis factor alpha
Tris	Tris hydroxymethyl aminomethane
TZD	Thiazolidinediones
™	Trademark
UCSC	University of California, Santa Cruz
USD	United States Dollar
UTR	Untranslated Region (3'/5')
Val	Valine
VCF	Variant Call Format
WHO	World Health Organization
www	World wide web

x g	Acceleration due to gravity
α	Alpha
β	Beta
γ	Gamma
µl	Micro-litre
µM	Micro-molar

LIST OF FIGURES

Page

LIST OF TABLES

CHAPTER ONE

LITERATURE REVIEW

1.1 Diabetes mellitus

Diabetes mellitus (DM) is recognized as one of the leading causes of disability and death worldwide (International Diabetes Federation, 2014). In 1998, the World Health Organization (WHO) estimated that there were 135 million people living with DM in the world, and by 2008 the estimated global prevalence had more than doubled (Danaei *et al.*, 2011). As of 2013, there are 387 million people living with DM in the world and 2.7 million in South Africa (International Diabetes Federation, 2014). Understanding the main contributors to disease burden is essential in the planning of health care facilities and programmes aimed at addressing this growing problem. This study aims to shed some light on this issue by investigating a specific risk factor that has been associated with DM in the black South African population.

DM describes a group of metabolic diseases of multiple aetiologies characterized by increased levels of glucose that results from defects in insulin action, insulin secretion, or both (Alberti and Zimmet, 1998). Glucose absorption is impaired in DM patients and glucose remains circulating in the blood, which is known as hyperglycaemia. The development of DM involves several pathological processes that range from autoimmune destruction of the β-cells of the pancreas, resulting in insulin deficiency, to abnormalities causing insulin resistance. The abnormalities in the metabolism of carbohydrate, fat and protein of diabetic individuals are a result of the insufficient action of insulin on target tissues (American Diabetes Association, 2008). Uncontrolled DM can lead to acute, life-threatening consequences, such as hyperglycaemia with ketoacidosis or the non-ketotic hyperosmolar syndrome. Long-term complications of DM include retinopathy with potential loss of vision; nephropathy leading to renal failure; peripheral neuropathy with risk of foot ulcers, amputations, and Charcot joints; and autonomic neuropathy causing gastrointestinal, genitourinary, and cardiovascular

symptoms, as well as sexual dysfunction (American Diabetes Association, 2008; Weinger *et al.*, 1995). Furthermore, people suffering from DM have an increased risk for other diseases, such as cardiac, peripheral arterial and cerebrovascular disease (Amod *et al.*, 2012; Mukherjee, 2009). DM is a chronic illness that requires continuing medical care and ongoing patient self-management education and support to prevent acute complications and to reduce the risk of long-term complications.

1.2 Prevalence and burden of disease

In 2013, there were an estimate of 387 million people living with DM in the world, with the majority aged between 40 and 59 years. It is estimated that the number of people with DM will increase by 55% by 2035 (International Diabetes Federation, 2014). Due to lifestyle changes, the prevalence of DM is drastically increasing in the Middle East, South-East Asia, Western Pacific and sub-Saharan Africa. Eighty percent of people affected are living in low- and middle- income countries. According to the International Diabetes Federation (IDF), Africa is estimated to have the highest increase of DM cases in the world with an increase of 109% predicted by 2035. With a national prevalence of 8.27%, there are an estimated 2.7 million people in South Africa currently diagnosed with DM and more than 1.2 million people with undiagnosed DM (International Diabetes Federation, 2014). Rapid urbanisation in South Africa is leading to increased lifestyle risk factors, such as a lack of physical activity and unhealthy diet which contribute to increased chronic disease rates (Steyn *et al.*, 1997).

DM exerts a heavy economic burden on society (Kirigia *et al.*, 2009), related to costs on the health system in managing the disease, indirect costs resulting from productivity losses due to patient disability and premature mortality, time spent by family members accompanying patients when seeking care, and intangible costs such as pain, anxiety, inconvenience and a negative influence on personal relationships (Kirigia *et al.*, 2009; WHO, 2015). In 2012, globally the burden of DM resulted in 5.1 million deaths and consumed 548 billion USD in health spending (Amod *et al.*, 2012). According to 2013 estimates for the Africa Region, at least 4 billion USD was spent on DM healthcare, and

this spending is expected to increase by around 58% by 2035 (International Diabetes Federation, 2014). In South Africa, the average DM related expenditure per person with diabetes is 935 USD annually. There is substantial evidence that DM is epidemic in many economically developing and newly industrialised countries. Therefore, there is an urgency for innovative research on DM to alleviate the burden that the disease has on our society and to diagnose and provide appropriate care to people with DM.

1.3 Gender and ethnic influences

Globally in 2013, there was little gender difference in the incidence of DM (International Diabetes Federation, 2014). There are about 14 million more men than women with DM (198 million men vs. 184 million women). However, this difference is expected to increase to 15 million (303 million men vs 288 million women) by 2035 (Gale and Gillespie, 2001). In South Africa, there is a prominent gender difference with 927,870 males suffering from DM compared to 1,718,180 females (International Diabetes Federation, 2014).

It has been established that in many countries of sub-Saharan Africa, including South Africa, the women are more likely to be obese or overweight than men and are therefore expected to have a higher prevalence of DM (Dugas *et al.*, 2009; Kamadjeu *et al.*, 2006; Njelekela *et al.*, 2009; Omar *et al.*, 1993; Puoane *et al.*, 2002). A study by Puoane *et al*. (2002) suggested that the predominant cause of malnutrition in adult South Africans is due to being overweight with high rates of abdominal obesity, especially in African women. The biggest challenge in obesity management in South Africa relates to the perceptions and positive values ascribed to obesity in the African community (Mvo *et al.*, 1999). Qualitative research by Mvo *et al.* (1999) identified that obesity in women is perceived to reflect wealth and happiness as well as the husband's ability to care for his wife. Although wide variations in the distribution of DM by gender have been documented in several review articles (BeLue *et al.*, 2009; Gale and Gillespie, 2001; Gill *et al.*, 2009; Tuei *et al.*, 2010), the possible causes of this heterogeneity have never been examined in detail.

The prevalence of DM has been found to differ between ethnic groups. A study by Cowie *et al.* (2010) suggested that the prevalence of DM was more than two times higher in non-Hispanic blacks and Mexican Americans versus non-Hispanic Caucasians (Cowie *et al.*, 2010). African Americans are disproportionately affected by DM with a prevalence of 18% (Cooke *et al.*, 2012) and are 1.8 times more likely to develop DM than Caucasians (Chlebowy *et al.*, 2013; Wei *et al.*, 2011). This may be due to a direct genetic propensity or unfavourable gene–environment interactions. The hypothesis has been proposed that modern lifestyle factors (especially those that promote obesity) may have a greater effect on African Americans than on Caucasians (Abate and Chandalia, 2003; Signorello *et al.*, 2007).

Many studies have been done on European-derived populations, and a few on African Americans, but there remains a gap for studies that address and investigate risk factors associated with African populations. In general, the black population in South Africa predominates over other population sub-groups. It was reported that the highest rates of obesity are among African women when compared to Caucasian, Asian and mixed ancestry (Puoane *et al.*, 2002). A high prevalence of DM has been reported in urban black populations in Cape Town and in those of Zulu descent (Levitt *et al.*, 1993; Omar *et al.*, 1993). The association of DM with urbanization has important implications in view of the large-scale urbanization occurring in southern Africa (Levitt *et al.*, 1993).

1.4 Diagnosis of diabetes mellitus

The three principal methods to diagnose DM include the fasting plasma glucose (FPG) test, random plasma glucose (RPG) test and oral glucose tolerance test (OGTT) (American Diabetes Association, 2008). Since 2011 the WHO and the American Diabetes Association (ADA) has also recognized haemoglobin A1c (HbA1c) as a means of diagnosing DM (World Health Organization, 2011). The Society for Endocrinology, Metabolism and Diabetes of South Africa (SEMDSA) has adopted and endorsed the report of the WHO on the use of HbA1c in the diagnosis of DM (Amod *et al.*, 2012).

1.4.1 FPG, RPG, and OGTT

The FPG test measurement requires the patient to fast overnight for at least 8 hours as well as a follow-up clinic visit which may cause some inconvenience but is inexpensive and risk-free. The accuracy of the test may, however, be affected by the patient not adhering to fasting, and certain medication can compromise the test. The RPG measurement is inexpensive, easily accomplished and free of risk, except for discomfort from phlebotomy (Barr *et al.*, 2002). The OGTT was originally described in 1922 as a measure of the ability to tolerate a supraphysiologic glucose load (Conn 1940). Since the 1970s, the WHO and other organizations interested in DM agreed on a standard dose and duration, but the ADA no longer recommends it for routine use in non-pregnant adults (International Expert, 2009).

1.4.2 HbA1c

Tests of glycated haemoglobin concentration yield a measure of chronic glycaemia from the slow, post-translational, non-enzymatic glycation of haemoglobin (Goldstein *et al.*, 1982). Clinical studies have indicated a strong correlation between the concentration of glycated haemoglobin and the mean level of blood glucose over the preceding one to three months (Nathan *et al.*, 1984; Svendsen *et al.*, 1982; Tahara and Shima, 1995). The HbA1c test should be performed using a method that is certified by the National Glycohemoglobin Standardization Program (NGSP) and standardized or traceable to the Diabetes Control and Complications Trial (DCCT) reference assay. The HbA1c test has several advantages compared to the OGTT and FPG test. It is more convenient, as individuals are not required to fast, has greater pre-analytical stability and fewer perturbations during illness or stress. It is extremely reliable compared with other tests of glycaemia. Some of the limitations, however, include higher cost, limited availability in certain regions of the developing world and some issues with standardization (American Diabetes Association, 2015). DM-specific complications can also be predicted with the HbA1c levels and provide the current basis for diabetic treatment decisions. Small elevations in HbA1c levels can predict future DM in people with impaired glucose tolerance and mild impairment in glycaemia (American Diabetes Association, 2015).

1.4.3 Diagnostic criteria for diabetes mellitus

The criteria for the diagnosis of DM as well as intermediate hyperglycaemia according to the 2012 SEMDSA guidelines for management of DM are outlined in Table 1.1. The diagnosis of DM should always be confirmed by a repeat test (preferably the same test) on a subsequent day, unless there is unequivocal hyperglycaemia with acute metabolic decompensation or obvious symptoms (Amod *et al.*, 2012).

Table 1.1: Criteria for the diagnosis of diabetes and categories of intermediate hyperglycaemia according to the SEMDSA guidelines (Amod *et al.*, 2012).

Diagnostic test	Impaired fasting glucose (IFG)	Impaired glucose tolerance (IGT)	Diabetes
Fasting plasma glucose (FPG) [1]	6.1-6.9 mmol/l	< 7.0 mmol/l	≥ 7.0 mmol/l; or
Two hour plasma glucose during oral glucose tolerance test (OGTT) [2]	< 7.8 mmol/l	7.8-11.0 mmol/l	≥ 11.1 mmol/l; or
Glycated haemoglobin (HbA1c)[3]	-	-	≥ 6.5%; or
Random plasma glucose[4]	-	-	≥ 11.1 mmol/l if classic symptoms of diabetes or hyperglycaemic crisis is present
1: "Fasting" is defined as no calorie intake for at least eight hours			
2: The test should be performed as described by WHO, using a glucose load containing the equivalent of 75 g anhydrous glucose dissolved in 250 ml water ingested over five minutes.			
3: Provided that the test method meets stringent quality assurance criteria, that the assay is standardised according to criteria aligned with the international reference values [NGSP -certified and standardised to the DCCT assay], and that there are no conditions present which preclude its accurate measurement.			
4: “Random” (casual) is defined as any time of day, without regard to time of last meal. The classic symptoms of hyperglycaemia include polyuria, polydipsia and weight loss. “Hyperglycaemic crisis” refers to diabetic ketoacidosis or hyperosmolar nonketotic hyperglycaemia.			

Diagnosis in symptomatic individuals: A single abnormal test is sufficient to confirm the diagnosis of DM in patients who present with classic symptoms of hyperglycaemia, such as polyuria, polydipsia and weight loss or unequivocal hyperglycaemia (glycaemic crisis: diabetic ketoacidosis or hyperosmolar non-ketotic hyperglycaemia) (Mitscherlich, 1841).
Diagnosis in asymptomatic individuals: The diagnosis of DM should not be based on a single abnormal test result in asymptomatic individuals. In these cases, a glucose-based test (OGTT or FPG) or the HbA1c test should be performed and if the test is abnormal, the same test must be repeated on a subsequent day to confirm the diagnosis (Hindorff *et al.*, 2009).

1.4.4 Criteria for diagnosis of prediabetes

An international Expert Committee sponsored by the ADA (2003), has identified an intermediate group of individuals whose glucose levels are too high to be considered normal, although not meeting the criteria to be diagnosed with DM. Patients with IFG and/or IGT are now referred to as having “prediabetes” indicating the relatively high risk of developing DM in the future. The criteria for diagnosis of IFG and IGT are provided in Table 1.1. In patients with prediabetes, identification of and, if appropriate, treatment of other cardiovascular disease risk factors is indicated. Testing to detect prediabetes should be considered in children and adolescents who are overweight or obese and who have two or more additional risk factors for DM. Without intervention, the cumulative incidence of DM being diagnosed in people with IFG is 37% to 38% over 5 to 6 years (de Vegt *et al.*, 2001).

1.5 Aetiological types of diabetes

According to the IDF, there are three main aetiological types of DM, Type 1 Diabetes Mellitus (T1DM), Type 2 Diabetes Mellitus (T2DM) and Gestational Diabetes Mellitus (GDM) (Amod *et al.*, 2012; International Diabetes Federation, 2014). Less common forms of diabetes include Maturity-Onset-Diabetes of the Young (MODY), genetic defects in insulin action, diseases of the exocrine pancreas, endocrinopathies and drug- or chemical-induced diabetes (Amod *et al.*, 2012).

1.5.1 Gestational diabetes mellitus

GDM is defined as any degree of glucose intolerance with onset or first recognition during pregnancy. Pregnant women develop a resistance to insulin and subsequent high blood glucose because the action of insulin is blocked, possibly due to hormones produced by the placenta. Complications due to GDM result in roughly 4% of all pregnancies, which amounts to 135,000 cases annually in the United States. The prevalence depends on the population studied but ranges from 1% to 14% of pregnancies (American Diabetes Association, 2008). GDM usually occurs around the 24^{th} week of pregnancy. Since GDM only develops later in pregnancy, the unborn baby is already well-formed but still growing and the immediate risk is thus less for women who develop GDM compared to those suffering from T1DM or T2DM before pregnancy. Nonetheless, high blood glucose levels can lead to serious complications in the mother and her infant. Women who become pregnant and are known to have DM that antedates pregnancy do not have GDM but have 'DM and pregnancy' and should be treated accordingly before, during, and after the pregnancy (Alberti and Zimmet, 1998). GDM in mothers usually disappears after delivery, however, these women have an increased risk to develop T2DM later in life. A range of severity is found in GDM that can be treated with a healthy diet, exercise or in some cases oral medication or insulin (IDF, 2013).

GDM is defined by glucose intolerance on the OGTT with onset during pregnancy (International Diabetes Federation, 2014). Formal systematic testing for GDM is usually done between 24 and 28 weeks of gestation (Alberti and Zimmet, 1998). Pregnant women who meet the WHO criteria for DM or IGT are classified as having GDM (See Table 1.1). After pregnancy (six weeks or more after delivery) the women should be reclassified based on OGTT as normal glucose tolerance, or IGT, or DM.

1.5.2 Type 1 diabetes mellitus

T1DM, previously described as insulin-dependent diabetes or juvenile-onset diabetes, accounts for 5% to 10% of all diagnosed DM cases (American Diabetes Association,

2015). T1DM is an autoimmune disease that is characterized by destruction of pancreatic β-cells, resulting in absolute insulin deficiency (Eisenbarth, 1986). The majority of cases are attributable to an autoimmune-mediated destruction of β-cells while a small minority of cases results from an idiopathic destruction of β-cells. The presence of autoantibodies against the pancreatic islet cells is the hallmark of T1DM, even though the role of these antibodies in the pathogenesis of the disease is not clear (Morran *et al.*, 2015).

In T1DM the rate of β-cell destruction is variable, being rapid mainly in infants or children and slower in adults. At the first manifestation of the disease, some patients may present with ketoacidosis while others can have modest fasting hyperglycaemia (American Diabetes Association, 2008). This disease affects people of all ages but is usually found in children or young adults. T1DM is considered a complex genetic trait, thus, multiple genetic loci contribute to susceptibility and environmental factors play a role in determining risk, however, these are still poorly defined. Patients diagnosed with T1DM are rarely obese, but the presence of obesity has been identified as a risk factor contributing to its development (Hypponen *et al.*, 2000; Johansson *et al.*, 1994). It has been hypothesized that both T1DM and T2DM are “one and the same disorder of insulin resistance set against different genetic backgrounds” (Wilkin, 2001). In a study by Terry Wilkins, MD, he argued that children who develop T1DM are genetically predisposed to develop the disease, but being overweight accelerates the process (Wilkin, 2001).

Several genetic risk factors are related to the autoimmune destruction of β-cells. The most prominent candidates identified as risk factors for T1DM are genes in the human leukocyte antigen A (HLA) region of chromosome 6. This region contains several hundred genes known to be involved in the immune response. The HLA class II genes have been found to be most strongly associated with the disease (Barrett *et al.*, 2009; Thomson, 1984). Other genes have also been identified to be associated with T1DM by genome-wide association studies (GWAS). They include chymotrypsinogen B1/2 (*CTRB1/2)*, interferon induced with helicase C domain 1 (*IFIH1)*, GLIS family zinc finger 3 (*GLIS3)*, and protein tyrosine phosphatase non-receptor type 2 (*PTPN2)* which are

also expressed in β-cells supporting the concept that genetic susceptibility to T1DM influences both the immune system and β-cell function (Bergholdt *et al.*, 2012).

The risk of developing T1DM in offspring where one or both parents are diabetic ranges from 2% to 30%. When only the mother is affected the risk is 2% to 3% which increases to 8% to 9% with an affected father. With both parents affected the risk increases to 30% (Amod *et al.*, 2012). Twin studies also provide evidence for the importance of environmental risk factors for T1DM. T1DM concordance rates for monozygotic twins are higher than those for dizygotic twins (approximately 30% vs. 10%) (Hirschhorn, 2003). However, most monozygotic twin pairs remain discordant. These concordance rates emphasize the importance of genetics in T1DM, but also clearly demonstrate that having certain combinations of genes is not sufficient to cause T1DM. Therefore, environmental triggers modulate the onset of T1DM in genetically susceptible individuals.

1.5.2.1 Identification of individuals at risk for developing T1DM

Individuals at an increased risk for developing T1DM can be identified by genetic markers and by serological evidence of an autoimmune pathologic process occurring in the pancreatic islets (American Diabetes Association, 2010). Studies have suggested that the measuring of islet autoantibodies in relatives of those with T1DM may identify individuals who are at risk for development of T1DM (Verge *et al.*, 1996). These tests together with education about DM symptoms and close follow-up may help with earlier identification of T1DM. Currently, there is a lack of accepted screening programs and relatives of patients diagnosed with T1DM should be advised to go for antibody testing for risk assessment (American Diabetes Association, 2015). Clinical studies are underway to test different methods of preventing T1DM in those with evidence of autoimmunity.

1.6 Type 2 diabetes mellitus

T2DM is a heterogeneous metabolic disorder characterized by two interrelated metabolic defects, namely an inadequate response to insulin secretion by β-cells in the pancreas and resistance to insulin action in multiple peripheral tissues (Groop and Pociot, 2014; Majithia *et al.*, 2014). T2DM, previously referred to as non-insulin dependent diabetes or adult-onset diabetes, is the most common type of DM and accounts for approximately 90% of all diabetes cases (Amod *et al.*, 2012). Ketoacidosis does not occur spontaneously in T2DM but usually arises in association with another illness such as an infection or stress (American Diabetes Association, 2008). Some individuals do not present with clinical symptoms of T2DM, but the degree of hyperglycaemia is sufficient to cause pathologic and functional changes in various target tissues thus, causing long-term damage. These individuals remain undiagnosed, causing the disease to manifest in mid-life (International Diabetes Federation, 2014).

1.6.1 Diagnosis and treatment of T2DM

T2DM should be diagnosed according to the 2012 SEMDSA guidelines (Amod *et al.*, 2012) based on the 2006 WHO recommendations with the addition of the use of HbA1c in diagnosis (World Health Organization, 2011) (Table 1.1). Depending on the disease state, daily dosage with insulin is usually not required to manage the disease (International Diabetes Federation, 2014). In many cases, nutrition management and an increase in physical activity are successful in improving insulin sensitivity. The aim is for patients to lose 5% to 10% of their body weight, maintain weight loss and prevent weight regain (Amod *et al.*, 2012). This is important to maintain healthy glucose levels and sustain long-term health and quality of life. Other individuals who have some residual insulin secretion may require exogenous insulin to maintain adequate glycaemic control while individuals with extensive β-cell destruction and hence no residual insulin secretion require insulin for survival (American Diabetes Association, 2008). The severity of the metabolic abnormality and consequently the degree of hyperglycaemia may vary over time and treatment should be adjusted accordingly (American Diabetes Association, 2008).

1.6.2 T2DM as a multifactorial disease

The development of T2DM is not well understood, and it is thought that the complex interactions between several genes and environmental factors (multifactorial) contribute to the disease aetiology. T2DM is classified as a polygenic disorder because of the many different combinations of gene defects that exist among diabetic patients (Imperato and Imperato, 2009). While genes may impart susceptibility to T2DM, environmental factors serve as the trigger for the clinical disease. Genetic factors may also determine the rate of disease progression and the secretion and action of insulin (Imperato and Imperato, 2009). The development of T2DM can, therefore, be considered to be the result of interaction between the environment and strong hereditary components, although the exact mechanism is still poorly understood.

1.6.3 Risk factors for T2DM

The descriptive epidemiology and the pattern of inheritance of T2DM provide ample evidence that the disease originates from an interaction between genetic and lifestyle risk factors (Wareham *et al.*, 2002). Several risk factors for T2DM have been identified, including age, sex, obesity and central obesity, low physical activity, smoking, diet including low amount of fiber and high amount of saturated fat, ethnicity, family history, history of gestational DM, history of the non-diabetic elevation of fasting or 2-hour glucose, elevated blood pressure, dyslipidaemia, and different drug treatments (diuretics, unselected β-blockers, etc.) (Mykkanen *et al.*, 1993; Noble *et al.*, 2011). However, the main factors affecting the prevalence as well as the development and severity of T2DM are diet and obesity (American Diabetes Association, 2008). A decrease in physical activity and an increase in energy consumption will promote the development of obesity. Obesity itself causes some degree of insulin resistance, and patients who are not obese by the traditional weight criteria may have an increased fat deposition in the abdominal area. People who are overweight or obese have added pressure on their body's ability to use insulin to properly control blood sugar levels and are therefore more likely to develop DM (Prendergast, 2014).

Women who have been diagnosed with GDM and individuals with hypertension or dyslipidaemia are more prone to develop T2DM (Vijayaraghavan, 2010). The prevalence also differs between different racial or ethnic groups and is associated with a strong genetic predisposition. However, the genetic aetiology or genetic predictors for T2DM is complex and not clearly defined.

1.6.4 Heredity of T2DM

A number of studies have been conducted on the heritability of T2DM and it was reported that both genetic and environmental factors play a role in the development of T2DM (Almgren *et al.*, 2011; Poulsen *et al.*, 1999). The risk of developing T2DM increases when there is a positive family history of the disease. The heritability data of T2DM comes from population, family and twin-based studies and ranges from 20% to 80% (Meigs *et al.*, 2000). Individuals who have one parent affected by T2DM have a 40% risk of developing diabetes in their lifetime while individuals who have two parents suffering from T2DM have a 70% risk (Tillil and Kobberling, 1987). The concordance rate for the development of T2DM differs among monozygotic and dizygotic twin pairs. For dizygotic twins the concordance rate observed ranges between 17% to 20% (Kaprio *et al.*, 1992; Newman *et al.*, 1987) while for monozygotic twins it ranges between 35% to 58% and rises to 88% when impaired glucose tolerance is included (Henkin *et al.*, 2003).

Studies have determined that the risk of developing T2DM is greater when the mother is affected compared to the father. The cause for this parent-of-origin effect is unknown, but it could be due to biased parent-of-origin transmission of T2DM risk alleles (Kong *et al.*, 2009; Small *et al.*, 2011). This confirms the concept of a multifactorial aetiology of T2DM. It supports the contribution of genetic as well as non-genetic aetiological components in the development of T2DM.

1.6.5 Genetic component

Single nucleotide polymorphisms (SNPs) have been identified in over 60 genes that are associated with T2DM. Familial aggregation and the high concordance rate between monozygotic twins support a strong genetic contribution to its aetiology (Imperato and Imperato, 2009). The prevalence of T2DM also varies in different ethnic groups due to shared alleles and differing environments (King and Rewers, 1993). The majority of SNPs were identified through GWAS mostly in the European population (Scott *et al.*, 2007; Sladek *et al.*, 2007), but also in Asian populations (Hara *et al.*, 2014; Ma *et al.*, 2013; Unoki *et al.*, 2008) and more recently in the Australian Aboriginal population (Anderson *et al.*, 2015). According to the ADA, the rates of diagnosed T2DM differ greatly in non-Hispanic Caucasians (7.6%), Asian Americans (9.0%), Hispanics (12.8%) and non-Hispanic blacks (13.2%). While environmental risk factors to T2DM onset are well known, knowledge of the genetic basis is incomplete (Ali, 2013).

However, despite the vast flow of genetic information including the identification of many gene mutations and a large array of SNPs in many genes involved in the metabolic pathways, a major complication is the fact that a single gene mutation or polymorphism will not impose the same effect among different individuals within a population or different populations. This variation is directly or indirectly affected by the genetic background at the individual, family or population levels and can be complicated further by interaction with environmental factors which are also highly variable (Kharroubi and Darwish, 2015).

Overall, only a handful of studies have used the genome-wide approach to identify genomic regions linked to or associated with T2DM in African populations (Chen *et al.*, 2005; Chikowore *et al.*, 2015; Osei-Hyiaman *et al.*, 2001) and, thus, T2DM associated SNPs have not been comprehensively explored for genetic prediction in African populations. A few studies that have been investigating risk factors of T2DM in the black South African population have been focusing on the adiponectin (*ADIPOQ)* gene (Olckers *et al.*, 2007; Schwarz *et al.*, 2008). The study by Olckers *et al*. identified the C-11377G alteration in the *ADIPOQ* gene to have a protective effect against T2DM in

black South Africans. Schwarz *et al.* performed a meta-analysis study to compare the effects of the C-11377G locus within the adiponectin gene in a black South African, a Cuban Hispanic and a German Caucasian cohort. They found that there is no significant difference between the black South African control and diabetic cohorts and thus C-11377G is not a significant risk factor within the South African population. The homozygous genotype for the risk factor allele may only be associated with increased diabetes risk in the Cuban Hispanic cohort (Schwarz *et al.*, 2008). The majority of people with T2DM live in economically less-developed regions in the world and it is crucial to focus research on these populations to determine which genetic alterations make these people groups susceptible to developing T2DM (International Diabetes Federation, 2014).

Through linkage studies, candidate gene studies and GWAS, more than 60 genes have been found to be associated with T2DM. The most extensively studied include transcription factor 7 like 2 (*TCF7L2)*, insulin receptor substrate (*IRS)*, potassium channel subfamily J member 11 (*KCJ11*), cyclin-dependent kinase inhibitor 2A (CDKN2A), fat mass and obesity-associated gene (*FTO)*, tumor necrosis factor alpha (*TNFα)* and peroxisome proliferator-activated receptor gamma *(PPARG)* (Barroso, 2005; Clausen *et al.*, 1995; Day *et al.*, 1998; Grant *et al.*, 2006; Hu *et al.*, 2009; Majithia *et al.*, 2014).

1.7 The *PPARG* gene

Peroxisome proliferator-activated receptors (PPARs) are part of the ligand-inducible transcription factors, and one of the most comprehensively studied members of this family (Costa *et al.*, 2010). The *PPAR* genes are expressed in the reproductive organs, cardiac tissue and the major insulin target organs such as the skeletal muscle, liver, and white and brown adipose tissue (Braissant *et al.*, 1996). These genes have been associated with different biological pathways, varying from insulin sensitization, lipid and glucose homeostasis to tissue injury and wound repair, inflammation, immunity and cell differentiation and proliferation (Costa *et al.*, 2010). Three different PPAR isotypes are

known, *PPARα*, *PPARβ* and *PPARγ* (*PPARG*), which are encoded by separate genes on chromosome 22, 6 and 3 respectively and expressed in a tissue-specific manner (Sher *et al.*, 1993). The PPAR isoforms are very homologous and share transcriptional targets but the physiological functions of each PPAR are distinctive (Costa *et al.*, 2010).

PPARG (ENSG00000132170) is a member of the nuclear receptor superfamily and a major transcriptional regulator of adipocyte differentiation (Majithia *et al.*, 2014; Makishima, 2005). The transcriptional activity of *PPARG* is modulated through conformational changes in *PPARG* which is highly expressed in adipocytes, skeletal muscle, liver, and kidney, and has been shown to regulate expression of genes that mediate adipocyte differentiation, energy metabolism, and insulin action (Fajas *et al.*, 1997). *PPARG* is an established target for anti-diabetic thiazolidinediones (TZDs), used as a medication for people suffering from T2DM. TZDs enhance insulin sensitivity and improve glycaemic control (Chiarelli and Marzio, 2008).

1.7.1 Expression of *PPARG*

The *PPARG* gene is located on chromosome 3p25 (OMIM 601487) and spans more than 140 kb. *PPARG* exists as three major protein isoforms (*PPARG-1, PPARG-2* and *PPARG-3*) produced by alternate promoter usage and alternative splicing at the 5' end (Fajas *et al.*, 1998; Tontonoz *et al.*, 1994). The *PPARG-1* isoform is expressed in most tissues, while *PPARG-2* expression is restricted to adipose tissue where it is crucial for regulation of adipocyte differentiation (Israelian-Konaraki and Reaven, 2005; Yanase *et al.*, 1997), but it can also be induced in other tissues by a high-fat diet (Medina-Gomez *et al.*, 2007). Proteins produced from *PPARG-2* contain an additional NH2-terminal, containing 30 additional amino acids compared to *PPARG-1* and *PPARG-3* (Tontonoz *et al.*, 1994; Zhu *et al.*, 1993). *PPARG-3* mRNA is directed by an independent promoter and expression is confined to adipose tissue and colon epithelium (Fajas *et al.*, 1998). A specific polymorphism is present in this region, the Pro12Ala variant (rs1801282), which has been associated with resistance to the risk of T2DM (Yen *et al.*, 1997).

The *PPARG* contains nine exons shown in Figure 1.1. Exons 1-6 are shared between all three *PPARG* isoforms (*PPARG-1, PPARG-2* and *PPARG-3*). *PPARG-1* contains additional untranslated exons A1 and A2 (total of eight exons, six being translated). *PPARG-2* is encoded by an additional exon B which is translated and produces the extra 30 amino acids (7 translated exons). *PPARG-3* contains only the untranslated exon A2 (total of 7 exons, six being translated). *PPARG-1* and 3 thus give rise to the same protein which is encoded by exon 1-6 since exon A1 and A2 are not translated (Fajas *et al.*, 1998).

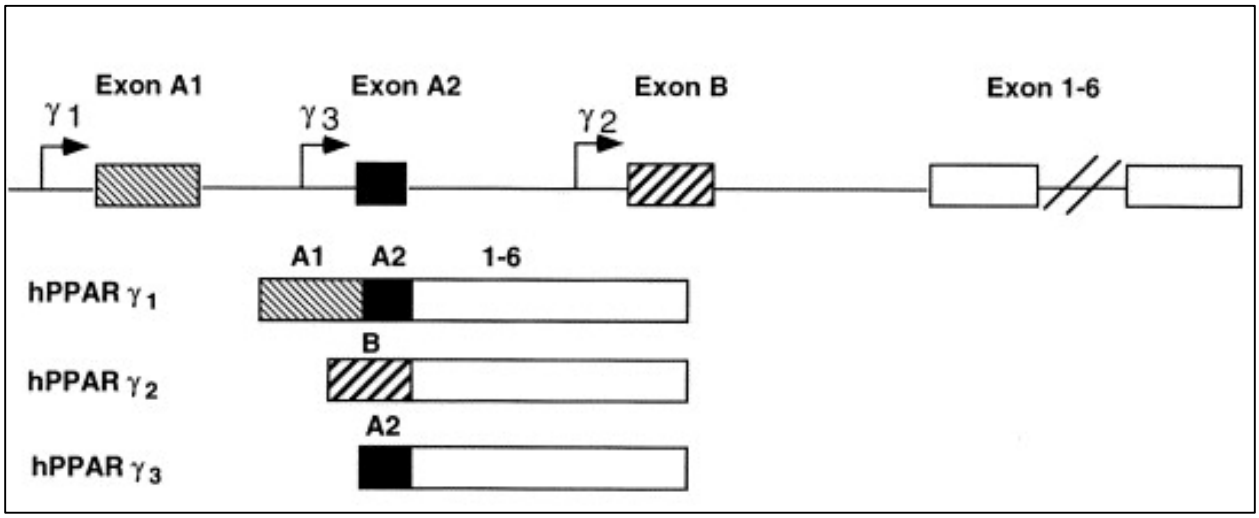


Figure 1.1: A schematic representation of the genomic structure of the 5' ends of the three *PPARG* isoforms (Copied from Fajas *et al*., 1998).

PPARG contains a number of distinct functional domains and transcription is regulated through the availability and binding potential of specific ligands. These functional domains include an activation function (AF) 1 region, a DNA-binding domain (DBD) with a C4-type zinc finger structure, a hinge region and a ligand-binding domain (LBD). TZDs are synthetic ligands which are insulin sensitizing drugs and used in the treatment of T2DM. The natural exogenous ligands are long-chain polyunsaturated fatty acids and the endogenous ligands include prostaglandins. *PPARG* DBD can regulate transcription of target genes by forming a heterodimer with retinoid X receptors (RXR). This complex then binds to a consensus element called the peroxisome proliferator response element

(PPRE), which consists of a two-hexanucleotide (AGGTCA or related sequence) direct repeat motif separated by a single nucleotide (5'-AGGTCA-N-AGGTCA-3') (Figure 1.2). Ligands bind to *PPARG* to induce a conformational change in the LBD which results in dissociation of the co-repressor complex and association of the co-activator complex, subsequently modulating the activity of *PPARG*. The *PPARG* gene activation improves insulin sensitivity and glucose, adiponectin, and fatty acid uptake (Savkur and Miller, 2006). Alterations in insulin signaling pathways play a central role in the pathogenesis of T2DM. This gene is of particular interest due to its pleiotropic functions that are crucial for the expression of genes involved in atherosclerosis, cancer, inflammation, glucose metabolism and adipogenesis (Capaccio *et al.*, 2010; Hummasti and Tontonoz, 2006).

PPARG activity is also controlled by amino acid modifications post-transcriptionally (Figure 1.2). SUMOylation is a post-translational modification involved in the regulation of protein function that plays an important role in a wide range of cellular processes. SUMOylation involves the covalent attachment of a member of the SUMO (small ubiquitin-like modifier) family of proteins to lysine residues in specific target proteins via an enzymatic cascade. SUMOylation of the AF1 region at Lys107 results in suppression of transactivation activity of *PPARG-2* while SUMOylation at Lys395 is required for transrepression of NF-kB (nuclear factor kappa-light-chain-enhancer of activated B cells) activity (Pascual *et al.*, 2005). Phosphorylation of the hinge region of *PPARG* by cyclin-dependent kinase 5 (*CDK5*) decreases the insulin-sensitizing capacity of *PPARG* without altering its adipogenic capacity. Phosphorylation by an MAP-kinase inhibits ligand-dependent *PPARG* activation (Camp and Tafuri, 1997). The browning of adipose tissue is associated with the deacetylation of *PPARG* by NAD-dependent deacetylase sirtuin 1 (Picard *et al.*, 2004). Thus, *PPARG* function is regulated by expression induction, ligand binding and protein modification.

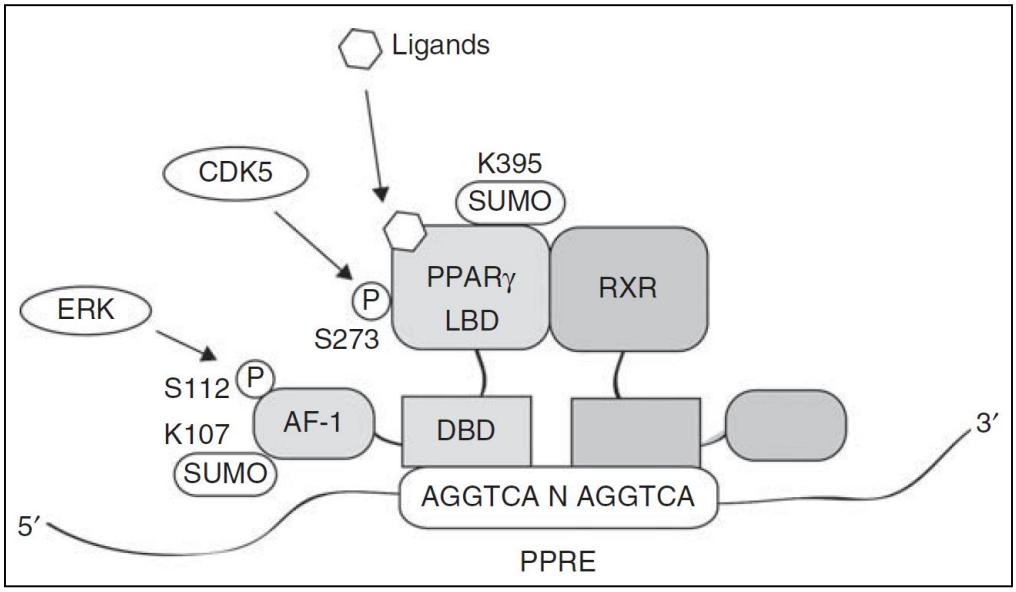


Figure 1.2: Structure and function of *PPARG*. Illustrated here is the heterodimer formation of *PPARG* and retinoid X receptors (RXR) when binding to the peroxisome proliferator response element (PPRE) and all its components. Also shown in this diagram are the sites of posttranslational modifications (Copied from Takada and Makishima, 2015).

1.7.2 *PPARG* Polymorphisms

A number of genetic variants have been identified in the *PPARG* gene. Two loss-of-function mutations (Val290Met and Pro467Leu) have been reported in three individuals with severe insulin resistance but normal body weight (Barroso *et al.*, 1999). A very rare gain-of-function mutation (Pro115Gln) has been identified and was associated with obesity but not with insulin resistance (Ristow *et al.*, 1998). A silent CAC478CAT mutation (Valve *et al.*, 1999) and the highly prevalent Pro12Ala polymorphism in *PPARG-2* (Figure 1.3) have also been described. The Pro12Ala polymorphism is one of the most documented gene variants and consistently associated with a reduced risk for T2DM (Gouda *et al.*, 2010).

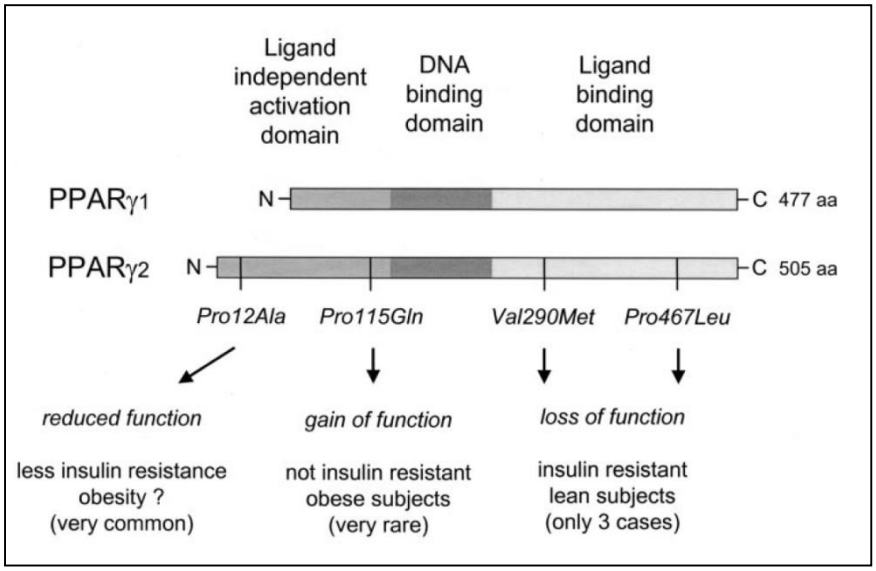


Figure 1.3: Organization of the *PPARG* gene. The two different isoforms are shown with the most well-known mutations (Copied from Stumvoll and Haring, 2002).

Yen *et al*. (1997) first identified the missense mutation caused by a cytosine (C) to guanine (G) mutation (CCA-to-GCA). This results in the amino acid substitution of alanine (Ala) for proline (Pro) at codon 12 in exon B of *PPARG-2* illustrated in Figure 1.3. The rare allele frequencies were observed to be 12% in Caucasians, 10% in Native Americans, 8% in Samoans, 4% in Japanese, 3% in African-Americans, 2% in Nauruans, and 1% in Chinese (Yen *et al.*, 1997). From this study, it was also concluded that the Pro12Ala polymorphism is present in diverse populations. The effect of this mutation on the individual is weak, but because of a prevalence of more than 75% of the high-risk Pro allele, the population-attributable risk is enormous (Stumvoll and Haring, 2002).

The Pro12Ala polymorphism is located in the ligand-independent activation domain and has been found to modulate the transcriptional activity of *PPARG* (Nemoto *et al.*, 2002). It leads to a diminished stimulation of *PPARG* target genes, subsequently lowered levels of adipose tissue accumulation, which, in turn, may improve insulin sensitivity (Miramontes González *et al.*, 2014). GWAS (Saxena *et al.*, 2007; Scott *et al.*, 2007; Zeggini *et al.*, 2007) and meta-analysis studies (Altshuler *et al.*, 2000; Gouda *et al.*,

2010) have provided decisive evidence that that the *PPARG* Pro12Ala polymorphism is positively associated with a reduction of 21% to 27% for the alanine allele in the risk of T2DM. The alanine genotype reportedly results in higher insulin sensitivity (Deeb *et al.*, 1998; Ek *et al.*, 2001; Muller *et al.*, 2003) by comparison with the proline genotype which has lower transcriptional activity.

Not all studies report evidence for an interaction between the Pro12Ala polymorphism and the nature of dietary fat intake on body mass index (BMI), insulin resistance or T2DM (Poirier *et al.*, 2000). It has been considered that physical activity and dietary fatty acids act synergistically with Pro12Ala to modulate insulin levels (Franks *et al.*, 2004). Intervention studies have indicated a greater improvement in glucose tolerance and insulin levels in individuals carrying the Ala genotype following a structured exercise and training program for 3 to 6 months compared to individuals homozygous for the Pro genotype (Adamo *et al.*, 2005; Kahara *et al.*, 2003; Weiss *et al.*, 2005). This data was supported by observations from the Finnish Diabetes Prevention Study where Ala allele carriers were less likely to develop T2DM than Pro allele homozygotes during a randomized controlled trial of lifestyle intervention in people with increased risk for T2DM (Lindi *et al.*, 2002).

Only a single study investigated the prevalence of the Pro12Ala SNP in a South African population. Vergotine *et al.* (2014) performed a cross-sectional study where 820 participants of mixed ancestry from Cape Town were genotyped for the Pro12Ala SNP. Of this study group, 222 had T2DM and 598 were non-diabetic. Their study confirmed an almost zero occurrence of the Pro12Ala SNP in the African population, but they also highlight the importance of further studies investigating this SNP in African populations.

Table 1.2 provides a summary of the data from multiple studies on the association of the Pro12Ala polymorphism and T2DM. The table indicates the study population, the population size and whether an association was made between the Pro12Ala polymorphism and T2DM.

Table 1.2: Summary of data from studies conducted to investigate the association of peroxisome proliferator-activated receptor gamma *(PPARG)* Pro12Ala polymorphism with type 2 diabetes mellitus (T2DM) and obesity.

Population	Population size	*PPARG* associated with T2DM	Study reference
Malaysian	241	Yes	(Paramasivam *et al.*, 2016)
Qatari	764	No	(Bener *et al.*, 2015)
Guwahati (Eastern) India	50	No	(Majumdar *et al.*, 2014)
South Africa (mixed ancestry)	820	No	(Vergotine *et al.*, 2014a)
Spain (Caucasians)	298	No	(Miramontes González *et al.*, 2014)
West Bengal (Eastern Indian)	400	No	(Pattanayak *et al.*, 2014)
China	453	No	(Ye *et al.*, 2014)
Chinese	3,146	Yes	(Wang *et al.*, 2013)
Italian	1,273	Yes	(Trombetta *et al.*, 2013)
Iranian	200	Yes	(Motavallian *et al.*, 2013)
Western India	50	Yes	(Shahrjerdi *et al.*, 2013)
Chinese Han	7,203	No	(Tong *et al.*, 2012)
North India	175	No	(Raza *et al.*, 2012)
France	4,676	Yes	(Lamri *et al.*, 2012)

Chinese Han	7,291	No	(Guo *et al.*, 2011)
South Indian	2,000	No	(Vimaleswaran *et al.*, 2010)
Tunis	675	No	(Ben Ali *et al.*, 2009)
Sweden	4,787	No	(Montagnana *et al.*, 2008)
Asian Sikh (Northern India)	918	Yes	(Sanghera *et al.*, 2008)
Qatari	850	No	(Badii *et al.*, 2008)
Spain	538	Yes	(Soriguer *et al.*, 2006)
Caucasians	32,000	No	(Tonjes *et al.*, 2006)
Brazil	377	Yes	(Tavares *et al.*, 2005)
Tunis	488	No	(Zouari Bouassida *et al.*, 2005)
France (Caucasians)	3,250	Yes	(Ghoussaini *et al.*, 2005)
Asia	3,938	No	(Tai *et al.*, 2004)
Polish	644	No	(Malecki *et al.*, 2003)
Canada (Québec)	720	Yes	(Robitaille *et al.*, 2003)

A recent study by Majithia *et al*. (2014) hypothesized that individuals in the general population might harbour rare, non-synonymous variants in *PPARG* and that some of these variants would alter function in adipocyte differentiation and, thus, be associated with T2DM. Using next generation sequencing the *PPARG* gene was sequenced in 19752 participants consisting of a T2DM and control cohort, from multiple studies and from different ethnic groups. They identified 49 novel non-synonymous *PPARG* variants

with nine of these causing reduced activities in adipocyte differentiation. Individuals carrying any of these loss of function (LOF) variants were associated with a significantly increased risk of developing T2DM (Majithia *et al.*, 2014).

In conclusion, *PPARG* is one of the major genes that have been identified to have a broad impact on the risk of common T2DM. The precise understanding of its mechanism may lead to novel diagnostic, preventive, and therapeutic approaches for improving the management of T2DM (Stumvoll and Haring, 2002).

1.8 Quantitative real-time polymerase chain reaction

Real-time polymerase chain reaction is a modification of conventional PCR that is rapidly changing the nature of how biomedical research is conducted . Real-time PCR was first introduced by Higuchi and co-workers in 1972 and has since rapidly increased in use (Higuchi *et al.*, 1992; Higuchi *et al.*, 1993). It allows for precise quantification of specific nucleic acids in a complex mixture even if the amount of starting material is at a very low concentration. This is accomplished by using fluorescent technology to monitor the amplification of a target sequence in real time. The amount of starting material present correlates with how quickly the amplification target reaches a threshold detection level.

Over the past decade, real-time PCR applications have become broadly used tools for the quantification of specific sequences in complex mixtures. For example, quantitative real-time PCR (qPCR) has been used for genotyping (Alker *et al.*, 2004; Cheng *et al.*, 2004; Gibson, 2006), quantifying viral load in patients (Ward *et al.*, 2004), assessing gene copy number in cancer tissue (Bieche *et al.*, 1998; Kindich *et al.*, 2005; Konigshoff *et al.*, 2003) and most commonly for studying gene expression levels by coupling it with a procedure called reverse transcription PCR.

Quantitative PCR technology is based on conventional PCR but incorporates different detection chemistries to allow amplification and detection of DNA in a single reaction

(Wong and Medrano, 2005). The simplicity, sensitivity and specificity together with its potential for high throughput has made qPCR the benchmark technology for the detection of DNA (Bustin, 2005).

For detection of the PCR product in real time, the use of a fluorescent dye is necessary. These dyes can either be nonspecific, such as fluorescent DNA-binding dyes (e.g., SYBR Green I) or sequence-specific probes (e.g., TaqMan® or molecular beacons). Most of these are based on fluorescent resonance energy transfer (FRET) to distinguish between different products.

1.8.1 Sequence-specific fluorescent probes

Sequence-specific fluorescent probe assays are ideal for applications where nonspecific amplification occurs or more than one target sequence is monitored in a single PCR reaction. Fluorophore-coupled nucleic acid probes are commonly used as a detection chemistry. Strand-specific probes will interact with the PCR products in a sequence-specific manner to provide information about a specific PCR product as it accumulates. A widely used strand-specific approach involves hydrolysis probes based on the 5' nuclease activity of *Taq* polymerase. The TaqMan® probes are a well-known example of a hydrolysis probe and have been used extensively in a wide range of studies (Heid *et al.*, 1996; Holland *et al.*, 1991).

The TaqMan® assay is a detection chemistry that makes use of probes which fluoresce upon probe hydrolysis to detect PCR product accumulation (Figure 1.4). These sequence-specific oligonucleotide probes are labelled with a reporter dye at the 5' end and a quencher at the 3' end (Gibson *et al.*, 1996). While the probe is intact, and the quencher is in close proximity to the reporter, it will reduce the reporter fluorescence intensity by FRET (Wong and Medrano, 2005). The quencher is selected based on its ability to specifically absorb the emitted spectra of the reporter and should be spaced in the probe to optimize the capture of that light. The probes are effectively used for allelic discrimination purposes. In a multiplex reaction, one probe is specific to the DNA sequence and usually, a mixture of several mutations can be detected at once. During

PCR, the *Taq* polymerase extends the primers and synthesizes the complementary strand which causes the 5'-exonuclease activity of the *Taq* polymerase to degrade the annealed probe. Degradation of the probe allows for the reporter molecule to fluoresce, as the reporter and quencher molecule are now separated (Heid *et al.*, 1996). The increase in reporter fluorescence is captured by the sequence detection instrument and displayed by the software. The amount of reporter fluorescence increase is proportional to the amount of product being produced for a given sample.

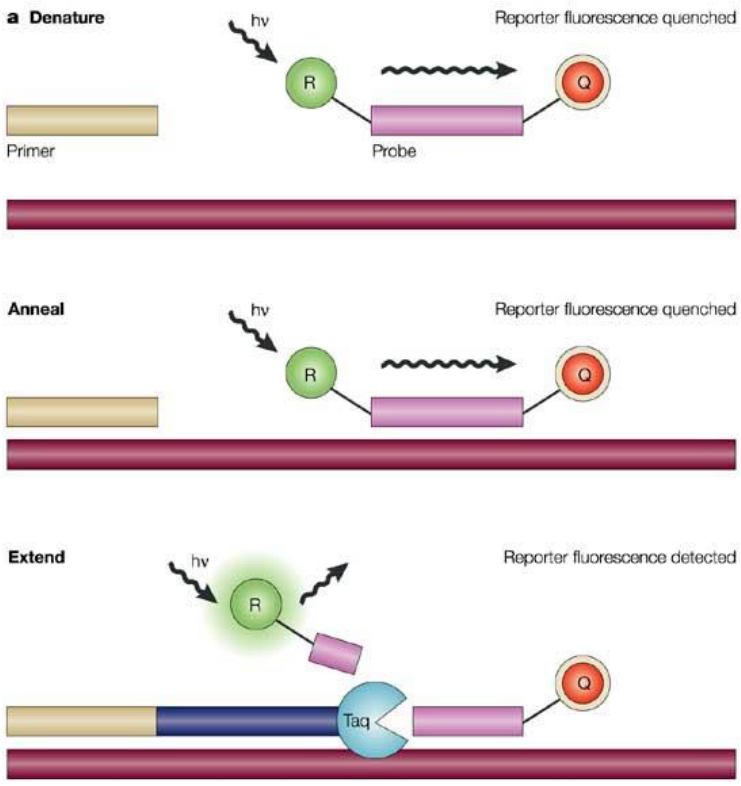


Figure 1.4: Schematic representation of the principle of the TaqMan® hydrolysis probe assay. The thermostable polymerases uses 5'-nuclease activity to cleave the hydrolysis probes during the amplicon extension step. This process separates the detectable reporter fluorophore (R) from the quencher (Q). Fluorescence is emitted when it is excited by an external light source (hv). The fluorescence emitted at each PCR cycle is proportional to the amount of product formed (Copied from Koch, 2004).

An advantage of using strand-specific probe assays is that multiple probes can be combined in a multiplex reaction that allows for information to be obtained about several target sequences from a single reaction. This is advantageous because both control and target sequences are amplified under identical conditions. A second advantage is that the use of strand-specific probes can help distinguish between products when there are two potential PCR products being produced by the same primer set. The hydrolysis probe system can be used for genotyping and to identify specific point mutations, SNPs and allelic variants (Cheng *et al.*, 2004; Eshel *et al.*, 2006; Livak, 2003; Marras *et al.*, 2003).

1.8.2 Advantages of using qPCR

Quantitative PCR collects data in the exponential growth phase while traditional PCR only measures at the end-point or plateau. The increase in reporter fluorescent signal is directly proportional to the number of amplicons generated. The dynamic range of detection is increased (qPCR method has a very large dynamic range of starting target molecule determination) and it is extremely accurate. Quantitative PCR requires 1000-fold less DNA than conventional assays. No post-PCR processing (such as electrophoresis) is necessary due to the closed system resulting in much faster and higher throughput assays and reduced risk of contamination. And it requires no data file conversion of bioinformatic pipeline analysis. This makes it the method of choice for genotyping, routine testing as well as on research cohorts of small to medium size.

1.9 Next Generation Sequencing

1.9.1 Introduction

The discovery by Sanger *et al.* (1977) of using dideoxy nucleotides for chain termination marked a milestone in the history of DNA sequencing. This concept laid the foundation for the development of automated Sanger sequencing, which has been the method of choice for DNA sequencing for almost 20 years (Ansorge *et al.*, 1987; Smith *et al.*, 1986). The use of nucleic acid sequencing has increased exponentially as the ability to

sequence has become more accessible to clinical and research laboratories across the world. This technique has been adapted to allow for longer DNA fragments and a higher level of parallelism, however, Sanger-based approaches have not been able to analyse DNA in a high-throughput manner. The undertaking of the Human Genome Project was the first major insight into DNA sequencing, with a 3 billion USD budget spanning over 13 years and completed in 2003 with the goal of determining all three billion base pairs making up the human genome (Collins *et al.*, 2003; Lander *et al.*, 2001). This project was completed using automated Sanger sequencing and there has been an increasing demand for cheaper and faster sequencing since the completion of the Human Genome Project. This has been the driving force behind the development of next-generation sequencing (NGS).

In recent years, the introduction of NGS technology has revolutionized how genomic studies are processed. NGS is an alternative to traditional Sanger sequencing that provides a much cheaper option for higher-throughput sequencing. In the process of NGS, millions of fragments of DNA from a single sample are sequenced simultaneously and thus allows for massively parallel sequencing. These sequencing technologies facilitate high-throughput sequencing that enables scientists to sequence an entire genome in a single day.

Due to the millions of short sequence reads generated by the NGS techniques, the bottleneck in sequencing has now shifted from sequence generation to data management and analysis. Data volume creates major challenges for storage, backup and analysis and this has highlighted the need for new algorithmic approaches to overcome the limitation of short read lengths. The development of streamlined, highly automated pipelines for data analysis is critical for the transition from technology adoption to accelerate research and consequent publications (Camerlengo *et al.*, 2012).

NGS facilitates the discovery of genes and regulatory elements that can be linked and associated with diseases. Discovery of disease-causing mutations in specific genes, using targeted sequencing can lead to disease diagnosis and/or early intervention.

RNA-sequencing can shed some light on gene expression profiles and information on the entire transcriptome of a sample in a single analysis (Mortazavi *et al.*, 2008; Wang *et al.*, 2008). This technique represents a useful alternative to microarrays for gene expression studies. As a result, NGS has rapidly become the technology of choice for most scientists conducting sequence projects.

1.9.2 Procedure

Several NGS platforms based on different chemistries have been developed in the past decade to offer low-cost and high-throughput sequencing. The following stepwise procedures are followed by all NGS platforms.

1. Template preparation (fragmentation)

Template preparation is the building of a DNA or cDNA library and amplification of that library. These sequencing libraries are constructed by fragmenting the DNA or cDNA template and then adding adapter sequences. Fragmentation is necessary due to the restriction on the length of sequence that can be produced, which at this stage ranges from 50 bp to ~1 kbp. Adapter sequences are synthetic oligonucleotides of a known sequence that are added onto the ends of each DNA fragment. Once the sequencing library is constructed, it is clonally amplified in preparation for sequencing.

2. Sequencing and imaging

After the amplified libraries are prepared, the nucleic acid sequence is usually obtained through sequencing by synthesis. The new DNA fragment is synthesized from the DNA library fragments that act as templates. The fragment is synthesized through repeated cycles of flooding the fragment with a specific nucleotide and then washing, in a sequential order. As the nucleotides are incorporated into the growing DNA strand the sequence is recorded as digital information.

3. Data Analysis

After sequencing, the raw data undergoes many different analysis steps. The amount of raw data produced by NGS technologies is remarkable, which necessitates many analytical and computational steps to convert this output into high quality usable information. There are analysis pipelines for NGS data that include the pre-processing of data to remove the adapter sequences and low-quality reads, mapping of processed data to a reference genome or *de novo* alignment of the sequencing reads, and analysis of the compiled sequence. Analysis of the sequences includes a wide variety of bioinformatics assessments, including genetic variant calling for detection of SNPs or indels, detection of novel genes or regulatory elements, and assessment of transcript expression levels. Analysis can also include identification of both somatic and germline mutation events that may contribute to the diagnosis of a disease or genetic condition. Many free online tools and software packages exist to perform the bioinformatics necessary to successfully analyse sequence data (Gogol-Doring and Chen, 2012). Despite this, data analysis is considered to be the major limiting step for using NGS.

1.9.3 Targeted sequencing

Targeted sequencing of specific genes or genomic regions is preferred to whole-genome or whole-exome sequencing where a particular disease or condition has been identified (Albert *et al.*, 2007; Hodges *et al.*, 2007). Targeted sequencing is much more affordable, allows for a higher coverage of regions of interest and also reduces sequencing time and cost (Xuan *et al.*, 2013). Sequencing panels have now been developed that target hundreds of genomic regions that are hotspots for specific disease-causing mutations. Such an example is a cancer hotspot panel that has been validated for use in clinical laboratories (Simen *et al.*, 2015; Singh *et al.*, 2013). This reduces the amount of data being produced and time to process all the data, by eliminating large regions of the genome that do not have an effect on the disease of interest. Targeted sequencing leads to rapid diagnosis and can aid in therapeutic decision-making in many genetic diseases.

1.9.4 Limitations of NGS

Compared to Sanger sequencing, NGS is much less costly in terms of time and money for the amount of sequence data generated; however it is still too expensive for many laboratories. The NGS platforms have a very high start-up cost and the individual sequencing reactions have an additional cost per genome. Short-sequencing read lengths and inaccurate sequencing of homopolymer regions on certain NGS platforms can lead to sequencing errors (Quail *et al.*, 2012). Due to large amounts of data produced by NGS platforms, the analysis is time-consuming and requires extensive knowledge of bioinformatics to gather accurate information from sequencing data, which is not readily available at all research institutions formerly using Sanger sequencing.

1.9.5 NGS Platforms

Sequencing technologies are evolving rapidly and several different benchtop high-throughput sequencing instruments are currently available.

1.9.5.1 Illumina

The Illumina sequencing system employs an array-based DNA sequencing-by-synthesis technology with reversible terminator chemistry (Bentley *et al.*, 2008). This approach involves the hybridization of template DNA fragments to a reaction chamber on an optically transparent solid surface (flow cell). DNA is synthesized by using reversible terminators (Turcatti *et al.*, 2008) which consist of four modified nucleotides labelled with different removable fluorescent dyes at the 3'-hydroxyl terminus, which are used for step-by-step synthesis. The flow cell contains eight independent lanes with millions of clonal clusters being generated in each lane that allows for multiple libraries to be sequenced in parallel. The GA IIx and HiSeq 2000 platforms yield increased read lengths and a much higher sequence output. The MiSeq workflow has the fewest manual steps, generates the highest throughput per run, has the lowest error rate and requires very little input-DNA (Loman *et al.*, 2012; Quail *et al.*, 2012). Regardless of the high-throughput of these platforms, the utility of Illumina systems is limited to short-read lengths which are due to the dephasing effect (Metzker, 2010). Failures in removing or

adding terminating moieties and increased and decreased efficacy of nucleotide incorporation in any cycle can cause overextension or incomplete extension of the growing strand along the template. This results in a leading strand or lagging strand dephasing. A decay in fluorescent signal, incomplete removal of fluorescent labels and incorporation of nucleotides without a fluorescent label can cause signal dephasing leading to base-call errors. As a result, with an increase in read length, the base substitution error rate will increase. Uneven read coverage has also been revealed across regions that are AT- and CT-rich with a bias towards the latter (Loman *et al.*, 2012). The MiSeq also has the longest run time (Loman *et al.*, 2012).

1.9.5.2 Roche 454

The 454 sequencing system is based on the combination of emulsion PCR and pyrosequencing technology (Margulies *et al.*, 2005). This process involves beads, carrying single-stranded template, being confined to individual emulsion droplets in which PCR amplification occurs to produce millions of copies of each template. These amplicon-bearing beads are then deposited into individual wells of a picotiter plate and solid-phase pyrosequencing is carried out (Margulies *et al.*, 2005). This method involves sequencing-by-synthesis and depends on the monitoring of luminescence emission in real time each time a pyrophosphate is released upon incorporation of nucleotides into the growing template strand (Margulies *et al.*, 2005). The long reads being sequenced by the 454 system is the main advantage of this technology (Loman *et al.*, 2012; Margulies *et al.*, 2005). The 454 platform is more expensive compared to other NGS platforms and is best suited for applications such as de novo assembly (Gilles *et al.*, 2011) and metagenomics (Loman *et al.*, 2012). However, the 454 technology has an inherent problem in the detection of homopolymers, which are stretches of the same nucleotide. This can lead to non-linearity between the signal intensity and the length of homopolymer stretches when more than three or four nucleotides are consecutively incorporated. Thus, the 454 platform has a relatively high error rate for calling insertions and deletions in homopolymers (Wommack *et al.*, 2008).

1.9.5.3 ABI/SOLiD (Life Technologies)

Massively parallel sequencing by hybridization–ligation, implemented in the supported oligonucleotide ligation and detection system (SOLiD) from Applied Biosystems, has been commercially available since 2006 (Shendure *et al.*, 2005). The ligation chemistry used in SOLiD is based on the polony sequencing technique that was published in the same year as the 454 method. Construction of sequencing libraries for analysis on the SOLiD instrument begins with an emulsion PCR single-molecule amplification step. The amplification products are transferred onto a glass surface where sequencing occurs by sequential rounds of hybridization and ligation with 16 dinucleotide combinations labelled with four different fluorescent dyes (each dye used to label four dinucleotides). Using the four dye encoding scheme, each position is effectively probed twice, and the identity of the nucleotide is determined by analysing the colour that results from two successive ligation reactions (Shendure *et al.*, 2005). Significantly, the two-base encoding scheme enables the distinction between a sequencing error and a sequence polymorphism: an error would be detected in only one particular ligation reaction, whereas a polymorphism would be detected in both. In comparison to the other NGS platforms, the SOLiD system presents the lowest error rate of which the most common error type is substitutions. It has also been shown in the SOLiD data that there is an underrepresentation of AT-rich regions (Harismendy *et al.*, 2009).

1.9.5.4 Ion Torrent Personal Genome Machine (PGM™)

Ion Torrent PGM™ is based on semiconductor technology detecting the protons released, as nucleotides are incorporated during synthesis (Rothberg *et al.*, 2011). First a library is constructed by fragmenting the DNA and ligating adapter sequences. After library preparation, the DNA fragments with specific adapter sequences are clonally amplified by emulsion PCR on the surface of 3-micron diameter beads, known as Ion Sphere Particles (Rothberg *et al.*, 2011). The templated beads are loaded into protonsensing wells that are fabricated on a silicon wafer and sequencing is primed from a specific location in the adapter sequence. Each of the four nucleotide bases is introduced consecutively as sequencing proceeds. When the nucleotide in the flow is

complementary to the template base directly downstream of the sequencing primer, the nucleotide is incorporated by the polymerase. As the different bases are incorporated, protons are released due to the hydrolysis of the nucleotide triphosphate, which causes the net liberation of a single proton for each nucleotide incorporated during that flow. A shift in the pH of the surrounding solution is produced by the release of the proton. This shift is directly proportional to the number of bases incorporated (Rothberg *et al.*, 2011). The change is detected by a sensor on the bottom of each well and is converted to a voltage which is digitized. After a nucleotide has flowed over the chip, the wells are washed to ensure that no nucleotides remain in the wells.

A signal processing software is used to change raw voltages into base calls. Raw data is converted into measurements of incorporation in each well for each nucleotide using a physical model. The physical model considers the polymerase rates, buffering effects and diffusion rates and is applied and fitted to the raw trace from each well and the incorporation signals are extracted. A base caller corrects the signals for phase and signal loss, normalizes to the key and generates correct base calls for each flow in each well to produce the sequencing reads (Rothberg *et al.*, 2011).

Each read is then sequentially passed through two signal-based filters to exclude low-accuracy reads. The first filter measures the fraction of flow in which an incorporation event was measured and the second filter measures the extent to which the observed values match those predicted by the phasing model. Lastly, an adaptation of the Phred method (Appendix B) is used to predict per-base quality values, which quantifies the concordance between the phasing model predictions and the observed signal (Rothberg *et al.*, 2011).

The biggest advantage of the Ion Torrent platform is that it delivers the fastest throughput and shortest run time. It is also the lowest-priced instrument and is notable for offering three different priced sequencing-chip reagents, which gives flexibility when designing experiments. This technology also requires a very low DNA-input for sequencing. However, negative comments on PGM™ include that the Ion Torrent

produces the shortest reads when compared to MiSeq (Illumina) and 454 GS Junior (Roche) platforms and has a higher error rate in homopolymer regions (Quail *et al.*, 2012). Ion torrent sequencing is not recommended for sequencing of regions that are extremely AT-rich, due to coverage bias as seen in some studies (Quail *et al.*, 2012).

Although the data output for Ion Torrent is still relatively low per chip, the fast turnaround time per chip makes this a very suitable technology for smaller, focused sequencing projects, 16S sequencing projects and SNP detection and validation as well as sequencing of small genomes. For each technology, there is a trade-off between advantages and disadvantages. The decision on which instrument to use will depend on many factors including available resources, available finances and the type of application being considered.

1.10 Conclusion

T2DM is a global epidemic that results in millions of deaths each year (Amod *et al.*, 2012). The number of people living with this disease worldwide is estimated to increase by 55% by 2035 and in Africa by 109%. There are approximately 2.7 million people in South Africa currently diagnosed with DM. The increase in obesity and T2DM is a result of lifestyle and dietary changes over the past few decades. The exact reasons for the prevalence of this disease in certain ethnic populations are still unidentified. Evidence for the genetic predisposition to T2DM has been observed in association studies in Caucasians. SNPs have been identified in over 50 genes that are associated with T2DM, but these risk-associated gene polymorphisms have not been comprehensively explored for genetic prediction in black South African populations.

We investigated the genetic risk factors of T2DM in the black, female South African population by performing NGS to identify the most common SNPs in the *PPARG* gene. Subsequently, we hypothesized that these results would either confirm mutations associated with T2DM in other global populations or identify novel mutations associated with T2DM in this population. The aim of this study was to screen the *PPARG* gene for

novel T2DM genetic risk factors and to determine the presence of a previously identified T2DM genetic risk factor, the Pro12Ala variant (rs1801282) in black South African women with T2DM.

The identification and description of a large number of novel genetic variants increasing susceptibility to T2DM will open up opportunities to translate this genetic information to the clinical practice and improve risk prediction. Genetic prediction models can be improved by increasing the precision of diagnosis of T2DM, by identifying low-frequency and rare genetic variants and by identifying risk variants in non-European ancestry since it has been shown that the greatest genetic variation is found in the recent African ancestry population (Hindorff *et al.*, 2009). NGS has a considerably better potential to find structural variation than conventional sequencing (Sanger sequencing) and will assist in and contribute to the understanding of the genetics of T2DM. Screening for T2DM can lead to earlier identification and treatment of asymptomatic diabetes, IFG or IGT and result in improved outcomes.

CHAPTER TWO

METHODOLOGY

2.1 Introduction

This chapter describes the study design, population group, sample selection and methodology of the study. The study procedure is outlined and the methods used is discussed. An overview is given of the data analysis pipelines used for Next Generation Sequencing (NGS) and detailed descriptions of the commands can be viewed in Appendix D. Lastly, the validity and reliability and the ethical considerations for this study are described.

2.2 Study design

For the purpose of this degree, an analytical case-control study was performed at the Department of Haematology and Cell Biology, Faculty of Health Sciences of the University of the Free State in Mangaung, South Africa.

2.3 Sample

The study population consisted of black female participants living in and around Mangaung, Free State. The languages spoken are Sesotho, Afrikaans, English, isiXhosa and Setswana. The Mangaung area covers more than 6,284 km² with a population of 747,431 people (Municipality: Mangaung Metropolitan Municipality, 2014).

A sample of convenience was taken from Type 2 diabetes (T2DM) patients that were recruited from the diabetes clinics of the Universitas Academic complex and Pelonomi Regional Hospital in Mangaung, which are both referral hospitals representing the Bloemfontein, Botshabelo and Thaba ‘Nchu black population. The patient samples were previously collected under the ethics number ECUFS 162/2012, according to the inclusion criteria below in section 2.3.1. Control subjects were also recruited as a convenience sample, but only after patient sampling was completed to determine the matching criteria ranges for age and body mass index.

All control samples were collected by the principal investigator. The control participants were recruited from the Universitas Hospital Staff as well as from outreaches arranged at various shopping centres in Mangaung. Participants who met the inclusion criteria (section 2.3.2) were asked to participate in the study. Informed consent was obtained from all participants in the language of their choice (refer Appendix A for English version). Thereafter, anthropometric measurements were taken and 10 ml blood drawn from all participants by a qualified professional nurse for a glycosylated haemoglobin (HbA1c) test and genetic testing. From this sample, participants with complete data sets for age, gender, weight and height as well as blood samples for genotype determination and HbA1C were included in the study. The study participants received feedback on their weight and height measurements, BMI calculations and on the glucose test results.

The calculation of the sample size for a genetic study requires the consideration of additional factors such as: the inheritance model (dominant, recessive or additive), the frequency of the alleles in the population, the overall disease risk in the population, in addition to the choice of an odds ratio (or relative risk), the power and the significance level (Cornell University, 2007). Another study by Hong and Park (2012) showed that 143 samples were sufficient to reach an 80% confidence level using a single SNP under a specific odds ratio of homozygotes to heterozygous alleles (Hong and Park, 2012). However, since no information exist on the ratio of homozygous to heterozygotes or the disease risk status of the allele, no specific calculation could be made. Thus the sample size was decided based on available funds. The study consisted of a total of 184 samples of which 93 were participants diagnosed with T2DM and 91 were control subjects.

2.3.1 Inclusion criteria for T2DM patients

The patient cohort included:

- participants that gave informed consent;
- black females;
- patients that were diagnosed with T2DM by a medical doctor;
- patients with a HbA1c above 6.5% upon first diagnosis;
- participants that were between 35 and 62 years of age.

2.3.2 Inclusion criteria for controls

The control cohort included:

- participants that gave informed consent;
- black females;
- participants that claimed they did not suffer from or had a previous history of T2DM;
- participants that were between 35 and 62 years of age;
- participants with a HbA1c below 6.5%;
- participants with a BMI above 25 (according to the T2DM patient cohort).

2.3.3 Exclusion criteria for all participants

Participants were excluded from the study if:

- participants were male;
- participants were diagnosed with Type 1 Diabetes Mellitus;
- participants were pregnant;
- participants were unable to give consent;
- participants were younger than 35 years or older than 62 years of age.

Both patient (n=93) and control (n=91) cohorts' samples were used to screen for the Pro12Ala mutation in the *PPARG* gene using qPCR. The cohorts used for screening were not individually matched but all met the requirements set by the inclusion criteria. Sixteen samples were randomly selected for NGS, eight patient samples and eight control samples. The control samples were individually matched to the patient samples according to race (self-declared black South African), gender (all females), age (± 2 years) and BMI (according to WHO categories, Table 2.1).

Table 2.1: BMI classification according to the World Health Organization (WHO, 2015).

BMI	Classification
< 18.5	Underweight
18.5 – 24.9	Normal
25 – 30	Overweight
> 30	Obese

2.4 Study procedure

A formal request was made to the Ethics Committee of the Faculty of Health Sciences at the University of the Free State (ECUFS) for ethical approval. Patient DNA was obtained from blood collected under Ethics number ECUFS 162/2012 with informed consent to use and store genetic material for future research on T2DM. For this study (ECUFS 53/2015), permission was obtained from each participant to store their blood and genetic material for future genetic studies on T2DM. Control subjects were recruited from Universitas Hospital Staff and from various shopping centres in Mangaung. Measurements were taken, informed consent was obtained and blood was drawn. HbA1C levels were tested at accredited private laboratories. DNA was isolated from control and patient blood samples.

A qPCR assay was designed and optimized for screening the Pro12Ala polymorphism in both cohorts. Quantitative PCR was performed at the University of the Free State, in the Department of Hematology and Cell Biology. Figure 2.1 is a schematic representation of the study procedure.

Ion Torrent AmpliSeq™ online facility was used to design primers that will capture the region of interest (*PPARG*). A cohort of 16 samples was sent for NGS at the DNA Sequencing Facility of the Faculty of Natural and Agricultural Sciences at the University of Pretoria. Thereafter, the primary sequence data analysis was performed according to the appropriate pipelines.

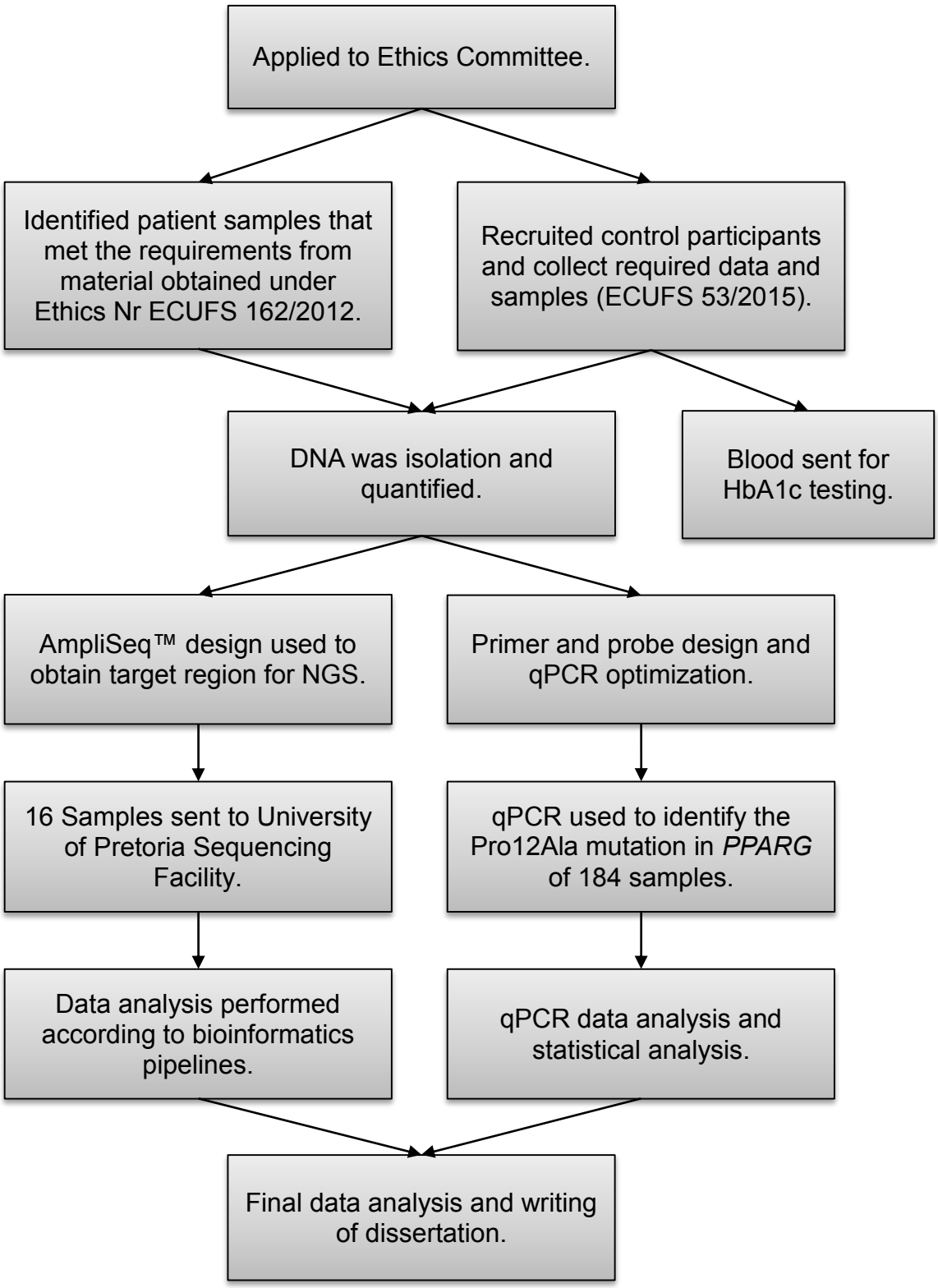


Figure 2.1: Flow chart illustrating the different steps of the study procedures.

2.5 Methods

2.5.1 Anthropometric measurements

Body Mass Index (BMI): A set of non-invasive quantitative techniques were used to calculate the BMI. The BMI is a measure of relative size based on the mass and height of an individual and provides a reliable indicator of body fatness (Shah and Braverman, 2012; Volgyi *et al.*, 2008). The actual body weight was measurement at the time of examination. The weight was measured with a periodically calibrated scale for accuracy. The subjects were weighed with light clothing without shoes. The direct height was measured using a stadiometer. Height was measured without shoes; the subject standing upright with feet together. The BMI was determined using participants weight (kg) over their height (m) squared.

2.5.2 HbA1c measurements

The HbA1c were analysed at accredited private laboratories. HbA1c is a reflection of the average blood glucose levels for the preceding two to three months and does not reflect recent changes in glucose levels. HbA1c is expressed as the percentage of total haemoglobin. An HbA1c level above 6.5% is required for the diagnosis of T2DM.

2.5.3 DNA isolation

Peripheral blood was collected in Ethylenediaminetetraacetic acid (EDTA) tubes (BD Vacutainer, Becton Dickinson, South Africa). Genomic DNA was extracted using the Wizard® Genomic DNA Purification Kit (Promega, USA) according to the manufacturer's instructions. Red blood cells were lysed by incubating 1 ml blood with 3 ml Cell Lysis Solution for 10 min at room temperature. The sample was centrifuged at 2,000 x g for 10 min and the supernatant was discarded. One ml Nuclei Lysis Solution was added to the pellet and mixed. Thereafter, 330 µl Protein Precipitation Solution was added and the sample was centrifuged at 2,000 x g for 10 min. The supernatant was transferred to a new tube containing 1 ml isopropanol. Two 70% ethanol wash steps were consequently performed. The ethanol was aspirated and the pellet was rehydrated in 50 µl TE buffer.

The quantity and quality of DNA were assessed using a Thermo Scientific NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, USA). DNA isolations were repeated for all samples until a concentration of 50 ng/µl or higher were reached.

2.5.4 Genotyping

Genotype analysis was performed using TaqMan® chemistry on a qPCR instrument on samples from both cohorts (n=184). The assay was developed to distinguish between the C and G alleles of the Pro12Ala variant (rs1801282) within the *PPARG-2* isoform.

2.5.4.1 Primer and probe design

The target selected for primer design was carefully chosen to include the Pro12Ala variant in an amplicon of between 50 and 150 bp. The target was selected to prevent secondary structures from forming. Forward and reverse primers, as well as probes, were designed using the Integrated DNA Technology (IDT) webpage (Integrated DNA Technology 2016). IDT PrimerQuest was used for the design and selection of primers and probes and IDT Oligo Analyzer allowed for the calculation of oligonucleotide parameters such as melting temperatures, hairpin loop formation and self-dimerization (Integrated DNA Technology 2016).

2.5.4.2 Quantitative PCR optimization

Conventional PCR was performed on a SimpliAmp™ Thermal Cycler (Applied Biosystems, US) to determine optimal conditions for primer binding by evaluating different temperatures and primer concentrations as well as to ensure detection of correct amplicon size. The annealing temperature ranged from 60°C to 66°C. The cycling conditions were 95°C for 5 min and then 35 repeats of a two-step cycling between 95°C for 10 sec and five different annealing temperatures for 45 sec. Thereafter optimized parameters were confirmed on the LightCycler® 480 qPCR instrument (Roche Diagnostics, Germany). The qPCR assay for the Pro12Ala SNP was optimized using two controls in duplicate. The wild type probe was designed according to the HG19 human genome sequence and optimized. The mutant control was manufactured by Inqaba Biotech industries and page quality control measures were taken (Appendix F).

2.5.4.3 Quantitative PCR

The lyophilized primers were reconstituted with TE into a 20 µM (20 pmol/µl) stock solution. The TaqMan® probes were diluted with TE to prepare 20 µM (20 pmol/µl) stock solutions of each probe.

Table 2.2: Tabulated are the forward and reverse primer sequences as well as the wild type and mutant probes.

Primers

Forward	5′-CCC TAT TCC ATG CTG TTA-3′	18 mer
Reverse	5′-CAG ACA GTG TAT CAG TGA-3′	18 mer

Probes

Wild type	HEX-TCTCCTATTGACCCAGAAAGCGATT--BBQ	25 mer
Mutant	6FAM-TCTCCTATTGACGCAGAAAGCGATT--BBQ	25 mer

A multiplex reaction mixture was set up containing 2x LightCycler® 480 Probe Master mix (Roche Diagnostics, Germany), 20 µM forward and reverse primers respectively, 20 µM HEX- and 6FAM- labelled TaqMan® probes respectively and 50 ng DNA template in a total volume of 20 µl. The reaction was conducted using the following conditions: 95°C for 5 min then 35 repeats of two-step cycling between 95°C for 10 sec and 60°C for 45 sec. Quantitative PCR results were analysed on the LightCycler® 480 qPCR Instrument (Roche Diagnostics, Germany) using the genotyping function.

2.5.5 Next Generation Sequencing

2.5.5.1 NGS Workflow

The *PPARG* gene was selected as a target region to design NGS primers using the Ion Torrent AmpliSeq™ Designer. The sequencing library was constructed using the AmpliSeq™ primers and the template was prepared from the automated library and loaded on the 316 Chip v2 BC. Samples were sequenced on the Ion Torrent PGM™ sequencer (Thermo Fisher Scientific, US). Finally, data was analysed, using a pipeline of bioinformatics tools.

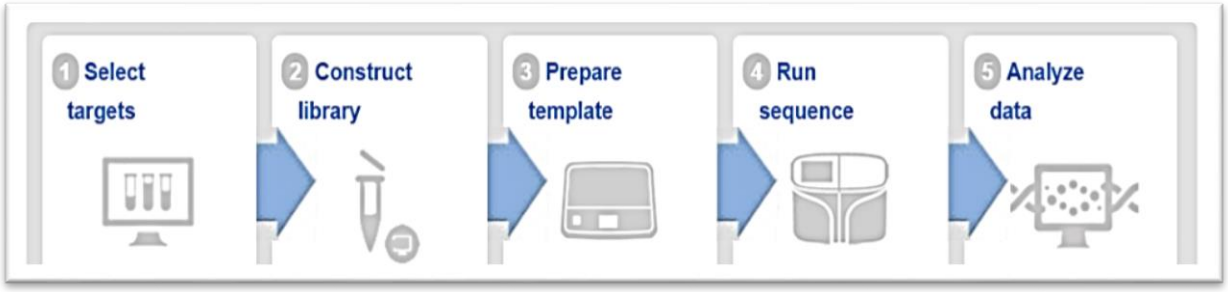


Figure 2.2: Schematic representation of the NGS workflow using the Ion Torrent technology (Thermo Fisher Scientific, US).

2.5.5.2 Ion AmpliSeq™ design

The sequencing library was constructed using the Ion Torrent AmpliSeq™ designer tool (Thermo Fisher Scientific, US). The Ion AmpliSeq™ technology delivers library construction for targeted sequencing of specific human genes or genomic regions. Based on ultrahigh-multiplex PCR, Ion AmpliSeq™ technology targets sets of genes by implementing one or two pools of primer pairs. The Ion AmpliSeq™ Custom Panels designer allowed for the selection of our target gene *PPARG* which was entered online using Ion AmpliSeq™ designer to create customized Ion AmpliSeq™ primer Panels (Life Technologies, US). AmpliSeq™ designer was used for the primer design of five genes, which included *PPARG* that was pooled with four other genes and sequenced on the same chip due to the vast capacity for sequencing per chip. For this study the focus was on *PPARG*. To optimize coverage, the AmpliSeq™ primers were designed with 100 bp spanning the ends of each exon. The design was based on the human GRCh37/hg19 reference genome from the UCSC Genome Browser.

The Ion AmpliSeq™ design amplified the *PPARG* ORF with total sequencing coverage of 99.62%. The target length of the *PPARG* gene, including the coding sequence and untranslated region (CDS + UTR), was 4,263 bp with 16 missed bp. The *PPARG* gene exons which were the input target for AmpliSeq™ (A) is illustrated in Figure 2.3. The region covered with the AmpliSeq™ design which includes the 100 glycosylated haemoglobin bp padding is illustrated in Figure 2.3 B. The rest of

the figure displays other *PPARG* sequences that were submitted to the database and SNPs that has been associated with the gene.

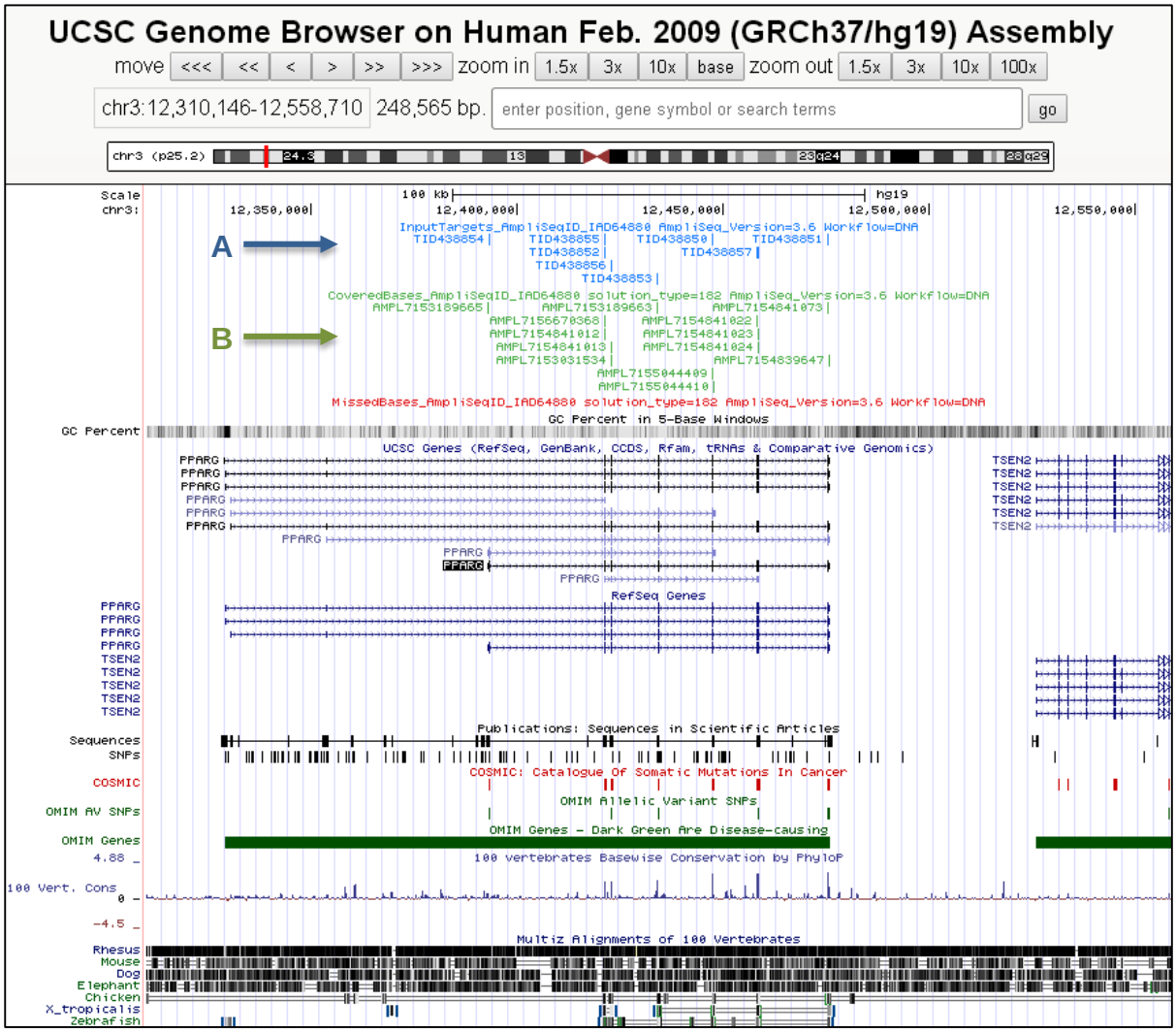


Figure 2.3: The outline of the Ion AmpliSeq™ design in the UCSC Genome Browser. A indicates the input target for the AmpliSeq™ design. B indicates the region covered with AmpliSeq™ design.

2.5.5.3 NGS using Ion Torrent PGM™

A total of 16 samples were sent for NGS on the Ion Torrent PGM™ platform at the University of Pretoria's Sequencing Facility. The samples consisted of eight patient samples and eight controls that were randomly selected and individually matched according to age, sex, BMI and race. The input DNA for the AmpliSeq™ design consisted of two pools of 10 ng DNA (20 ng) containing 74 amplicons and 72

amplicons respectively. The Ion Torrent PGM™ platform was used for sequencing. The 316 Chip v2 BC was selected with a sequencing depth of at least 100x. The Hi-Q™ sequencing kit (Thermo Fisher Scientific, US) was used according to Thermo Fisher Scientific standard protocol. Amplicon sizes ranged from 125 bp to 375 bp. DNA dilutions were prepared as required by the University of Pretoria Sequencing Facility. DNA samples were diluted to a final concentration of 100 ng/µl with a total of 10 µl per microcentrifuge tube.

2.6 Data analysis

2.6.1 Population group

The data from the patient and control cohorts were statistically analysed. The statistical data on age, BMI and HbA1c within each cohort of the study was reported as minimum and maximum values, medians, means, standard deviations (SD) and frequency tables (Microsoft Excel. Microsoft, 2013 software for Windows 2007, Copyright ©).

2.6.2 Genotyping

Allele frequencies were calculated for the C and G allele respectively. The Fisher's exact test was used to calculate whether there is a significant association between the two genotypes observed. This was done by using a 2x2 contingency table (In-silico: Project support 2016).

2.6.3 Next Generation Sequencing analysis

This section describes the bioinformatics analysis pipeline used for the NGS data. General terminology which describes important concepts, terms and general background on shell script programming that is used throughout the data analysis can be found in Appendix B. Figure 2.4 outlines the software that was used for the different steps in the analysis pipeline. Background on the specific software used and how it was applied to our data is given in the subsequent sections. Appendix C contains a summarized table listing all the software used for NGS data analysis.

Additional information pertaining to the analysis such as the basic commands, the parameters that were selected and the commands written for analysis of the data are included in Appendix D.

2.6.3.1 Automated analysis

Torrent Suite™, NextGENe® v.2.3.4 (SoftGenetics, State College, PA, USA) software was used to perform the primary analysis. This is preinstalled on the Torrent Server for automated sequencing data analysis. Files obtained were processed by Ion Torrent Suite™ software at the University of Pretoria Sequencing Facility. Reads were acquired in unmapped BAM files from which primers and adapters were already removed.

2.6.3.2 Manual Analysis

Linux bash was used to run all commands and scripts. Shell scripts were used since existing tools were too slow or too restricted in that they only allowed editing and analysis of one sample at a time. Shell scripts allowed for multiple samples to be analysed. The University of the Free State High-Performance Computing (HPC) cluster was used to upload files for processing to increase the speed with which analysis was done.

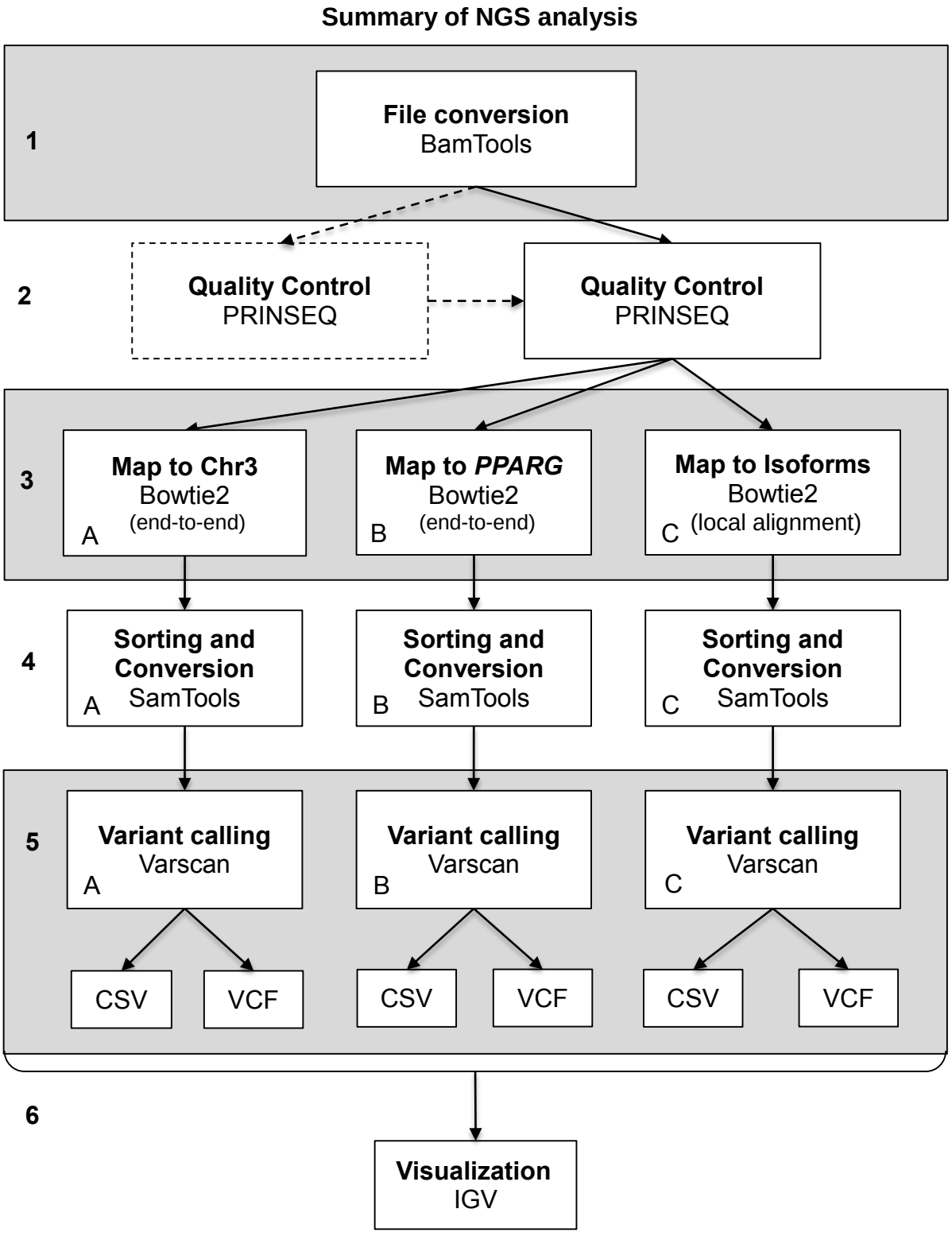


Figure 2.4: Outline of NGS data analysis pipeline showing the different computational tools used for the various steps. The dotted line represents the initial QC that was performed without the use of parameters. Thereafter parameters were included and analysis continued.

1. File conversion

The sequencing facility (University of Pretoria Sequencing Facility) provided the raw reads from the Ion Torrent PGM™ platform in BAM file format. BamTools was used to convert the BAM files to FASTQ files which are the required format for input into PRINSEQ for quality control (QC).

1.1 BamTools

The BamTools API provides programmers and end-users with an easily accessible, user-friendly interface for generating and querying BAM files (Barnett *et al.*, 2011). The BamTools software suite facilitates research analysis and data management using BAM files. Features in BamTools can be tailored specifically for NGS data analysis. These features include alternative indexing formats, conversion of alignment data to other formats such as FASTQ or BED files, basic coverage outputs and the ability to split a BAM into multiple files in a single command (Barnett *et al.*, 2011).

1.2 Conversion of BAM files to FASTQ files

The data received from the University of Pretoria consisted of a compilation of the sequencing data of all five genes including the *PPARG* gene of interest. During the mapping of the sequence reads to the *PPARG* reference gene (section 3) the sequences specific to *PPARG* was filtered out.

The command line created in Linux was used for the conversion of the 16 BAM files to FASTQ files using BamTools. The BAM files consisted of controls and patients designated in the script below by C and P respectively. The control and patient cohorts each contained 8 samples labelled [1..8] (implying that the action will be performed for sample 1 to 8 respectively). The script was written to contain a nested loop that converted all the files for the controls and patients in series. This was done by inserting a variable ($) for files containing a C and a P and a variable for the individual samples [1..8].

Command for conversion of BAM files to FASTQ files for controls and patients:

```
for e in C P
      for i in [1..8]
      do
             bt2_conv_cmd="bamtools convert -format fastq -in
             filename-$e$i.bam -out filename-$e$i.fastq"
             echo $bt2_conv_cmd
             eval $bt2_conv_cmd

      done
done
```

2. Quality Control

2.1 PRINSEQ

PRINSEQ (PReprocessing and INformation of SEQuence data) is currently one of the best methods for analysing NGS data (Shrestha *et al.*, 2014). It is easily configurable and provides a user-friendly interface. PRINSEQ was chosen to filter, reformat, and trim our sequence data according to specific parameters to improve the downstream analysis. The summary statistics of the output file (FASTA or FASTQ) was generated in a table as well as in graphical form (Schmieder and Edwards, 2011). For our data, the stand-alone command line was used for the latest versions of prinseq-lite (version 0.20.4) and prinseq-graphs (version 0.6).

The prinseq-lite version is primarily designed for data pre-processing and does not generate summary statistics in graphical form. The lite version is a standalone perl script (prinseq-lite.pl) that does not require any non-core perl modules for processing. Prinseq-graphs are used to generate readable graphs that can be interpreted for human reading and analysis. The graphs version is also a standalone perl script (prinseq-graphs.pl) and generates graphs (PNG files) and HTML report files.

2.2 Quality assessment

First the raw data was processed with PRINSEQ without doing any QC, to assess the overall sequences quality in order to choose the appropriate parameters for

efficient data trimming and filtering. Prinseq-lite was used to generate data. The script was written to contain a nested loop that generated data for the controls and patients in series (Appendix D). The output format for running prinseq-lite is gd files. Summary statistics were graphically presented with prinseq-graph. This command used for prinseq-graph was similar to the command used to run prinseq-lite consisting of a nested loop to analyse controls and patients simultaneously (Appendix D).

2.3 Quality controlled data

After data was empirically evaluated, the following parameters were included in the command to obtain sufficiently high quality data. Parameters were chosen based on quality and quantity of the sequencing reads. Ion Torrent reads typically have lower quality bases towards the 3' end of each read (Bragg *et al.*, 2013), and may also contain adapter or primer sequences at the 3'. With our data, the adapter and primer sequences were already removed using the Ion Torrent Suite software and trimming only had to be done on the low quality bases from the 3' end of each read. Sequence reads were right trimmed using a minimum Q-score of 30. The average quality is calculated over a sliding window from the 5’ to 3’ end removing the 3’ ends when quality is below the defined threshold. This was done due to the decline of quality at 3’ ends of sequencing reads. A Phred quality value of 30 corresponds to an error probability of 0.0001 thus, the base call accuracy is 99.9%. Shorter sequences were also trimmed with a minimum length of 40 bp. This was an arbitrary value chosen to make mapping more specific since shorter sequences map less specific.

3. Mapping

Mapping of the sequence to the reference sequence is the most important step in the SNP detection process (Horner *et al.*, 2010). Mapping of sequence reads was performed after the data was trimmed in PRINSEQ. Bowtie2 (Langmead and Salzberg, 2012; Langmead *et al.*, 2009) was chosen for mapping of sequences. It is a well-cited mapping program used in many studies (Bhuvaneshwar *et al.*, 2015; Hatem *et al.*, 2013; Horner *et al.*, 2010).

3.1 Indexing and conversion

Bowtie2 is a fast and efficient tool used for aligning sequence reads to long reference sequences (Langmead *et al.*, 2009). It is especially good at aligning reads of between 50 bp and 1000 bp to relatively long genomes and is optimized for the read lengths and error modes yielded by the Ion Torrent, Illumina HiSeq 2000 and Roche 454 instruments. Bowtie2 supports gapped, local and paired-end alignments (Langmead *et al.*, 2009). The output file for alignments performed in Bowtie2 is SAM format which enables interoperation with other tools such as SAMtools.

The build module of Bowtie2-build creates index files of the reference sequence, similar to Burrows Wheeler Aligner (BWA) (Li and Durbin, 2009). These files together constitute the index; they are all that is needed to align reads to the reference. The original FASTA files are no longer used by Bowtie2 once the index is built.

SAMTools (Li *et al.*, 2009) provide a set of utilities for manipulating alignments in the SAM or BAM format, including sorting, merging, indexing and allows for retrieving of reads in any regions swiftly. SAMtools was used for conversion (samtools_conv) of SAM to BAM files and for sorting (samtools_sort) and indexing (samtools_index) quality controlled data which is necessary for analysis in IGV.

3.2 Map to references

Bowtie-build was first used to build an index file using Chromosome 3 (Chr3), the full *PPARG* gene and the three *PPARG* isoforms as the reference. Thereafter a script was written to map sequences using Bowtie2. The input option for Bowtie2 was FASTQ files and the output file after mapping is in SAM format. For mapping sequences to Chr3 a sensitive end-to-end alignment was done and 16 processors (CPU's) were used for rapid results. In the case of mapping to the complete *PPARG* gene, Bowtie2 commands were altered to substitute the reference sequence to which mapping should be done with the FASTA file of the entire *PPARG* gene (gi568815595). SAMtools was used to convert SAM files to BAM files. The BAM files were required as the input format for SAMtools to create mpileup files and thereafter the files were sorted and indexed (Appendix D).

The procedure was repeated for mapping to the three *PPARG* isoforms with minor changes. Indexes were built for all three isoforms using Bowtie-build. The commands for running Bowtie2 were edited slightly to substitute end-to-end alignment with local alignment and the reference sequence was substituted with the FASTA files for the three *PPARG* isoforms (Isoform 1: ENST00000397015; Isoform 2: ENST00000287820; Isoform 3: ENST00000397015).

4. Sorting and Conversion

The variant calling features of VarScan 2 for multiple samples (mpileup2snp) expect input in SAMtools mpileup format. The building of an mpileup file requires the following: BAM files that have been sorted and indexed using the *sort* and *index* commands in SAMtools and the reference sequence to which reads were aligned, in FASTA format, SAMtools was used in both cases.

5. Variant calling

Variant calling is the process of identifying genetic variation in sequencing data, such as single nucleotide variants, copy number variations, structural variants, indels and inversions. VarScan 2 (Koboldt *et al.*, 2012) is a software tool developed at the Genome Institute at Washington University for the detection of variants in NGS data. VarScan 2 is written in Java and was executed from the command line.

The VarScan 2 command expects an input file in SAMtools mpileup format from sequence alignments in BAM format. It also requires the reference sequence to which reads were aligned, in FASTA format. VarScan 2 creates two output files, one is a Variant Call Format (VCF) file and a Comma Separated Value (CSV) file. A command was written for variant calling using the default parameters for Varscan 2 (Appendix D). By default, VarScan 2 requires a minimum coverage of 33, minimum Phred base quality of 20, allele frequency of at least 8%, and a P-value of <0.05. Variants with a variant allele frequency of >75% are called homozygous. First, Chr3 was used as the reference sequence to which reads aligned. Thereafter the process was repeated for reference against the *PPARG* full gene (B) and using the three isoforms (C) as the reference

6. Visual Inspection of mapped region and SNPs

The Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011; Thorvaldsdóttir *et al.*, 2013) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. IGV supports a wide range of different data types, including NGS data and genomic annotations. Each patient and control sorted BAM and VCF file was uploaded in IGV together with the relevant reference sequence file (FASTA) and visually inspected for SNPs.

IGV user interface (Figure 2.5).

1. The toolbar provides easy access to frequently used functions.
2. The red box on the chromosome ideogram indicates the portion of the chromosome which is displayed.
3. The ruler reflects the visible portion of the chromosome or gene. The span lists the number of bases currently displayed.
4. IGV displays data in horizontal rows called tracks. Typically, each track represents one sample or experiment. Track names are listed in the far left panel. Tracks can be added and removed by right clicking and selecting ‘Load track’ or ‘Remove track’.
5. IGV can display features, such as genes, in tracks. By default, IGV displays data in one panel and features in another. Data and features panels can be combined by selecting the option on the General tab.
6. VCF file formats display variations in sequence. The bar across the top of the plot displaying the bars marking variant calls.
7. Coverage Track: IGV dynamically calculates and displays the default coverage track for an alignment file. When IGV is zoomed to the alignment read visibility threshold (by default, 30 KB), the coverage track displays the depth of the reads displayed at each locus as a grey bar chart. If a nucleotide differs from the reference sequence in greater than 20% of quality weighted reads, IGV colours the bar in proportion to the read count of each base (A, C, G, T).
8. Alignment Track: When zoomed into the alignment read visibility threshold, by default 30 KB, IGV shows the reads.

9. Reference genome sequence track: The sequence is represented by coloured bars or coloured letters, depending on zoom level, with adenine in green, cytosine in blue, guanine in yellow, and thymine in red. With the reference genome sequence track, you can optionally display a 3-band track that shows a 3-frame translation of the amino acid sequence for the corresponding nucleotide sequence. The translation is shown for the strand indicated. Amino acids are displayed as blocks coloured in alternating shades of grey. Methionines are coloured green, and all stop codons are coloured red.

Figure 2.5: Illustration of the IGV user interface. The numbers represent different features of IGV and the descriptions are given above.

2.7 Validity and reliability

The validity of a measuring instrument refers to the degree to which the instrument being used measures what it is intended to measure (Leedy and Ormrod, 2010). A calibrated electronic scale is recommended for direct weight measurement and was used to assure the validity of anthropometric measurements (DeBruyne *et al.*, 2011). Height measured was done with a stadiometer, which is a standard direct method (DeBruyne *et al.*, 2011). The measurement of HbA1c levels was validated by the World Health Organization (WHO) as a means to diagnose T2DM (World Health Organization, 2011). Quantitative PCR was performed using the TaqMan® detection chemistry which is a well-established method for sensitive and specific genotyping (Heid *et al.*, 1996). The Ion Torrent PGM, based on Ampliseq™ libraries, is fast, cheap, high-throughput and a highly efficient NGS strategy. It fulfils conditions required for the systematic detection of genomic variants in the most prevalent genes of a disease and is ready to be deployed in clinical laboratories (Millat *et al.*, 2014).

Reliability refers to the steadiness with which a measuring instrument yields a certain outcome when the entity being measured hasn't changed (Leedy and Ormrod, 2010). Reliability will be assured by standardizing the measuring instruments and a registered dietician or nurse took the anthropometric measurements. Controls used in the qPCR assay were either sequenced or synthetically manufactured and ran in duplicate to ensure reliability. The Ion Torrent software provides a high sensitivity mutation detection of gene variants composed of substitutions and indels. The Ion 316™ Chips yield 99.99% consensus accuracy.

2.8 Ethical considerations

Ethical approval for this study was obtained from the Ethics Committee of the Faculty of Health Sciences, at the University of the Free State before the onset of the study (Ethics reference number ECUFS 53/2015).

Approval and permission to conduct the study were obtained from the CEO's of the institutions involved. Written informed consent (Appendix A) was obtained from all

participants, providing them with an information letter in which the procedures were explained in simple and understandable terms in the language of their choice (Sotho, English or Afrikaans). The risks associated with drawing of blood (physical discomfort, potential risks for infection or bruising) were included in informed consent. Consent was also obtained to store and use the participant's DNA sample for genetic research. Participants were also informed that the results of this study may be published. Participation in the study was voluntary and participants were free to withdraw from the study at any time. Confidentiality was retained during all stages of the research by ensuring that no names are disclosed, or written down in the data analysis spreadsheet. No names were used during data analysis and in the discussion of results, only numbers were used.

2.9 Conclusion

The methodology for conducting this study was described in this chapter. The study population and the selection of samples were discussed, as well as the procedures used and data that were collected during this study. The methods that were used are explained, as well as the methods used for statistical analysis. The larger part of this chapter contains the bioinformatics pipelines that were used for the analysis of NGS data. This section is a detailed outline of the steps that were followed to obtain high-quality sequences that led to accurate results. The detail of the commands used are described in Appendix D. Lastly, this chapter ends with the validity and reliability of the study and the ethical aspects that were considered.

CHAPTER THREE

RESULTS AND DISCUSSION

3.1 Study population

The study population and samples were collected as described in the Methodology chapter. Tables containing information and data related to the results chapter are provided in Appendix G. Descriptive statistics were used to analyse data.

3.1.1 T2DM patients cohort

The frequencies of the age distribution, body mass index (BMI) and HbA1c for the patient cohort is illustrated in Figure 3.1, 3.2 and 3.3 with the data distribution in Table 3.1. The age of the participants ranged between 35 and 61 years, with the largest grouping in the 50-59 years category and an average of 50.61 years for the patient cohort. The BMI of the participants ranged from normal to obese class 3, with no individuals in the underweight category. The largest groupings of individuals were found in obese class 1 and obese class 3. The mean BMI was 35.78 (SD ± 7.55) kg/m² and falls within the obese class 2 group which is high, but was expected for individuals suffering from T2DM. The HbA1c levels of the patient cohort were divided as being optimal (<5.7%) or sub-optimal (≥5.7%) since all patients had already been diagnosed with T2DM by a medical doctor. The average HbA1c for this cohort was 8.74%.

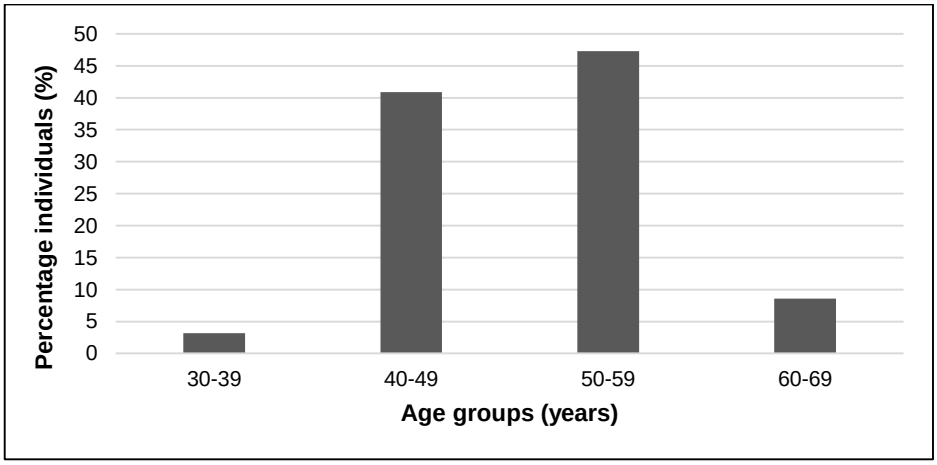


Figure 3.1: The graph shows the distribution of patients in each age group (n=93).

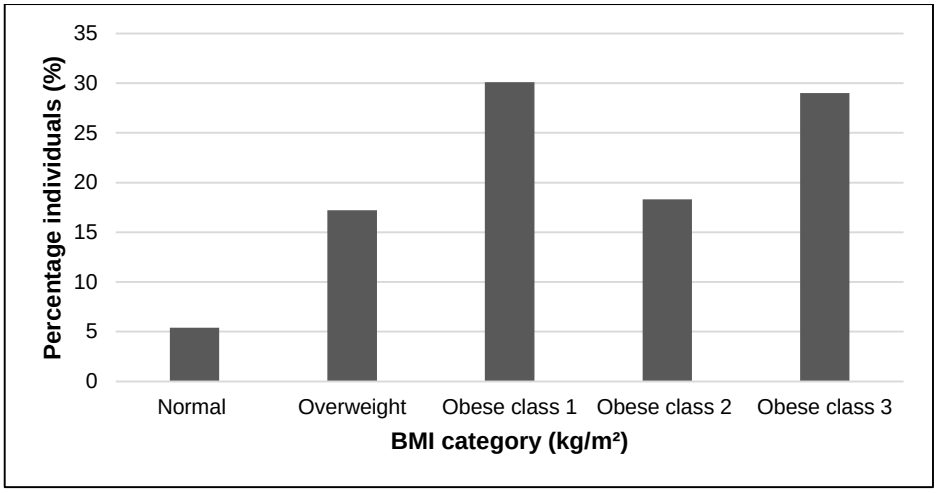


Figure 3.2: The graph shows the percentage patients in each BMI category (n=93). The BMI ranges (kg/m²) for each category is as follows: Normal (18.5-24.9), Overweight (25-29.9), Obese class 1 (30-34.9), Obese class 2 (35-39.9) and Obese class 3 (>40).

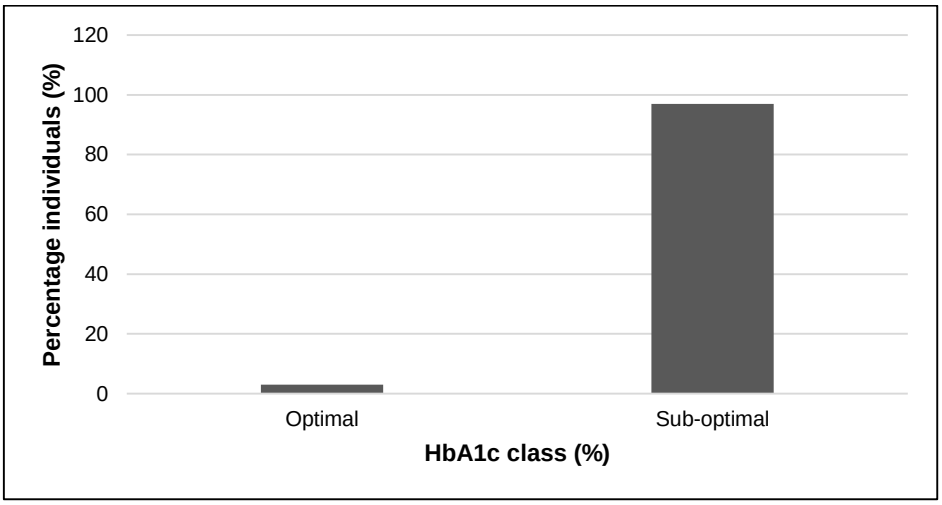


Figure 3.3: The graph shows the dispersal of patients HbA1c levels (n=93). The patient HbA1c levels (%) are classified as being Optimal (<5.7) or Sub-optimal (≥5.7).

Table 3.1: Summary of data distribution of patients with T2DM (n=93).

Variable	Min	Max	Median	Mean (SD)
Age	35	61	51	50.61 (6.51)
Weight (kg)	52.3	130.6	84.7	88.52 (20.33)
Height (cm)	141.5	172.0	157.0	157.07 (5.88)
BMI (kg/m²)	21.5	52.2	34.4	35.78 (7.55)
HbA1c (%)	5.3	15.8	8.3	8.74 (2.33)

3.1.2 Control cohort

The frequencies of the age distribution, BMI and HbA1c for the control group are illustrated in Figure 3.4, 3.5 and 3.6 with the data distribution in Table 3.2. The age of the participants ranged between 36 and 62 years, with a mean age of 48.96 years. As indicated in Table 3.2, the age distribution showed almost equal frequencies between the 40-49 and 50-59 year age groups. The BMI of control individuals ranged from normal to obese class 3. The participants mainly fell into the overweight and obese class 1 group with an average BMI of 32.33 (SD ± 6.19) kg/m². All control participants had HbA1c levels below 6.50%. The average HbA1c for this cohort was 5.50%.

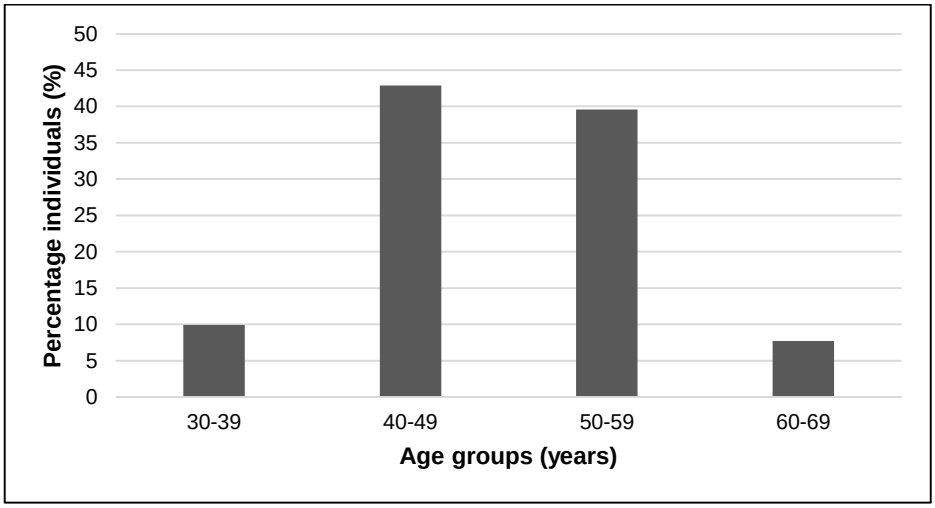


Figure 3.4: The graph shows the distribution of controls in each age group (n=91).

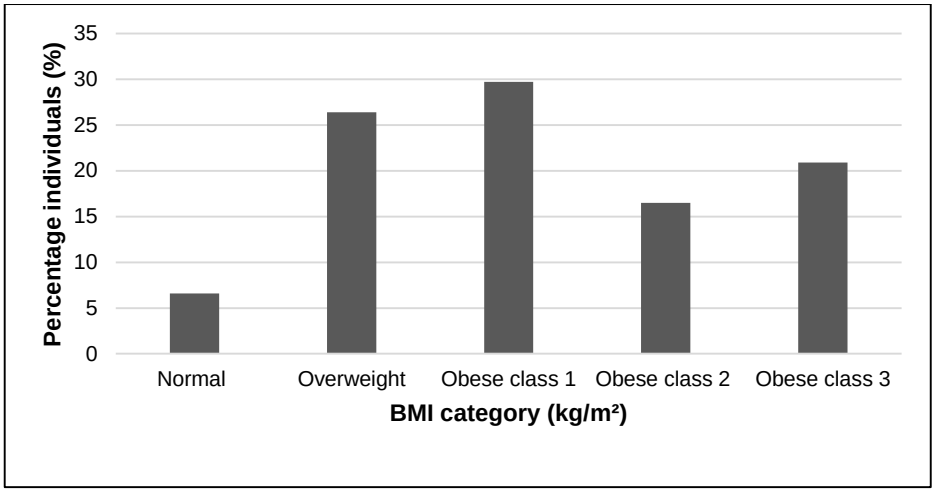


Figure 3.5: The graph shows the percentage controls in each BMI category (n=91). The BMI ranges for each category is as follows: Normal (18.5-24.9), Overweight (25-29.9), Obese class 1 (30-34.9), Obese class 2 (35-39.9) and Obese class 3 (>40).

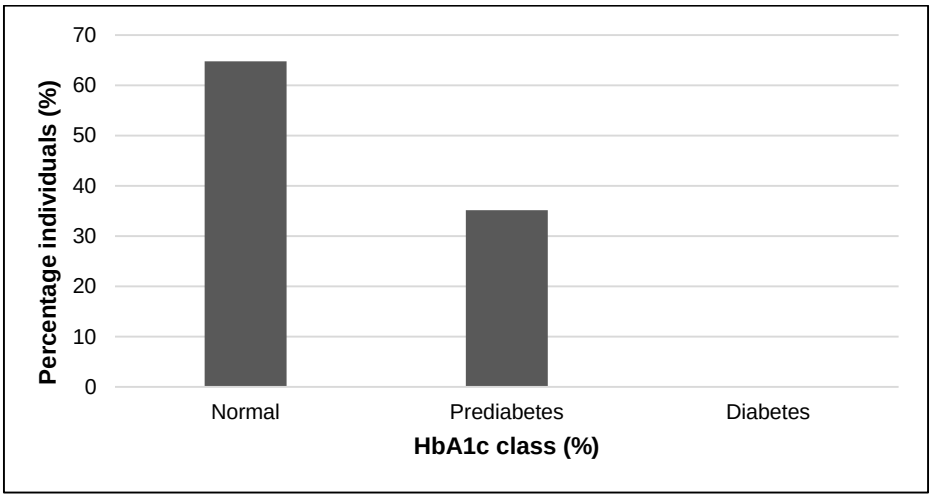


Figure 3.6: The graph shows the dispersal of the controls in the different HbA1c classes (n=91). The HbA1c classes are Normal (<5.7), Prediabetes (5.7-6.4) and Diabetes (>6.4).

Table 3.2: Summary of data distribution of controls participants (n=91).

Variable	Min	Max	Median	Mean (SD)
Age	36	62	49	48.96 (±7.45)
Weight (kg)	52.0	120.2	82.2	83.63 (16.14)
Height (cm)	145.2	176.0	157.0	157.78 (6.49)
BMI (kg/m²)	21.0	47.3	32.3	32.33 (6.19)
HbA1c (%)	4.3	6.4	5.5	5.50 0.41)

3.1.3 Conclusion

Figure 3.1 and 3.4 show an almost equal distribution of age groups for both patient and control groups and a similar average age for both cohorts. This is to be expected since the control group was selected to match the patient group for gender, age and BMI. The patient and control groups both had the majority individuals falling into the obese class 1 group. The mean BMI for the patient group was slightly higher than that of the control group. This can be explained by the control group being recruited as a convenience sample from public supermarkets selected to fall within the age and BMI range of the T2DM patient group.

The mean HbA1c level of 8.7% indicate that most patients' glucose levels are not controlled sufficiently and not within the normal range. This is of great concern, because it increases the patients risk to micro- and macro vascular complications or even death (Monami *et al.*, 2013). The HbA1c levels of the control cohort were all below 6.4%. Once again this is expected since the control participants were excluded from the study if their HbA1c levels were 6.5% or higher. Individuals with prediabetes were included in the control group according to the diagnosis of T2DM which require an HbA1c above 6.5%. It is alarming that 35% of the control subjects that view themselves as having no blood sugar problems, is prediabetic (HbA1c 5.7% to 6.4%). The general public has little knowledge about diabetes prevention and the basic principles of nutrition, physical activity and blood glucose control. Health education is crucial to people suffering from T2DM as well as to the general public, as adequate knowledge of the disease is associated with good metabolic control and prevention of complications (Afridi and Khan, 2003).

3.2 Quantitative PCR

3.2.1 Primer and probe design

Primers and probes were designed using the IDT web access page (Integrated DNA Technology 2016). The primers were both 18 base pairs in size for adequate specificity. Melting temperatures for the two primers differed with only 0.4°C, thus allowing both primers to bind simultaneously and efficiently to amplify the product. The melting temperatures (Tm) of the primer pair was below the ideal range, however, primer binding was still specific to the target region with successful amplification and no unspecific amplification. The GC content was 44.4%, which allows for complexity while still maintaining a unique sequence. The primer sequences do not contain regions with four or more consecutive G residues. The ΔG values for hairpin formation (self-dimerization) and cross-dimerization are all above the acceptable ΔG values thus not favouring the formation of any secondary structures. Primers were successfully designed according to the various parameters outlined in Table 3.3. The primers were designed to produce an amplicon of 115 bp.

Table 3.3: The parameters to which forward and reverse primers were selected.

Parameter for primer design	Ideal	Design for Forward	Design for Reverse
Primer length	15-30 bp	18 bp	18 bp
Melting Temperature	Tm 58-60°C	Tm 47.7°C	Tm 48.1°C
GC content	35-65 %	44.4%	44.4%
Avoid hairpin	ΔG > -2 kcal.mol^{-1} Less than 4 bp bind	ΔG ≥ 1.2 kcal.mol^{-1}	ΔG ≥ -1.08 kcal.mol^{-1}
Avoid cross-dimerization	ΔG > -5 kcal.mol^{-1}	ΔG ≥ -4.89 kcal.mol^{-1}	ΔG ≥ -4.89 kcal.mol^{-1}
Primers to probe binding	No complementarity	✓	✓
Cross homologs	Specifically binds to target of interest	Yes	Yes
Blast (NCBI)	No complementary binding	✓	✓
Amplicon size	50-150	115 bp	

Probes were designed according to the parameters outlined in Table 3.4. Both probes were 25 bp in length with the reporter dye on the 5' end and the quencher on 3' end. The Tm was 61.3°C and 62.2°C for HEX (wild type) and 6FAM (mutant) probes respectively with a GC content of 44% for both probes. No G was placed on the terminal of the probe due to the quenching effect of G. Probes are not identical and low complementation will prevent dimers from forming.

Table 3.4: The table sets out the variables (parameters) to which wild type and mutant probes were selected.

Parameter	Ideal	Wild type probe (HEX)	Mutant probe (6FAM)
Probe length	15-30 bp	25 bp	25 bp
Melting temperature	Tm 65–72°C 5-10°C > Tm of primers	Tm is 61.3°C	Tm is 62.2°C
GC content	30-60%	44%	44%
Terminal G	Should be avoided (due to quenching effect)	✓	✓
Primer and probe complementation	Less than 4 bp complementarity	✓	✓
Dyes and Quenchers	Dye and Q must have FRET	✓	✓

The wild type, Pro12 allele-specific probe was labelled with HEX fluorescent dye and the mutant, Ala12 allele-specific probe was labelled with 6FAM fluorescent dye. A BBQ quencher was attached to both probes as suggested by the manufacturer (Roche Diagnostics, Germany). The primers and probes were synthesized by TIB Molbiol (Roche Diagnostics, Germany). Figure 3.7 is a diagram illustrating where the forward and reverse primers annealed to the exon B of *PPARG*. The sequence where the probe (wild type) annealed to the exon is shown and the Pro12Ala SNP position is indicated in the block.

Figure 3.7: Illustration of where primers and probes bind to exon B of *PPARG*-2. The boundaries for exon B is shown in yellow. The primer sequences are indicated by the green (forward) and blue (reverse complement) colours. The wild type probe sequence is shown in red with the Pro12Ala SNP position depicted in the purple block.

3.2.2 Quantitative PCR optimization

In order to achieve optimal genotyping results the annealing temperatures of the primers and probes had to be optimized. The cycling conditions used for optimization were used as suggested by the qPCR Master mix manufacturer (Roche Diagnostics, Germany). Five annealing temperatures were tested and the results are presented in Figure 3.8. The temperature selected for optimal primer annealing in the qPCR assay was 60°C (lane 2). Primers annealed optimally at this temperature where a single, high intensity fragment can be observed. At 64°C and 66°C no fragment was observed indicating that amplification at this temperature was unsuccessful. The no template control (NTC) was negative, indicating that no contamination was observed. The optimized cycling conditions were as follows: 95°C for 5 min and then 35 repeats of a two-step cycling between 95°C for 10 seconds and 60°C for 45 seconds. These conditions were used for all patient and control samples.

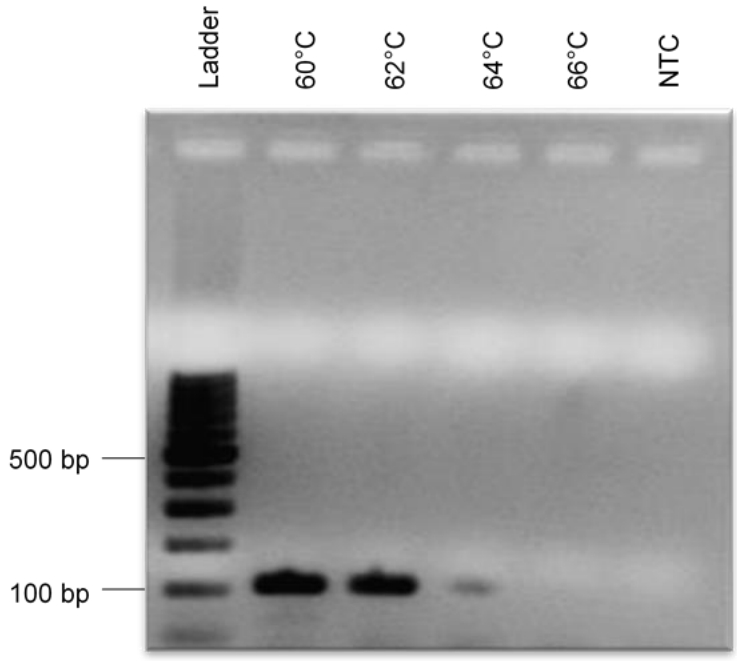


Figure 3.8: Agarose gel image showing the results for optimization of primer annealing at different temperatures. A 1% TAE agarose gel visualized by ethidium bromide staining. In the first lane 1 µl of a 100 bp + 50 bp DNA Ladder (www.sibenzyne.com) was loaded. Lane 2-5 shows the PCR products of 115 bp for temperatures ranging from 60°C to 66°C. The no template control (NTC) is indicated in lane 6.

3.2.3 Quantitative PCR results

The complete patient and control cohort (n = 184 samples) was screened for the Pro12Ala variant using the optimized qPCR assay. The frequencies of the different genotypes observed during the screening process are shown in Table 3.5. A total of 98.9 % of the patients presented with the homozygous Pro/Pro allele and only a single individual presented with the heterozygous Pro/Ala allele of the *PPARG* Pro12Ala (rs1801282) genotype (Figure 3.10). All control participants carried the Pro/Pro genotype as indicated in the table. Figure 3.9 shows an amplification plot of the qPCR assay conducted on 10 patients. The fluorescence on the y-axis corresponds to the wavelength used to detect the FAM dye. The amplification profiles with a crossing time of between 13-16 cycles represent of the mutant control labelled with the FAM dye (done in duplicate). The amplification profiles with a

crossing time of between 25-28 cycles represent the duplicate heterozygote sample (F049). All the other samples showing no amplification on the plot, are homozygous for the wildtype allele labelled with HEX, since only FAM dye is detected under this wavelength. The heterozygote crossing time or Ct value is much later than the control because only half the copies of DNA in the heterozygote sample are detected with the FAM dye. A similar plot for the HEX wavelength was studied with similar results, but with all the homozygote wildtype samples at Ct value of between 22 and 25 cycles and the heterozygote Ct value of 30. The non- template control shows no amplification in either the HEX or FAM wavelength pots indicating that there was no contamination present. Figure 3.10 shows the scatter plot corresponding to the amplification plot of the ten patient samples. The mutant probe was labelled with the FAM dye and the wild type probe was labelled with the HEX dye. The FAM wavelengths are indicated on the y-axis and the HEX wavelengths on the x-axis. The green arrow indicates the mutant control, labelled with the FAM dye and has the highest fluorescence at the x-axis. The single heterozygote Pro/Ala genotype detected by the assay is shown by the red circle. The orange block indicates all the homozygote wild type control and samples, labelled with the HEX dye. The NTC is indicated by the blue arrow confirming that there was no contamination present in these samples. The complete data set containing the genotype of each individual patient and control is set out in Appendix E in Table 6 and 7.

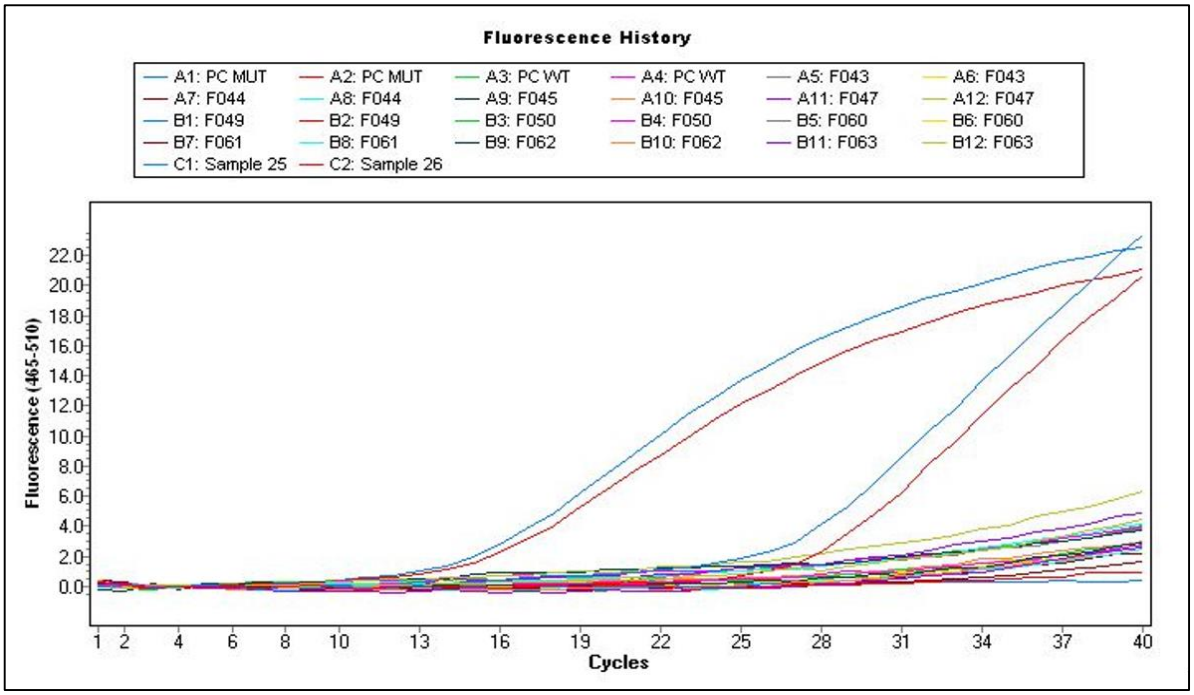


Figure 3.9: Example of the qPCR results shown as an amplification plot of ten patient samples. The fluorescence is indicated on the y-axis and the cycle number on the x-axis. The fluorescent range in this plot correlates with the wavelength of the FAM dye.

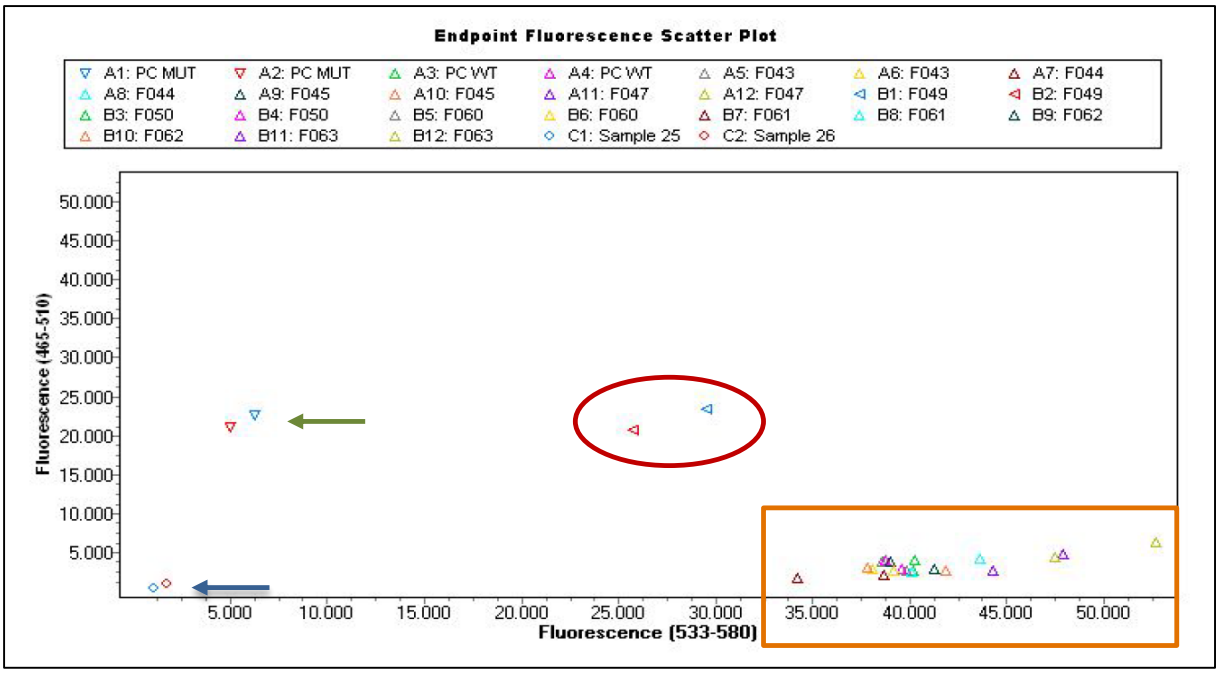


Figure 3.10: Example of the qPCR results shown as a scatter plot of ten patient samples. The mutant probe was labelled with FAM, and the wild type probe with

HEX. The FAM signals are indicated on the y-axis and the HEX signals on the x-axis. The no template control (NTC) is indicated by the blue arrow. The green arrow indicates the mutant control homozygote, labelled with the FAM dye and has the highest fluorescence at the x-axis and the lowest on the y-axis. The single heterozygote Pro/Ala genotype detected by the assay is shown by the red circle. The orange block indicates all the homozygote wild type control and samples labelled with the HEX dye.

Table 3.5: The distribution of genotype frequencies of the Pro12Ala variant for patient and control participants.

Genotype	Frequency	Percentage
Patients		
Pro/Pro (C/C)	92	98.9%
Pro/Ala (C/G)	1	1.1%
Ala/Ala (G/G)	0	0%
Total	93	

Genotype	Frequency	Percentage
Controls		
Pro/Pro (C/C)	91	100%
Pro/Ala (C/G)	0	0%
Ala/Ala (G/G)	0	0%
Total	91	

3.2.4 Statistical analysis of qPCR results

3.2.4.1 Allele frequencies

The allele frequencies are depicted in Table 3.6. In the control cohort, the C allele frequency was 1 and the G allele frequency 0. In the patient cohort, the C allele frequency was 0.99 and the G allele frequency 0.01.

Table 3.6: Allele frequency in the control and patient cohort.

	Controls	Patients
Allele C	C: 182/182 = 1	C: 184/186 = 0.99
Allele G	G: 0/182 = 0	G: 2/186 = 0.01

3.2.4.2 Fisher's exact test

For the Fisher's exact test the total for each genotype per cohort was used (In-silico: Project Support 2016). Observation A and B was the patient and control cohort respectively, Group 1 being the C/C genotype and Group 2 the C/G genotype. The results indicated a two-tailed p-value of >0.9999. The groups were found to be not significantly different. Thus the genotype shows no association to the T2DM or control group.

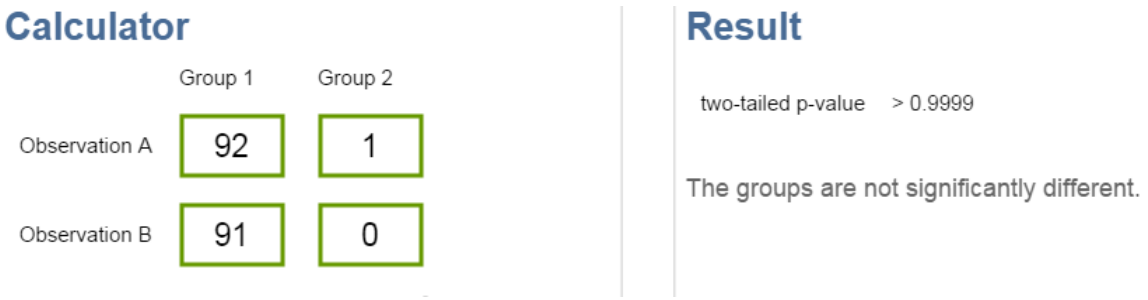


Figure 3.11: The calculation and results for the Fisher's exact test to determine if the two genotypes observed in the population are significantly associated with each other.

Results from the Fisher exact test calculated a two-tailed p-value of > 0.9999, indicating the SNP is not statistically significant (Figure 3.11). Since only a single patient was found to exhibit the heterozygous genotype (Figure 3.10), this was expected. The remaining patients and control cohort have presented the homozygous Pro/Pro genotype. Contrary to previous research the *PPARG-2* Pro12Ala polymorphism did not prove to have any relation to T2DM in the back population in this study.

3.2.5 Conclusion

Published research confirmed that the *PPARG-2* Pro12Ala polymorphism is related to T2DM, with the Ala allele having the protective effect against T2DM (Gouda *et al.*, 2010). Several studies indicated that the *PPARG* Pro12Ala polymorphism was associated with T2DM in a French population with obese subjects (Ghoussaini *et al.*, 2005), an African American population (Kao *et al.*, 2003) an Iranian population (Motavallian *et al.*, 2013), a community of Khatri Sikhs in Northern India (Sanghera *et al.*, 2008) and a Western Indian population (Shahrjerdi *et al.*, 2013). Gouda *et al.* (2010) conducted a meta-analysis involving 32,894 cases and 47,456 controls from 60 studies, concluding that the *PPARG* Pro12Ala polymorphism is positively associated with a reduction in the risk of T2DM. Findings from these populations regarding the *PPARG-2* Pro12Ala polymorphism are in contrast to the results obtained in this study from the South African black female population.

The small sample size of this study may be one explanation for the contradiction to literature, since most studies that did not find an association between the Pro12Ala variant and T2DM had sample sizes less than 1,000 participants (Bener *et al.*, 2015; Majumdar *et al.*, 2014; Motavallian *et al.*, 2013). The studies that found an association were mostly larger studies with sample sizes above 1,000 participants (Ghoussaini *et al.*, 2005; Lamri *et al.*, 2012; Trombetta *et al.*, 2013). However, many meta-analysis studies have also not found a significant association between T2DM and the Pro12Ala SNP (Guo *et al.*, 2011; Tong *et al.*, 2012; Tonjes *et al.*, 2006).

To add to the contradiction of the *PPARG* polymorphism association to T2DM, the absence of the homozygotic Ala/Ala genotype and the very low prevalence of the heterozygotic Pro/Ala genotype in this mainly obese, black, female group with T2DM were also found in other studies. Populations that also did not present with a significant association between the Pro12Ala variant and T2DM included Caucasians from Spain (Miramontes González *et al.*, 2014), individuals from the Eastern region of India (Majumdar *et al.*, 2014), Qatari population (Badii *et al.*, 2008), Tunisians (Zouari Bouassida *et al.*, 2005), a Chinese population (Ye *et al.*, 2014), a Polish population (Malecki *et al.*, 2003) and Asian populations (Al-Safar *et al.*, 2015; Tai *et al.*, 2004).

Literature as well as the results from this study show that the *PPARG-2* Pro12Ala polymorphism cannot be associated with T2DM nor did it have a significant presence in any of the two cohorts studied. It could be that the *PPARG-2* Pro12Ala polymorphism is rare in the black South African population. According to the 1000 Genomes Project this variant is much more common in the European and American populations compared to the African and East Asian populations as illustrated in Figure 3.12 (Genomes Project *et al.*, 2012; Vergotine *et al.*, 2014b).

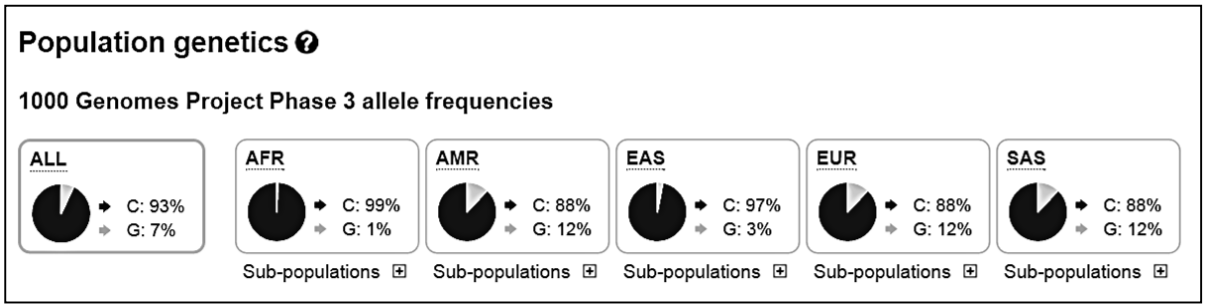


Figure 3.12: Population genetics for variant rs1801282. The G allele has a frequency of 7% in the general population (ALL), but decreases to 1% in the African (AFR) population and also in the East Asian (EAS) population. The frequency of the G allele is much more prevalent in the American (AMR), European (EUR) and South Asian (SAS) populations (Ensembl.org).

Another possibility can also be that the Pro12Ala SNP is gender specific. A study in a Tunisian population provided evidence that the Pro12Ala polymorphism is associated with obesity in non-diabetic males, but not with females (Ben Ali *et al.*, 2009). A study by Hsiao and Lin (2015), have also revealed a sex-bias with regards to the Pro12Ala SNP. In contrary to the findings by Ben Ali *et al.* (2009), this study indicated that the Pro12Ala polymorphism significantly predicts overweight, BMI, and total cholesterol in female but not male Taiwanese subjects (Hsiao and Lin, 2015). In the Spanish population female carriers of the Ala/Ala genotype were also proven to be more insulin sensitive and have better lipid profiles than the subjects with the Pro/Pro genotype. However this was not found in the male Spanish population (Gonzalez Sanchez *et al.*, 2002). Further investigation into this hypothesis is necessary to make accurate conclusions on the possibility that the Pro12Ala polymorphism is associated with a gender bias.

Nevertheless, the *PPARG-2* Pro12Ala polymorphism did not prove to be a good T2DM predictive marker in black South African females from this study. This highlights the need for large scale Genome Wide Association Studies (GWAS) on chronic diseases in Africa, since data from European and Asian populations are not always applicable on individuals from African ancestry.

3.3 Next Generation Sequencing

3.3.1 Data Analysis

Samples selected for NGS were randomly selected from the larger cohort (Appendix E). A total of 16 samples were selected, consisting of eight T2DM samples and 8 control samples. The T2DM patients were individually matched to a control according to age and BMI.

Approximately 2 GB of raw read data containing all five genes were received from the sequencing facility. A total of 2,7 million reads, with an average read length of 221 bp were collected from the 16 data files. Mapping of the sequence reads to the *PPARG* reference gene filtered out the sequences of interest as only reads that align to the *PPARG* reference will be mapped.

3.3.1.1 Data trimming

Automated analysis was performed with Torrent Suite™. All primers and adapters were trimmed off during this analysis. All data files were received from the University of Pretoria's Sequencing Facility as unmapped BAM files.

3.3.1.2 Quality Control (QC)

Downstream analysis is often compromised by low-quality sequences, sequence contamination and sequencing artefacts which lead to misassembled and inaccurate data (Schmieder and Edwards, 2011). The quality control of NGS datasets should include the investigation of sequence length, quality score, GC content and sequence complexity, artefacts, contamination and sequence duplication. The initial

processing should include trimming of sequence ends and filtering of unwanted sequences.

All FASTQ files were processed with prinseq-lite online to determine the quality of the data. Prinseq-graph produced readable graphs as PNG files which show the summary statistics that was produced. Examining the summary statistics is the simplest method to do QC on the data. These graphs were generated for each sample using PRINSEQ.

First the raw data was analysed with PRINSEQ without including any QC parameters, this was done to determine the overall quality of the sequence reads and to select the appropriate parameters for QC. The graphs and tables generated by prinseq-graph for the raw analysis was compared with those after proper QC was done with PRINSEQ. The different graphs produced for one sample, (Control 1), are outlined as an example to illustrate and describe the difference between the data before and after QC and can be viewed in Appendix G.

In this study effective QC of NGS data has been achieved using PRINSEQ by considering four basic areas namely: length distribution, base quality, sequence duplication, and the presence of tag sequences. After the raw data was evaluated it was decided to select for stringent QC parameters. This increased the overall quality of the data by trimming low quality reads and consequently the downstream analysis was more accurate. Trimming with PRINSEQ filtered out between 7% and 12% of data which did not meet the standard set by the stringent parameters.

The length distribution of the reads was used as a quality measure of the library preparation. The closer the read lengths are to a normal distribution the better the result (Prinseq.sourceforge.net, 2016). After QC the mean sequence length changed from 225 bp to 191 bp. The minimum length changed from 25 bp to 40 bp as was specified by the parameters. Trimming of the 3’ ends (Q-score below 30) have caused the maximum length to change from 524 bp to 452 bp. Length range and mode length was also affected by the QC. The shift in M, 1SD and 2SD can be observed when Figure 3.13 and 3.14 are compared.

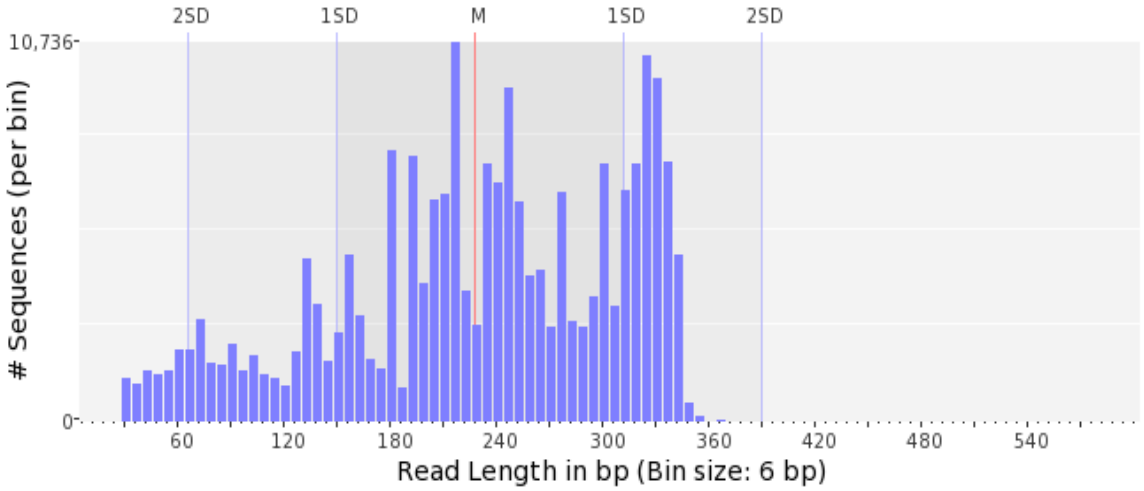


Figure 3.13: Length distribution in a graphical format before QC was performed.

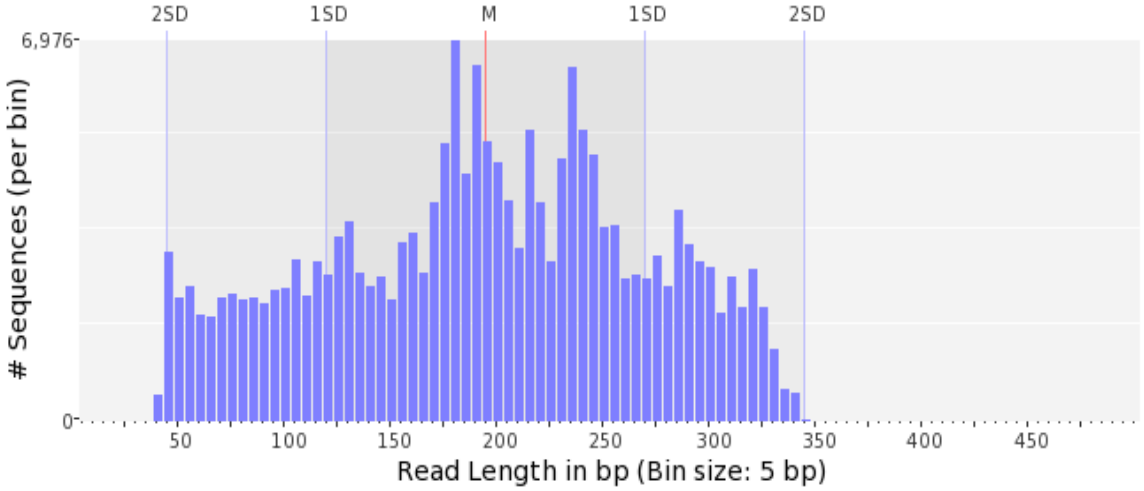


Figure 3.14: Length distribution in graphical format after QC was performed.

As with conventional Sanger sequencing, NGS data is also produced with a degradation of quality across the read (Bragg *et al.*, 2013). The quality scores for Ion Torrent PGM™ sequencers are Phred-based ranging from 0 to 40. Together with the decrease in quality across the read, regions with homopolymer stretches also tend to have lower quality scores (Bragg *et al.*, 2013).

In PRINSEQ, box plots map the quality scores across the reads. The x-axis indicates the absolute position if all reads are no longer than 100 bp and the relative position if any read is longer than 100 bp (Schmieder and Edwards, 2011). The data showed

improved quality scores across sequence reads after QC was performed. Before QC sequences with quality scores below 30 are present and after QC was done the sequences with low quality scores (below 30) at the ends were trimmed. The majority of the sequences have high mean quality score after QC was done.

The number of expected sequence duplicates depend on the type of library sequenced, the depth of the library, and the sequencing technology used. Exact duplicates are identical sequence copies, whereas 5' or 3' duplicates are sequences that are identical with the 5' or 3' end of a longer sequence (Prinseq.sourceforge.net, 2016). In this dataset the number of 5’ duplicates was higher than the number of 3’ duplicates due to the trimming at the 3’ end of sequence reads. The number of exact duplicates was higher than the number of 5' and 3' duplicates. Many exact duplicates were observed due to the use of AmpliSeq™ designed primers which amplified many of the same sequences compared to the usual fragmentation of input DNA.

Tag sequences are artefacts present at the ends of reads such as adapters and primer sequences (Schmieder and Edwards, 2011). Adapters were added to our sequence reads during the pre-amplification with the primer-based method AmpliSeq™. The equal distribution of the different nucleotides in the data indicate that no sequence tags were present. This was expected due to pre-analysis with Ion Torrent Suite™.

3.3.1.3 Mapping

Mapping was performed against five different reference sequences using Bowtie2. Firstly the data was mapped to Chr3. Mapping was done to Chr3 to get an overall view of all the data mapped to the chromosome. Chromosome mapping was also done to obtain the specific coordinates of each mutation according to the University of California, Santa Cruz (UCSC) Genome Browser. Thereafter mapping was performed against the entire *PPARG* gene (gi568815595) for closer investigation into the specific mutations found in *PPARG*. Mapping against the *PPARG* gene was performed to have a more manageable dataset to confirm that no mutations were missed by mapping against Chr3. It also enabled investigation into the 3’ and 5’ UTR’s and the 100 bp padding regions. Lastly, mapping was done to the ORF of the

three different splice isoforms of the *PPARG* gene (Isoforms 1: ENST00000397015. Isoform 2: ENST00000287820 and Isoform 3: ENST00000397015) to determine in which of the isoforms mutations were present.

Table 3.7: Tabulated is the different references used for mapping sequence reads using Bowtie2. Mapping was performed against Chr3, *PPARG* full gene sequence and the three different isoforms of the *PPARG* gene.

Reference	Database	Accession number
Chromosome 3		
PPARG gene	NCBI	gi568815595
Isoform 1	Ensembl	ENST00000397015
Isoform 2	Ensembl	ENST00000287820
Isoform 3	Ensembl	ENST00000397015

3.3.1.4 Variant calling

VarScan 2 calls somatic variants (SNPs and indels) using a heuristic method and a statistical test based on the number of aligned reads supporting each allele. Varscan 2 was chosen above other published variant callers, due to the robust heuristic approach which calls variants that meet the desired thresholds for read depth, base quality, variant allele frequency, and statistical significance. A study by Stead et al (2013) evaluated several somatic mutation callers including MuTect, Strelka, and VarScan 2. It was found VarScan 2 performed best overall with sequencing depths of 100x, 250x, 500x and 1000x required to accurately identify variants present at 10%, 5%, 2.5% and 1% respectively (Stead *et al.*, 2013).

VCF and CSV files were created for all 16 samples by Varscan 2. The VCF file format is used to store DNA polymorphism data as SNPs, indels and structural variants. VCF files must be indexed for viewing in IGV. VCF files were used during visualization in IGV for manual analysis, to visually confirm and curate each individual SNP. CSV files allow data to be saved in table structured format in plain text. CSV files can be opened with a spreadsheet programs, such as Microsoft Excel or Google Spreadsheets and produce a summary of all SNPs called. The CSV files

were used to briefly examine the variants found in each sample. All the variant calls had p-values <0.05, a minimum coverage of 33, a minimum Phred base quality of 20 and an allele frequency of at least 8% as was specified by the parameters when Varscan 2 was run. Examples of the visualization of the NGS data can be viewed in Appendix G.

3.3.2 Discussion on NGS data analysis

Mapping against Chr3 and against the whole *PPARG* gene revealed a total of 47 mutations, most of which are in the introns, the UTR (5'/3') or in the 100 bp padding region outside the *PPARG* gene at the 3' and 5' end. A full list of all mutations found including mutations in intronic regions and UTRs are available in Appendix H. These SNPs were not described or reported on because they did not specifically contribute to the research question of this study. However, three heterozygous mutations were found in the coding region (ORF) of *PPARG* in three different individuals. Mapping against the ORF of the isoforms revealed that the mutations were present in all three isoforms and all three SNPs are silent mutations.

The SNPs were present in two controls and one patient. The SNP found in the one control (C3) and the one patient (P4) were identical. This SNP is an A to G substitution on the third codon position of the Serine amino acid (rs41516544). The SNP found in the other control is a C to T substitution also on the third codon of the Histidine amino acid (rs3856806). All three mutations were present in exon 6 of *PPARG*. Exon 6 is one of the common exons shared between all three isoforms thus substantiating why the SNPs were present in all three isoforms. Table 3.8 contain all the information on the three synonymous SNPs.

Additionally, 28 mutations were detected in the intronic region of *PPARG* due to the 100 bp padding added to each exon. These mutations were present in both the patient and control cohorts and had high read depths of up to 2,000x. Three individuals in the control cohort showed an identical mutation in the 3'UTR region of *PPARG.* The read depth for these mutations ranged between 69 and 439x. Mutations were also identified outside the *PPARG* gene on the 3' and 5' ends. Nine of these mutations were identified throughout both cohort, but most had a relatively

low read depth of below 50x. Lastly four individuals, two patients and two controls, showed an identical mutation in the A1 untranslated exon present in *PPARG-1* and *PPARG-3* isoforms. These mutations had high read depths of 401, 196, 428 and 520 respectively. Only the mutations in the coding region of *PPARG* were described in this study.

3.3.2.1.1 Variant rs41516544

This variant was present in one control and one patient with a read depth of 561x and 1579x respectively. The mutation is located on Chr3 position 12475497 and is caused by a substitution of an A to a G nucleotide (TCA>TCG) in the protein coding region of *PPARG* (Table 3.8)**.** This is a synonymous mutation in the third position of the codon coding for the Serine amino acid. The variant has been described by the Ensembl database in The 1000 Genomes Project (Phase 3) (Genomes Project *et al.*, 2012). According to the population genetics on Ensembl, the ancestral base pair on this position is a G. The A bp is present in 99% of the total population and the G bp in 1% of the population shown in Figure 3.15. The frequency of the G allele is the highest in the African population and absent from the East Asian, European and South Asian populations. The G allele is present in 2% of the African population and the A bp in 98% of the African population. This SNP has no clinical significance and is possibly just a rare population variant.

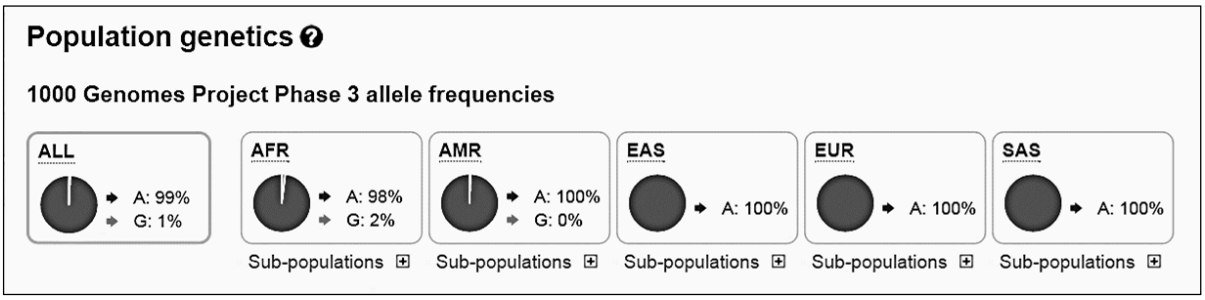


Figure 3.15: Population genetics for variant rs41516544. The G allele is present in the general population (ALL), but increases specifically in the African (AFR) population. The frequency of the G allele is also shown in the American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS) populations (Ensembl.org).

3.3.2.1.2 Variant rs3856806 (C1431T)

This variant was present in one control with a read depth of 400x. The SNP is caused by the substitution of a C with a T nucleotide (CAC>CAT) at position 12475557 on Chr3. This is a silent mutation in the third position of a Histidine amino acid. The mutation was also previously described on the Ensembl database. According to The 1000 Genomes Project (Phase 3) on Ensemble the ancestral allele is a C and is still

present in the largest part of the population, with a small percentage of people having the substituted T allele at that position indicated in Figure 3.16 (Genomes Project *et al.*, 2012).

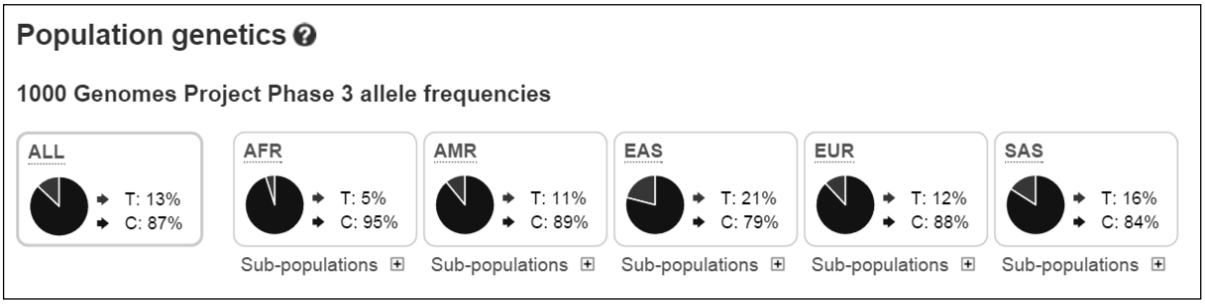


Figure 3.16: Population genetics for variant rs3856806. The frequency of the C allele is higher than that of the T allele in the general population (ALL). The frequency of the T allele in the African (AFR) population is the lowest when compared to the American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS) populations (Ensembl.org).

This SNP is a well described silent polymorphism in *PPARG* gene. The C1431T SNP has been associated with increased BMI and fat, increased leptin levels, increased resistin levels and a reduced risk of T2DM (Costa *et al.*, 2009; Haseeb *et al.*, 2009; Meirhaeghe *et al.*, 1998; Pattanayak *et al.*, 2014; Tai *et al.*, 2004; Valve *et al.*, 1999)**.** Many studies have shown that independently C1431T does not have a significant effect, but in combination with other variants such as Pro12Ala have a greater effect on obesity, an increase in fat mass, HDL cholesterol levels and decreased risk for T2DM (Bego *et al.*, 2011; Doney *et al.*, 2004; Liu *et al.*, 2015; Valve *et al.*, 1999).

Vergotine *et al.* (2014) identified the C1431T variant in 23.8% of a mixed ancestry population in South Africa, however the genotype distribution did not differ significantly between the diabetic and non-diabetic group. More recently, the polymorphism has been associated with increased risk of diabetic retinopathy in an Egyptian population (Ramadan *et al*., 2016).

The clinical significance of this SNP is benign. The condition associated with the SNP is Glioma susceptibility 1 (*GLM1*) or Glioblastoma (Zhou *et al.*, 2000). In a study by Zhou *et al.* (2000) they investigated the over-representation of the C1431T variant compared to the control cases, however, the exact mechanism of this association is still unknown. It was proposed that this SNP and the Pro12Ala variant may be acting in a low penetrance predisposing manner for the development of glioneural tumors. These associations were however only found in American patients with sporadic glioblastoma multiform and were not yet confirmed in other populations (Zhou *et al.*, 2000).

3.3.3 Conclusion

In-depth analysis of the NGS data revealed three synonymous polymorphisms in the *PPARG* coding region of three different individuals. Two of these variants were identical and were observed in a control and a patient, the other variant was detected in a control sample. The two variants identified were previously described and only one variant showed an association with T2DM in several populations, however, this was usually in combination with an additional variant such as Pro12Ala. This is most likely a common variant with a small effect on the population.

Table 3.8: Information on the three synonymous mutations found in the NGS data analysis. The table shows the position of the mutation on the chromosome as well as the position on the different isoforms. The reference and alternative base pair is shown and the mutation type, amino acid change and the codon change.

Participant	Position on chromosome	Isoform	Position on isoform	Exon	Reference	Alternate	Mutation type	AA change	AA	Codon change
C3	12475497	1	1287	Exon 6	A	G	Silent	No	Serine	TCA-TCG
		2	1371							
		3	1287							
C6	12475557	1	1347	Exon 6	C	T	Silent	No	Histidine	CAC-CAT
		2	1431							
		3	1347							
P4	12475497	1	1287	Exon 6	A	G	Silent	No	Serine	TCA-TCG
		2	1371							
		3	1287							

CHAPTER FOUR

CONCLUSION

Type 2 Diabetes Mellitus (T2DM) is a global epidemic that results in millions of deaths each year (Amod *et al.*, 2012). Obesity and diabetes is a growing problem in South Africa and the number of individuals with T2DM is increasing daily. Therefore, the identification of individuals that are at a high risk for developing T2DM is very important and of great interest to health care providers. Novel gene polymorphisms that are associated with chronic diseases of lifestyle should be identified especially in black African populations. This ethic group belongs to a haplogroup that harbours the most ancient human lineages and has the highest level of variation (Chen *et al.*, 1995). This study was significant because African populations have been neglected in research and by fine mapping casual variants within the black South African population this could be used to identify loci shared across ethnic groups globally. Various polymorphisms associated with T2DM have been identified and described in the *PPARG* gene (Barroso *et al.*, 1999; Ristow *et al.*, 1998; Valve *et al.*, 1999; Yen *et al.*, 1997).

The aim of this study was to screen the *PPARG* gene for novel T2DM genetic risk factors and to determine the presence of a previously identified T2DM genetic risk factor, the Pro12Ala variant (rs1801282), in black South-African women with T2DM. The prevalence of the *PPARG* Pro12Ala polymorphism has never been researched in a black, female population group in Mangaung, Free State. Subsequently, we hypothesized that these results would either confirm mutations associated with T2DM in other global populations or identify novel mutations associated with T2DM in this population.

Quantitative PCR was used to determine the presence of the *PPARG* Pro12Ala polymorphism in the black female South Africans with T2DM. After genotyping of 184 individuals which consisted of 93 T2DM patients and 91 control subjects, only a single individual (patient) presented with the heterozygote Pro/Ala genotype. The rest of the study population all contained the homozygous Pro/Pro genotype. The

presence of the homozygotic Ala/Ala genotype was not present in this black, female population in Mangaung in the Free State Province, South Africa. Although the results obtained from this study represent a small subset of the population, the allele frequency of the SNP suggests that it may not be a significant predictor of T2DM in the black South African population. It is very likely that the Ala/Ala genotype is rare in this population. Vergotine *et al*. (2014) found that the Ala/Ala allele was not significantly associated with T2DM in South African individuals with mixed ancestry. These findings are contrary to literature (Ghoussaini *et al.*, 2005; Sanghera *et al.*, 2010; Shahrjerdi *et al.*, 2013; Wang *et al.*, 2013) which have shown Pro12Ala to be significantly associated with development of T2DM, which makes this research significant. The controversial findings related to this polymorphism may be attributable to population differences that may be genetic. The *PPARG* Pro12Ala SNP would therefore not be a suitable biomarker for early risk prediction of T2DM in this population.

Next Generation Sequencing (Ion Torrent technology) was used to sequence the *PPARG* gene in eight patients and eight control subjects to identify novel population-specific polymorphisms in a black female South African population. A bioinformatic data analysis pipeline was successfully developed and applied to sequences from a case control T2DM study. This shows that the computational bioinformatic steps that is usually seen as one of the biggest complications with NGS, can successfully be applied by biologists. Data analysis on the sequences of the various samples revealed mutations in three individuals, one patient and two controls. Two individuals (a patient and a control) contained an identical variant (rs41516544) with a different variant present in the remaining control. Both these variants were synonymous variants in the coding region of *PPARG* that has already been described. The mutations did not show any clinical significance but one rs3856806 has been associated with T2DM. It has been suggested that this variant has a protective effect against T2DM but usually in combination with an additional variant (Bego *et al.*, 2011; Doney *et al.*, 2004; Liu *et al.*, 2015; Valve *et al.*, 1999).

To conclude we were unable to observe a significant correlation between the well-described *PPARG* Pro12Ala polymorphism and T2DM in our sample consisting of 184 black female participants. A single well-defined polymorphism was detected in

one of the 16 participants selected for NGS that relates specifically to T2DM, but no unique (novel) population-specific polymorphism were identified that might be associated with T2DM specifically in the black female population of South Africa

The strengths of this study is that one gene was investigated for a detailed representation of the polymorphisms and risk factors associated with diabetes found in this specific gene. One ethnic group controlled for BMI and age was investigated, giving specific information regarding the variants associated with that group. Fine mapping is usually published without control cohorts, the inclusion of age, BMI and HbA1c in the control group strengthens the association of mutations with T2DM. This was helpful as an association could be made between the genotyping results of the patient and control cohort. The NGS data could be compared between the patient and control subject as they were individually matching according to the specific criteria.

The limitation is that one gene is being investigated in a disease that is multifactorial and thus will not give a complete picture of risk factors associated with T2DM in a population. A relatively small sample size was selected for NGS due to the depth of sequencing, financial constraints and data volumes to analyse. Another limitation is that only females are included in the study, the addition of males could give a more comprehensive view of variants present in the whole population.

Further studies are required to confirm these observations. The prevalence of certain common variants associated with T2DM such as the Pro12Ala variant might be very different in the black South African population from Mangaung compared to the populations in Asia or Europe. This study support the broader thesis that the genetic background for the African population are very diverse and cannot be directly extrapolated using genetic variants from other ethnicities. It remains an important need for the identification of population-specific variants that are specifically linked to the black population in South Africa. The identification and description of a large number of novel genetic variants increasing susceptibility to T2DM will open up opportunities to translate this genetic information to the clinical practice and improve risk prediction.

SUMMARY

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder which is caused by a combination of an inadequate response to insulin secretion and a resistance to insulin action. The peroxisome proliferator-activated receptor gamma (*PPARG*) gene has been established as one of the major genes to have an impact on the risk of T2DM. The Pro12Ala polymorphism is one of the most common mutations found within *PPARG* and has been described in many different populations.

However, it has not yet been established whether the Pro12Ala variant has a significant association with T2DM in the black, female South African population. The aim of this study was to screen for novel T2DM genetic risk factors in the *PPARG* gene and to determine the presence of previously identified T2DM genetic risk factors in black South-African women with T2DM. Quantitative PCR was performed on 184 black female South African participants that consisted of 93 patients diagnosed with T2DM and 91 control participants. Quantitative PCR was used to screen for the presence of the Pro12Ala polymorphism in the *PPARG* gene. Next Generation Sequencing (NGS) was performed on eight patients and eight control samples which were individually matched according to age and body mass index (BMI). NGS was used to identify novel polymorphisms which might be associated with T2DM and to detect the prevalence of previously described variants within the *PPARG* gene.

The qPCR genotyping results showed that of the 184 participants, 183 had the Pro/Pro genotype and only one had the heterozygous Pro/Ala genotype. The Ala/Ala genotype was not detected in this study population. Although the study sample is only a small representation of the total population, it can be derived from the results that it is likely that the Ala/Ala genotype is rare in the population.

Additionally, NGS results identified two variants within three individuals of the selected sample. The one variant (rs41516544) did not show any clinical relevance and is probably just a rare population variant. The other variant (rs3856806) is a well-described polymorphism and has been associated with having a protective

effect against T2DM and was present in a control participant. This variant might be significant in its association to T2DM in the black South African population but will have to be further investigated in future studies.

Keywords: AmpliSeq™; Ion Torrent PGM™; Next generation sequencing; NGS data-analysis; Peroxisome proliferator-activated receptor gamma (*PPARG*); Pro12Ala; qPCR; Type 2 diabetes mellitus.

OPSOMMING

Tipe 2 diabetes mellitus (T2DM) is 'n kroniese metaboliese afwyking wat veroorsaak word deur 'n kombinasie van 'n onvoldoende reaksie op insulien afskeiding en 'n weerstand teen insulien aksie. Die peroxisoom proliferator-geaktiveerde reseptor-gamma 2 (*PPARG-2*) geen word beskou as een van die belangrikste gene wat 'n impak op die risiko van T2DM het. Die Pro12Ala polimorfisme is een van die mees algemene mutasies wat in die *PPARG* geen aangetref word en is al beskryf in verskillende bevolkingsgroepe.

Dit is nog nie vasgestel of die Pro12Ala variant 'n beduidende assosiasie met T2DM in die swart, vroulike populasie van Suid-Afrika het nie. Die doel van hierdie studie was om te toets vir nuwe T2DM genetiese risikofaktore in die *PPARG* geen en om die teenwoordigheid van voorheen geïdentifiseerde T2DM genetiese risikofaktore in swart Suid-Afrikaanse vroue met T2DM te bepaal. Reële-tyd polimerase ketting reaksie (Reële-tyd PKR) is uitgevoer op 184 swart vroulike Suid-Afrikaanse deelnemers wat bestaan uit 93 individue wat met T2DM gediagnoseer is en 91 individue wat dien as kontroles. Reële-tyd PKR is gebruik om te toets vir die teenwoordigheid van die Pro12Ala polimorfisme in die *PPARG* gene. Volgende Generasie Volgordebepaling (VGV) is uitgevoer op agt pasiënte en agt kontrole individue wat ooreengestem het in terme van ouderdom en liggaamsmassa-indeks (BMI). VGV is gebruik om nuwe polimorfismes te identifiseer wat moontlik geassosieer kan word met T2DM en om die voorkoms van variante wat voorheen beskryf is in *PPARG* te bepaal.

Die Reële-tyd PKR genotipering resultate het getoon dat van die 184 deelnemers 183 die Pro/Pro genotype vertoon en net een die heterosigotiese Pro/Ala genotype het. Die Ala/Ala genotipe is nie teenwoordig in hierdie studiepopulasie nie. Alhoewel die studiepopulasie slegs 'n klein verteenwoordiging van die totale bevolking is dui die resultate daarop dat die Ala/Ala genotipe waarskynlik baie skaars in die bevolking is.

VGV resultate het 2 variante geïdentifiseer in drie individue van die studiegroep. Die een variant (rs41516544) het geen kliniese relevansie nie en is waarskynlik net 'n seldsame bevolkings variant. Die ander variant (rs3856806) is 'n goed beskryfde polimorfisme wat geassosieer is met 'n beskermende effek teen T2DM. Hierdie variant kan moontlik betekenisvol wees in die swart Suid-Afrikaanse bevolking, maar sal eers verder ondersoek moet word deur toekomstige studies.

Kernwoorde: AmpliSeq™; Ion Torrent PGM™; Peroxisoom proliferator-geaktiveerde reseptor-gamma (*PPARG*); Pro12Ala; Reële-tyd PKR; Tipe 2 diabetes mellitus; Volgende generasie sequentiëring; VGV data analise.

REFERENCES

Abate, N. & Chandalia, M. 2003. The impact of ethnicity on type 2 diabetes. *J Diabetes Complications,* 17**,** 39-58.

Adamo, K. B., Sigal, R. J., Williams, K., Kenny, G., Prud'homme, D. & Tesson, F. 2005. Influence of Pro12Ala peroxisome proliferator-activated receptor gamma2 polymorphism on glucose response to exercise training in type 2 diabetes. *Diabetologia,* 48**,** 1503-9.

Afridi, M. A. & Khan, M. N. 2003. Role of health education in the management of diabetes mellitus. *J Coll Physicians Surg Pak,* 13**,** 558-61.

Al-Safar, H., Hassoun, A., Almazrouei, S., Kamal, W., Afandi, B. & Rais, N. 2015. Association of the Genetic Polymorphisms in Transcription Factor 7-Like 2 and Peroxisome Proliferator-Activated Receptors- gamma 2 with Type 2 Diabetes Mellitus and is Interaction with Obesity Status in Emirati Population. *J Diabetes Res,* 2015**,** 129695.

Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., Richmond, T. A., Middle, C. M., Rodesch, M. J., Packard, C. J., Weinstock, G. M. & Gibbs, R. A. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods,* 4**,** 903-5.

Alberti, K. G. & Zimmet, P. Z. 1998. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabet Med,* 15**,** 539-53.

Ali, O. 2013. Genetics of type 2 diabetes. *World J Diabetes,* 4**,** 114-23.

Alker, A. P., Mwapasa, V. & Meshnick, S. R. 2004. Rapid real-time PCR genotyping of mutations associated with sulfadoxine-pyrimethamine resistance in Plasmodium falciparum. *Antimicrob Agents Chemother,* 48**,** 2924-9.

Almgren, P., Lehtovirta, M., Isomaa, B., Sarelin, L., Taskinen, M. R., Lyssenko, V., Tuomi, T. & Groop, L. 2011. Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia,* 54**,** 2811-9.

Altshuler, D., Hirschhorn, J. N., Klannemark, M., Lindgren, C. M., Vohl, M. C., Nemesh, J., Lane, C. R., Schaffner, S. F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T. J., Daly, M., Groop, L. & Lander, E. S. 2000. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet,* 26**,** 76-80.

American Diabetes Association 2008. Diagnosis and classification of diabetes mellitus. *Diabetes Care,* 31 Suppl 1**,** S55-60.

American Diabetes Association 2010. Diagnosis and classification of diabetes mellitus. *Diabetes Care,* 33 Suppl 1**,** S62-9.

American Diabetes Association 2015. (2) Classification and diagnosis of diabetes. *Diabetes Care,* 38 Suppl**,** S8-S16.

Amod, A., Motala A, Levitt N, Berg J, Young M, Grobler N, Heilbrunn A, Distiller L, Pirie F, Dave J, Huddle K, Jivan D, Paruk I, May W, Raal D, Blom D, Ascott-Evans B, Brown S, Mollentze W, Rheeder P, Tudhope L, Van Rensburgh G, Ganie Y, Carrihill M, Rauff S, Van Zyl D, Randeree H, Khutsoane D, Joshi P, Raubenheimer P & Guideline Committee 2012. The 2012 SEMDSA Guideline for the Management of Type 2 Diabetes. *Journal of Endocrinology, Metabolism and Diabetes of South Africa,* 17(1)**,** S1-S94.

Anderson, D., Cordell, H. J., Fakiola, M., Francis, R. W., Syn, G., Scaman, E. S., Davis, E., Miles, S. J., McLeay, T., Jamieson, S. E. & Blackwell, J. M. 2015. First genome-wide association study in an Australian aboriginal population provides insights into genetic risk factors for body mass index and type 2 diabetes. *PLoS One,* 10**,** e0119333.

Ansorge, W., Sproat, B., Stegemann, J., Schwager, C. & Zenke, M. 1987. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res,* 15**,** 4593-602.

Badii, R., Bener, A., Zirie, M., Al-Rikabi, A., Simsek, M., Al-Hamaq, A. O., Ghoussaini, M., Froguel, P. & Wareham, N. J. 2008. Lack of association between the Pro12Ala polymorphism of the PPAR-gamma 2 gene and type 2 diabetes mellitus in the Qatari consanguineous population. *Acta Diabetol,* 45**,** 15-21.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. & Marth, G. T. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics,* 27**,** 1691-2.

Barr, R. G., Nathan, D. M., Meigs, J. B. & Singer, D. E. 2002. Tests of glycemia for the diagnosis of type 2 diabetes mellitus. *Ann Intern Med,* 137**,** 263-72.

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D. J., Stevens, H., Todd, J. A., Walker, N. M., Rich, S. S. & Type 1 Diabetes Genetics, C. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet,* 41**,** 703-7.

Barroso, I. 2005. Genetics of Type 2 diabetes. *Diabet Med,* 22**,** 517-35.

Barroso, I., Gurnell, M., Crowley, V. E., Agostini, M., Schwabe, J. W., Soos, M. A., Maslen, G. L., Williams, T. D., Lewis, H., Schafer, A. J., Chatterjee, V. K. & O'Rahilly, S. 1999. Dominant negative mutations in human PPARgamma associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature,* 402**,** 880-3.

Bego, T., Dujic, T., Mlinar, B., Semiz, S., Malenica, M., Prnjavorac, B., Ostanek, B., Marc, J. & Causevic, A. 2011. Association of PPARG and LPIN1 gene polymorphisms with metabolic syndrome and type 2 diabetes. *Med Glas (Zenica),* 8**,** 76-83.

BeLue, R., Okoror, T. A., Iwelunmor, J., Taylor, K. D., Degboe, A. N., Agyemang, C. & Ogedegbe, G. 2009. An overview of cardiovascular risk factor burden in sub-Saharan African countries: a socio-cultural perspective. *Global Health,* 5**,** 10.

Ben Ali, S., Ben Yahia, F., Sediri, Y., Kallel, A., Ftouhi, B., Feki, M., Elasmi, M., Haj-Taieb, S., Souheil, O., Sanhagi, H., Slimane, H., Jemaa, R. & Kaabachi, N. 2009. Gender-specific effect of Pro12Ala polymorphism in peroxisome proliferator-activated receptor gamma-2 gene on obesity risk and leptin levels in a Tunisian population. *Clin Biochem,* 42**,** 1642-7.

Bener, A., Zirie, M., Al-Hamaq, A., Nawaz, Z., Samson, N. & Mohammad, R. 2015. Impact of the Pro12Ala polymorphism of the PPARgamma2 gene on diabetes and obesity in a highly consanguineous population. *Indian J Endocrinol Metab,* 19**,** 77-83.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara, E. C. M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K.,

Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M*., et al.* 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456**,** 53-9.

Bergholdt, R., Brorsson, C., Palleja, A., Berchtold, L. A., Floyel, T., Bang-Berthelsen, C. H., Frederiksen, K. S., Jensen, L. J., Storling, J. & Pociot, F. 2012. Identification of novel type 1 diabetes candidate genes by integrating genome-wide association data, protein-protein interactions, and human pancreatic islet gene expression. *Diabetes,* 61**,** 954-62.

Bhuvaneshwar, K., Sulakhe, D., Gauba, R., Rodriguez, A., Madduri, R., Dave, U., Lacinski, L., Foster, I., Gusev, Y. & Madhavan, S. 2015. A case study for cloud based high throughput analysis of NGS data using the globus genomics system. *Computational and Structural Biotechnology Journal,* 13**,** 64-74.

Bieche, I., Olivi, M., Champeme, M. H., Vidaud, D., Lidereau, R. & Vidaud, M. 1998. Novel approach to quantitative polymerase chain reaction using real-time detection: application to the detection of gene amplification in breast cancer. *Int J Cancer,* 78**,** 661-6.

Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. & Tyson, G. W. 2013. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Computational Biology,* 9**,** e1003031.

Braissant, O., Foufelle, F., Scotto, C., Dauca, M. & Wahli, W. 1996. Differential expression of peroxisome proliferator-activated receptors (PPARs): tissue distribution of PPAR-alpha, -beta, and -gamma in the adult rat. *Endocrinology,* 137**,** 354-66.

Bustin, S. A. 2005. Real-Time PCR. *In:* FUCHS, J. & PODDA, M. (eds.) *Encyclopedia of Diagnostic Genomics and Proteonomics.* Marcel Dekker.

Bustin, S.A.,Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J. & Wittwer, C. T. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem*, 55, 611-22.

Camerlengo, T., Ozer, H. G., Onti-Srinivasan, R., Yan, P., Huang, T., Parvin, J. & Huang, K. 2012. From sequencer to supercomputer: an automatic pipeline for managing and processing next generation sequencing data. *AMIA Jt Summits Transl Sci Proc,* 2012**,** 1-10.

Camp, H. S. & Tafuri, S. R. 1997. Regulation of peroxisome proliferator-activated receptor γ activity by mitogen-activated protein kinase. *Journal of Biological Chemistry,* 272**,** 10811-10816.

Capaccio, D., Ciccodicola, A., Sabatino, L., Casamassimi, A., Pancione, M., Fucci, A., Febbraro, A., Merlino, A., Graziano, G. & Colantuoni, V. 2010. A novel germline mutation in peroxisome proliferator-activated receptor gamma gene associated with large intestine polyp formation and dyslipidemia. *Biochim Biophys Acta,* 1802**,** 572-81.

Chen, Y., Kittles, R., Zhou, J., Chen, G., Adeyemo, A., Panguluri, R. K., Chen, W., Amoah, A., Opoku, V., Acheampong, J., Agyenim-Boateng, K., Eghan, B. A., Jr., Nyantaki, A., Oli, J., Okafor, G., Ofoegbu, E., Osotimehin, B., Abbiyesuku, F., Johnson, T., Fasanmade, O., Rufus, T., Furbert-Harris, P., Daniel, H. I., Berg, K. A., Collins, F. S., Dunston, G. M. & Rotimi, C. N. 2005. Calpain-10 gene polymorphisms and type 2 diabetes in West Africans: the Africa America Diabetes Mellitus (AADM) Study. *Ann Epidemiol,* 15**,** 153-9.

Chen, Y. S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S. & Wallace, D. C. 1995. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet,* 57**,** 133-49.

Cheng, J., Zhang, Y. & Li, Q. 2004. Real-time PCR genotyping using displacing probes. *Nucleic Acids Res,* 32**,** e61.

Chiarelli, F. & Marzio, D. D. 2008. Peroxisome proliferator-activated receptor-γ agonists and diabetes: Current evidence and future perspectives. *Vascular Health and Risk Management,* 4**,** 297-304.

Chikowore, T., Conradie, K. R., Towers, G. W. & van Zyl, T. 2015. Common Variants Associated with Type 2 Diabetes in a Black South African Population of Setswana Descent: African Populations Diverge. *OMICS*, 19(10):617-26.

Chlebowy, D. O., Hood, S. & LaJoie, A. S. 2013. Gender differences in diabetes self-management among African American adults. *West J Nurs Res,* 35**,** 703-21.

Clausen, J. O., Hansen, T., Bjorbaek, C., Echwald, S. M., Urhammer, S. A., Rasmussen, S., Andersen, C. B., Hansen, L., Almind, K., Winther, K. & et al. 1995. Insulin resistance: interactions between obesity and a common variant of insulin receptor substrate-1. *Lancet,* 346**,** 397-402.

Collins, F. S., Morgan, M. & Patrinos, A. 2003. The Human Genome Project: lessons from large-scale biology. *Science,* 300**,** 286-90.

Conn, J. W. 1940. The necessity of a standard preparatory diet. *The American Journal of the Medical Sciences,* 199**,** 555-563.

Cooke, J. N., Ng, M. C., Palmer, N. D., An, S. S., Hester, J. M., Freedman, B. I., Langefeld, C. D. & Bowden, D. W. 2012. Genetic risk assessment of type 2 diabetes-associated polymorphisms in African Americans. *Diabetes Care,* 35**,** 287-92.

Cornell University, (2007). *Cornell Statistical Consulting Unit*. [online] Available at: https://www.cscu.cornell.edu/news/statnews/stnews71.pdf [Accessed 24 May 2016].

Costa, V., Casamassimi, A., Esposito, K., Villani, A., Capone, M., Iannella, R., Schisano, B., Ciotola, M., Di Palo, C., Corrado, F. C., Santangelo, F., Giugliano, D. & Ciccodicola, A. 2009. Characterization of a novel polymorphism in PPARG regulatory region associated with type 2 diabetes and diabetic retinopathy in Italy. *J Biomed Biotechnol,* 2009**,** 126917.

Costa, V., Gallo, M. A., Letizia, F., Aprile, M., Casamassimi, A. & Ciccodicola, A. 2010. PPARG: Gene Expression Regulation and Next-Generation Sequencing for Unsolved Issues. *PPAR Res,* 2010.

Cowie, C. C., Rust, K. F., Byrd-Holt, D. D., Gregg, E. W., Ford, E. S., Geiss, L. S., Bainbridge, K. E. & Fradkin, J. E. 2010. Prevalence of diabetes and high risk for diabetes using A1C criteria in the U.S. population in 1988-2006. *Diabetes Care,* 33**,** 562-8.

Danaei, G., Finucane, M. M., Lu, Y., Singh, G. M., Cowan, M. J., Paciorek, C. J., Lin, J. K., Farzadfar, F., Khang, Y. H., Stevens, G. A., Rao, M., Ali, M. K., Riley, L. M., Robinson, C. A., Ezzati, M. & Global Burden of Metabolic Risk Factors of Chronic Diseases Collaborating, G. 2011. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. *Lancet,* 378**,** 31-40.

Day, C. P., Grove, J., Daly, A. K., Stewart, M. W., Avery, P. J. & Walker, M. 1998. Tumour necrosis factor-alpha gene promoter polymorphism and decreased insulin resistance. *Diabetologia,* 41**,** 430-4.

de Vegt, F., Dekker, J. M., Jager, A., Hienkens, E., Kostense, P. J., Stehouwer, C. D., Nijpels, G., Bouter, L. M. & Heine, R. J. 2001. Relation of impaired fasting and postload glucose with incident type 2 diabetes in a Dutch population: The Hoorn Study. *JAMA,* 285**,** 2109-13.

DeBruyne, L., Pinna, K. & Whitney, E. 2011. *Nutrition and Diet Therapy*, Cengage Learning.

Deeb, S. S., Fajas, L., Nemoto, M., Pihlajamaki, J., Mykkanen, L., Kuusisto, J., Laakso, M., Fujimoto, W. & Auwerx, J. 1998. A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nat Genet,* 20**,** 284-7.

Doney, A. S., Fischer, B., Cecil, J. E., Boylan, K., McGuigan, F. E., Ralston, S. H., Morris, A. D. & Palmer, C. N. 2004. Association of the Pro12Ala and C1431T variants of PPARG and their haplotypes with susceptibility to Type 2 diabetes. *Diabetologia,* 47**,** 555-8.

Dugas, L. R., Carstens, M. A., Ebersole, K., Schoeller, D. A., Durazo-Arvizu, R. A., Lambert, E. V. & Luke, A. 2009. Energy expenditure in young adult urban informal settlement dwellers in South Africa. *Eur J Clin Nutr,* 63**,** 805-807.

Eisenbarth, G. S. 1986. Type I diabetes mellitus. A chronic autoimmune disease. *N Engl J Med,* 314**,** 1360-8.

Ek, J., Andersen, G., Urhammer, S. A., Hansen, L., Carstensen, B., Borch-Johnsen, K., Drivsholm, T., Berglund, L., Hansen, T., Lithell, H. & Pedersen, O. 2001. Studies of the Pro12Ala polymorphism of the peroxisome proliferator-activated receptor-gamma2 (PPAR-gamma2) gene in relation to insulin sensitivity among glucose tolerant caucasians. *Diabetologia,* 44**,** 1170-6.

Ensembl.org,. "Rs1801282 (SNP) - Population Genetics - Homo Sapiens - Ensembl Genome Browser 83". N.p., 2015 [Accessed 18 Oct. 2015].

Ensembl.org,. "Rs3856806 (SNP) - Population Genetics - Homo Sapiens - Ensembl Genome Browser 83". N.p., 2015 [Accessed 26 Oct. 2015].

Ensembl.org,. "Rs41516544 (SNP) - Population Genetics - Homo Sapiens - Ensembl Genome Browser 83". N.p., 2015 [Accessed 23 Oct. 2015].

Eshel, R., Vainas, O., Shpringer, M. & Naparstek, E. 2006. Highly sensitive patient-specific real-time PCR SNP assay for chimerism monitoring after allogeneic stem cell transplantation. *Lab Hematol,* 12**,** 39-46.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res,* 8**,** 175-85.

Fajas, L., Auboeuf, D., Raspe, E., Schoonjans, K., Lefebvre, A. M., Saladin, R., Najib, J., Laville, M., Fruchart, J. C., Deeb, S., Vidal-Puig, A., Flier, J., Briggs, M. R., Staels, B., Vidal, H. & Auwerx, J. 1997. The organization, promoter analysis, and expression of the human PPARgamma gene. *J Biol Chem,* 272**,** 18779-89.

Fajas, L., Fruchart, J. C. & Auwerx, J. 1998. PPARgamma3 mRNA: a distinct PPARgamma mRNA subtype transcribed from an independent promoter. *FEBS Lett,* 438**,** 55-60.

Franks, P. W., Luan, J., Browne, P. O., Harding, A. H., O'Rahilly, S., Chatterjee, V. K. & Wareham, N. J. 2004. Does peroxisome proliferator-activated receptor gamma genotype (Pro12ala) modify the association of physical activity and dietary fat with fasting insulin level? *Metabolism,* 53**,** 11-6.

Gale, E. A. & Gillespie, K. M. 2001. Diabetes and gender. *Diabetologia,* 44**,** 3-15.

Genomes Project, C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. & McVean, G. A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

Ghoussaini, M., Meyre, D., Lobbens, S., Charpentier, G., Clement, K., Charles, M. A., Tauber, M., Weill, J. & Froguel, P. 2005. Implication of the Pro12Ala polymorphism of the PPAR-gamma 2 gene in type 2 diabetes and obesity in the French population. *BMC Med Genet,* 6**,** 11.

Gibson, N. J. 2006. The use of real-time PCR methods in DNA sequence variation analysis. *Clin Chim Acta,* 363**,** 32-47.

Gibson, U. E., Heid, C. A. & Williams, P. M. 1996. A novel method for real time quantitative RT-PCR. *Genome Res,* 6**,** 995-1001.

Gill, G. V., Mbanya, J. C., Ramaiya, K. L. & Tesfaye, S. 2009. A sub-Saharan African perspective of diabetes. *Diabetologia,* 52**,** 8-16.

Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T. & Martin, J. F. 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics,* 12**,** 245.

Gogol-Doring, A. & Chen, W. 2012. An overview of the analysis of next generation sequencing data. *Methods Mol Biol,* 802**,** 249-57.

Goldstein, D. E., Parker, K. M., England, J. D., England, J. E., Wiedmeyer, H.-M., Rawlings, S. S., Hess, R., Little, R. R., Simonds, J. F. & Breyfogle, R. P. 1982. Clinical Application of Glycosylated Hemoglobin Measurements. *Diabetes,* 31**,** 70-78.

Gonzalez Sanchez, J. L., Serrano Rios, M., Fernandez Perez, C., Laakso, M. & Martinez Larrad, M. T. 2002. Effect of the Pro12Ala polymorphism of the peroxisome proliferator-activated receptor gamma-2 gene on adiposity, insulin sensitivity and lipid profile in the Spanish population. *Eur J Endocrinol,* 147**,** 495-501.

Gouda, H. N., Sagoo, G. S., Harding, A. H., Yates, J., Sandhu, M. S. & Higgins, J. P. 2010. The association between the peroxisome proliferator-activated receptor-gamma2 (PPARG2) Pro12Ala gene variant and type 2 diabetes mellitus: a HuGE review and meta-analysis. *Am J Epidemiol,* 171**,** 645-55.

Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., Styrkarsdottir, U., Magnusson, K. P., Walters, G. B., Palsdottir, E., Jonsdottir, T., Gudmundsdottir, T., Gylfason, A., Saemundsdottir, J., Wilensky, R. L., Reilly, M. P., Rader, D. J., Bagger, Y., Christiansen, C., Gudnason, V., Sigurdsson, G., Thorsteinsdottir, U., Gulcher, J. R., Kong, A. & Stefansson, K. 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet,* 38**,** 320-3.

Groop, L. & Pociot, F. 2014. Genetics of diabetes--are we missing the genes or the disease? *Mol Cell Endocrinol,* 382**,** 726-39.

Guo, W. L., Tang, Y., Han, X. Y. & Ji, L. N. 2011. [Meta-analysis of the association of Pro12Ala polymorphism of peroxisome proliferator activated receptor gamma gene with type 2 diabetes in Chinese Han population]. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao,* 33**,** 593-9.

Hara, K., Fujita, H., Johnson, T. A., Yamauchi, T., Yasuda, K., Horikoshi, M., Peng, C., Hu, C., Ma, R. C., Imamura, M., Iwata, M., Tsunoda, T., Morizono, T., Shojima, N., So, W. Y., Leung, T. F., Kwan, P., Zhang, R., Wang, J., Yu, W., Maegawa, H., Hirose, H., consortium, D., Kaku, K., Ito, C., Watada, H., Tanaka, Y., Tobe, K., Kashiwagi, A., Kawamori, R., Jia, W., Chan, J. C., Teo, Y. Y., Shyong, T. E., Kamatani, N., Kubo, M., Maeda, S. & Kadowaki, T. 2014. Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum Mol Genet,* 23**,** 239-46.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S. & Frazer, K. A. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol,* 10**,** R32.

Haseeb, A., Iliyas, M., Chakrabarti, S., Farooqui, A. A., Naik, S. R., Ghosh, S., Suragani, M. & Ehtesham, N. Z. 2009. Single-nucleotide polymorphisms in peroxisome proliferator-activated receptor gamma and their association with plasma levels of resistin and the metabolic syndrome in a South Indian population. *J Biosci,* 34**,** 405-14.

Hatem, A., Bozdağ, D., Toland, A. E. & Çatalyürek, Ü. V. 2013. Benchmarking short sequence mapping tools. *BMC Bioinformatics,* 14**,** 1-25.

Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. 1996. Real time quantitative PCR. *Genome Res,* 6**,** 986-94.

Henkin, L., Bergman, R. N., Bowden, D. W., Ellsworth, D. L., Haffner, S. M., Langefeld, C. D., Mitchell, B. D., Norris, J. M., Rewers, M., Saad, M. F., Stamm, E., Wagenknecht, L. E. & Rich, S. S. 2003. Genetic epidemiology of insulin resistance and visceral adiposity. The IRAS Family Study design and methods. *Ann Epidemiol,* 13**,** 211-7.

Higuchi, R., Dollinger, G., Walsh, P. S. & Griffith, R. 1992. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology (N Y),* 10**,** 413-7.

Higuchi, R., Fockler, C., Dollinger, G. & Watson, R. 1993. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology (N Y),* 11**,** 1026-30.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. & Manolio, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A,* 106**,** 9362-7.

Hirschhorn, J. N. 2003. Genetic epidemiology of type 1 diabetes. *Pediatr Diabetes,* 4**,** 87-100.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J. & McCombie, W. R. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet,* 39**,** 1522-7.

Holland, P. M., Abramson, R. D., Watson, R. & Gelfand, D. H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proceedings of the National Academy of Sciences,* 88**,** 7276-7280.

Hong, E. P. & Park, J. W. 2012. Sample size and statistical power calculation in genetic association studies. *Genomics Inform,* 10**,** 117-22.

Horner, D. S., Pavesi, G., Castrignanò, T., De Meo, P. D. O., Liuni, S., Sammeth, M., Picardi, E. & Pesole, G. 2010. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics,* 11**,** 181-197.

Hsiao, T. J. & Lin, E. 2015. The Pro12Ala polymorphism in the peroxisome proliferator-activated receptor gamma (PPARG) gene in relation to obesity and metabolic phenotypes in a Taiwanese population. *Endocrine,* 48**,** 786-93.

Hu, C., Zhang, R., Wang, C., Wang, J., Ma, X., Lu, J., Qin, W., Hou, X., Wang, C., Bao, Y., Xiang, K. & Jia, W. 2009. PPARG, KCNJ11, CDKAL1, CDKN2A-CDKN2B, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 are associated with type 2 diabetes in a Chinese population. *PLoS One,* 4**,** e7643.

Hummasti, S. & Tontonoz, P. 2006. The peroxisome proliferator-activated receptor N-terminal domain controls isotype-selective gene expression and adipogenesis. *Mol Endocrinol,* 20**,** 1261-75.

Hypponen, E., Virtanen, S. M., Kenward, M. G., Knip, M., Akerblom, H. K. & Childhood Diabetes in Finland Study, G. 2000. Obesity, increased linear growth, and risk of type 1 diabetes in children. *Diabetes Care,* 23**,** 1755-60.

IDT. SciTools OligoAnalyzer. 2009; Accessed 25/08/2015. Available from: http://www.idtdna.com/analyzer/Applications/OligoAnalyzer/.

Imperato, P. J. & Imperato, G. H. 2009. The role of genealogy and clinical family histories in documenting possible inheritance patterns for diabetes mellitus in the pre-insulin era: part 1. The clinical case of Josephine Imperato. *J Community Health,* 34**,** 400-18.

International Diabetes Federation 2014. IDF Diabetes Atlas, 6th ed. *Brussels, Belgium.* International Diabetes Federation, 2014.

International Expert, C. 2009. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes Care,* 32**,** 1327-34.

In-silico: *Project support for life sciences,* 2016. Available from: <http://in-silico.net/tools/statistics/fisher_exact_test>. [10 January 2016].

Israelian-Konaraki, Z. & Reaven, P. D. 2005. Peroxisome proliferator-activated receptor-alpha and atherosclerosis: from basic mechanisms to clinical implications. *Cardiol Rev,* 13**,** 240-6.

Johansson, C., Samuelsson, U. & Ludvigsson, J. 1994. A high weight gain early in life is associated with an increased risk of type 1 (insulin-dependent) diabetes mellitus. *Diabetologia,* 37**,** 91-4.

Kahara, T., Takamura, T., Hayakawa, T., Nagai, Y., Yamaguchi, H., Katsuki, T., Katsuki, K., Katsuki, M. & Kobayashi, K. 2003. PPARgamma gene polymorphism is associated with exercise-mediated changes of insulin resistance in healthy men. *Metabolism,* 52**,** 209-12.

Kamadjeu, R. M., Edwards, R., Atanga, J. S., Kiawi, E. C., Unwin, N. & Mbanya, J. C. 2006. Anthropometry measures and prevalence of obesity in the urban adult population of Cameroon: an update from the Cameroon Burden of Diabetes Baseline Survey. *BMC Public Health,* 6**,** 228.

Kao, W. H. L., Coresh, J., Shuldiner, A. R., Boerwinkle, E., Bray, M. S. & Brancati, F. L. 2003. Pro12Ala of the Peroxisome Proliferator–Activated Receptor-γ2 Gene Is Associated With Lower Serum Insulin Levels in Nonobese African Americans: The Atherosclerosis Risk in Communities Study. *Diabetes,* 52**,** 1568-1572.

Kaprio, J., Tuomilehto, J., Koskenvuo, M., Romanov, K., Reunanen, A., Eriksson, J., Stengard, J. & Kesaniemi, Y. A. 1992. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia,* 35**,** 1060-7.

Kharroubi, A. T. & Darwish, H. M. 2015. Diabetes mellitus: The epidemic of the century. *World J Diabetes,* 6**,** 850-67.

Kindich, R., Florl, A. R., Jung, V., Engers, R., Muller, M., Schulz, W. A. & Wullich, B. 2005. Application of a modified real-time PCR technique for relative gene copy number quantification to the determination of the relationship between NKX3.1 loss and MYC gain in prostate cancer. *Clin Chem,* 51**,** 649-52.

King, H. & Rewers, M. 1993. Global estimates for prevalence of diabetes mellitus and impaired glucose tolerance in adults. WHO Ad Hoc Diabetes Reporting Group. *Diabetes Care,* 16**,** 157-77.

Kirigia, J. M., Sambo, H. B., Sambo, L. G. & Barry, S. P. 2009. Economic burden of diabetes mellitus in the WHO African region. *BMC Int Health Hum Rights,* 9**,** 6.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L. & Wilson, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res,* 22**,** 568-76.

Koch, W. H. 2004. Technology platforms for pharmacogenomic diagnostic assays. *Nat Rev Drug Discov,* 3**,** 749-761.

Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K. T., Jonasdottir, A., Frigge, M. L., Gylfason, A., Olason, P. I., Gudjonsson, S. A., Sverrisson, S., Stacey, S. N., Sigurgeirsson, B., Benediktsdottir, K. R., Sigurdsson, H., Jonsson, T., Benediktsson, R., Olafsson, J. H., Johannsson, O. T., Hreidarsson, A. B., Sigurdsson, G., Consortium, D., Ferguson-Smith, A. C., Gudbjartsson, D. F., Thorsteinsdottir, U. & Stefansson, K. 2009. Parental origin of sequence variants associated with complex diseases. *Nature,* 462**,** 868-74.

Konigshoff, M., Wilhelm, J., Bohle, R. M., Pingoud, A. & Hahn, M. 2003. HER-2/neu gene copy number quantified by real-time PCR: comparison of gene amplification, heterozygosity, and immunohistochemical status in breast cancer tissue. *Clin Chem,* 49**,** 219-29.

Lamri, A., Abi Khalil, C., Jaziri, R., Velho, G., Lantieri, O., Vol, S., Froguel, P., Balkau, B., Marre, M. & Fumeron, F. 2012. Dietary fat intake and polymorphisms at the PPARG locus modulate BMI and type 2 diabetes risk in the D.E.S.I.R. prospective study. *Int J Obes (Lond),* 36**,** 218-24.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A.,

Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M.*, et al.* 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

Langmead, B. & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9**,** 357-9.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.

Leedy, P. D. & Ormrod, J. E. 2010. *Practical Research: Planning and Design.*, USA: Pearson Education.

Levitt, N. S., Katzenellenbogen, J. M., Bradshaw, D., Hoffman, M. N. & Bonnici, F. 1993. The prevalence and identification of risk factors for NIDDM in urban Africans in Cape Town, South Africa. *Diabetes Care,* 16**,** 601-7.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

Lindi, V. I., Uusitupa, M. I., Lindstrom, J., Louheranta, A., Eriksson, J. G., Valle, T. T., Hamalainen, H., Ilanne-Parikka, P., Keinanen-Kiukaanniemi, S., Laakso, M., Tuomilehto, J. & Finnish Diabetes Prevention, S. 2002. Association of the Pro12Ala polymorphism in the PPAR-gamma2 gene with 3-year incidence of type 2 diabetes and body weight change in the Finnish Diabetes Prevention Study. *Diabetes,* 51**,** 2581-6.

Liu, M., Zhang, J., Guo, Z., Wu, M., Chen, Q., Zhou, Z., Ding, Y. & Luo, W. 2015. [Association and interaction between 10 SNP of peroxisome proliferator-activated receptor and non-HDL-C]. *Zhonghua Yu Fang Yi Xue Za Zhi,* 49**,** 259-64.

Livak, K. J. 2003. SNP genotyping by the 5'-nuclease reaction. *Methods Mol Biol,* 212**,** 129-47.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J. & Pallen, M. J. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol,* 30**,** 434-9.

Ma, R. C., Hu, C., Tam, C. H., Zhang, R., Kwan, P., Leung, T. F., Thomas, G. N., Go, M. J., Hara, K., Sim, X., Ho, J. S., Wang, C., Li, H., Lu, L., Wang, Y., Li, J. W., Wang, Y., Lam, V. K., Wang, J., Yu, W., Kim, Y. J., Ng, D. P., Fujita, H., Panoutsopoulou, K., Day-Williams, A. G., Lee, H. M., Ng, A. C., Fang, Y. J., Kong, A. P., Jiang, F., Ma, X., Hou, X., Tang, S., Lu, J., Yamauchi, T., Tsui, S. K., Woo, J., Leung, P. C., Zhang, X., Tang, N. L., Sy, H. Y., Liu, J., Wong, T. Y., Lee, J. Y., Maeda, S., Xu, G., Cherny, S. S., Chan, T. F., Ng, M. C., Xiang, K., Morris, A. P., Consortium, D., Keildson, S., Mu, T. C., Hu, R., Ji, L., Lin, X., Cho, Y. S., Kadowaki, T., Tai, E. S., Zeggini, E., McCarthy, M. I., Hon, K. L., Baum, L., Tomlinson, B., So, W. Y., Bao, Y., Chan, J. C. & Jia, W. 2013. Genome-wide association study in a Chinese population identifies a susceptibility locus for type 2 diabetes at 7q32 near PAX4. *Diabetologia,* 56**,** 1291-305.

Majithia, A. R., Flannick, J., Shahinian, P., Guo, M., Bray, M. A., Fontanillas, P., Gabriel, S. B., Go, T. D. C., Project, N. J. F. A. S., Consortium, S. T. D., Consortium, T. D. G., Rosen, E. D., Altshuler, D. & Go, T. D. C. 2014. Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc Natl Acad Sci U S A*.

Majumdar, N., Bhowmick, A., Sarkar, P., Doley, R., Kalita, I., Medhi, S. S. & Banu, S. 2014. Study on distribution of Pro12Ala single nucleotide polymorphism of PPARγ2 gene in randomly sampled diabetic population from Guwahati city. *Italian Journal of Medicine,* 9**,** 157-162.

Makishima, M. 2005. Nuclear receptors as targets for drug development: regulation of cholesterol and bile acid metabolism by nuclear receptors. *J Pharmacol Sci,* 97**,** 177-83.

Malecki, M. T., Frey, J., Klupa, T., Skupien, J., Walus, M., Mlynarski, W. & Sieradzki, J. 2003. The Pro12Ala polymorphism of PPARgamma2 gene and susceptibility to type 2 diabetes mellitus in a Polish population. *Diabetes Res Clin Pract,* 62**,** 105-11.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-80.

Marras, S. A., Kramer, F. R. & Tyagi, S. 2003. Genotyping SNPs with molecular beacons. *Methods Mol Biol,* 212**,** 111-28.

Medina-Gomez, G., Gray, S. L., Yetukuri, L., Shimomura, K., Virtue, S., Campbell, M., Curtis, R. K., Jimenez-Linan, M., Blount, M., Yeo, G. S., Lopez, M., Seppanen-Laakso, T., Ashcroft, F. M., Oresic, M. & Vidal-Puig, A. 2007. PPAR gamma 2 prevents lipotoxicity by controlling adipose tissue expandability and peripheral lipid metabolism. *PLoS Genet,* 3**,** e64.

Meigs, J. B., Cupples, L. A. & Wilson, P. W. 2000. Parental transmission of type 2 diabetes: the Framingham Offspring Study. *Diabetes,* 49**,** 2201-7.

Meirhaeghe, A., Fajas, L., Helbecque, N., Cottel, D., Lebel, P., Dallongeville, J., Deeb, S., Auwerx, J. & Amouyel, P. 1998. A genetic polymorphism of the peroxisome proliferator-activated receptor gamma gene influences plasma leptin levels in obese humans. *Hum Mol Genet,* 7**,** 435-40.

Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet,* 11**,** 31-46.

Millat, G., Chanavat, V. & Rousson, R. 2014. Evaluation of a new NGS method based on a custom AmpliSeq library and Ion Torrent PGM sequencing for the fast detection of genetic variations in cardiomyopathies. *Clin Chim Acta,* 433**,** 266-71.

Miramontes González, J. P., Cieza Borrella, C., Mayoral, R., Corral Gudino, L., Makena Hoghtower, C. & González Sarmiento, R. 2014. PPAR gamma pro12Ala polymorphism and type 2 diabetes: a study in a spanish cohort. *Journal of Genetics Study,* 2.

Mitscherlich, E. 1841. Ueber die chemische Verwandtschaftskraft. *Annalen der Physik,* 129**,** 95-117.

Monami, M., Adalsteinsson, J. E., Desideri, C. M., Ragghianti, B., Dicembrini, I. & Mannucci, E. 2013. Fasting and post-prandial glucose and diabetic complication. A meta-analysis. *Nutr Metab Cardiovasc Dis,* 23**,** 591-8.

Montagnana, M., Fava, C., Nilsson, P. M., Engstrom, G., Hedblad, B., Lippi, G., Minuz, P., Berglund, G. & Melander, O. 2008. The Pro12Ala polymorphism of the PPARG gene is not associated with the metabolic syndrome in an urban population of middle-aged Swedish individuals. *Diabet Med,* 25**,** 902-8.

Morran, M. P., Vonberg, A., Khadra, A. & Pietropaolo, M. 2015. Immunogenetics of type 1 diabetes mellitus. *Mol Aspects Med,* 42**,** 42-60.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods,* 5**,** 621-8.

Motavallian, A., Andalib, S., Vaseghi, G., Mirmohammad-Sadeghi, H. & Amini, M. 2013. Association between PRO12ALA polymorphism of the PPAR-gamma2 gene and type 2 diabetes mellitus in Iranian patients. *Indian J Hum Genet,* 19**,** 239-44.

Mukherjee, D. 2009. Peripheral and cerebrovascular atherosclerotic disease in diabetes mellitus. *Best Pract Res Clin Endocrinol Metab,* 23**,** 335-45.

Muller, Y. L., Bogardus, C., Beamer, B. A., Shuldiner, A. R. & Baier, L. J. 2003. A functional variant in the peroxisome proliferator-activated receptor gamma2 promoter is associated with predictors of obesity and type 2 diabetes in Pima Indians. *Diabetes,* 52**,** 1864-71.

Municipality: Mangaung Metropolitan Municipality. [Online]. Available from: http://led.co.za/municipality/mangaung-metropolitan-municipality [Accessed September 11th, 2014].

Mvo, Z., Dick, J. & Steyn, K. 1999. Perceptions of overweight African women about acceptable body size of women and children. *Curationis,* 22**,** 27-31.

Mykkanen, L., Kuusisto, J., Pyorala, K. & Laakso, M. 1993. Cardiovascular disease risk factors as predictors of type 2 (non-insulin-dependent) diabetes mellitus in elderly subjects. *Diabetologia,* 36**,** 553-9.

Nathan, D. M., Singer, D. E., Hurxthal, K. & Goodson, J. D. 1984. The clinical information value of the glycosylated hemoglobin assay. *N Engl J Med,* 310**,** 341-6.

Nemoto, M., Sasaki, T., Deeb, S. S., Fujimoto, W. Y. & Tajima, N. 2002. Differential effect of PPARgamma2 variants in the development of type 2 diabetes between native Japanese and Japanese Americans. *Diabetes Res Clin Pract,* 57**,** 131-7.

Newman, B., Selby, J. V., King, M. C., Slemenda, C., Fabsitz, R. & Friedman, G. D. 1987. Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. *Diabetologia,* 30**,** 763-8.

Njelekela, M. A., Mpembeni, R., Muhihi, A., Mligiliche, N. L., Spiegelman, D., Hertzmark, E., Liu, E., Finkelstein, J. L., Fawzi, W. W., Willett, W. C. & Mtabaji, J. 2009. Gender-related differences in the prevalence of cardiovascular disease risk factors and their correlates in urban Tanzania. *BMC Cardiovasc Disord,* 9**,** 30.

Noble, D., Mathur, R., Dent, T., Meads, C. & Greenhalgh, T. 2011. Risk models and scores for type 2 diabetes: systematic review. *BMJ,* 343**,** d7163.

Olckers, A., Towers, G. W., van der Merwe, A., Schwarz, P. E., Rheeder, P. & Schutte, A. E. 2007. Protective effect against type 2 diabetes mellitus identified within the ACDC gene in a black South African diabetic cohort. *Metabolism,* 56**,** 587-92.

Omar, M. A., Seedat, M. A., Motala, A. A., Dyer, R. B. & Becker, P. 1993. The prevalence of diabetes mellitus and impaired glucose tolerance in a group of urban South African blacks. *S Afr Med J,* 83**,** 641-3.

Osei-Hyiaman, D., Hou, L., Zhiyin, R., Zhiming, Z., Yu, H., Amankwah, A. A. & Harada, S. 2001. Association of a novel point mutation (C159G) of the CTLA4 gene with type 1 diabetes in West Africans but not in Chinese. *Diabetes,* 50**,** 2169-71.

Paramasivam, D., Safi, S. Z., Qvist, R., Abidin, I. B. Z., Hairi, N. N. M. & Chinna, K. 2016. Role of PPARG (Pro12Ala) in Malaysian type 2 diabetes mellitus patients. *International Journal of Diabetes in Developing Countries***,** 1-8.

Pascual, G., Fong, A. L., Ogawa, S., Gamliel, A., Li, A. C., Perissi, V., Rose, D. W., Willson, T. M., Rosenfeld, M. G. & Glass, C. K. 2005. A SUMOylation-dependent pathway mediates transrepression of inflammatory response genes by PPAR-gamma. *Nature,* 437**,** 759-63.

Pattanayak, A. K., Bankura, B., Balmiki, N., Das, T. K., Chowdhury, S. & Das, M. 2014. Role of peroxisome proliferator-activated receptor gamma gene polymorphisms in type 2 diabetes mellitus patients of West Bengal, India. *Journal of Diabetes Investigation,* 5**,** 188-191.

Picard, F., Kurtev, M., Chung, N., Topark-Ngarm, A., Senawong, T., Machado De Oliveira, R., Leid, M., McBurney, M. W. & Guarente, L. 2004. Sirt1 promotes fat mobilization in white adipocytes by repressing PPAR-gamma. *Nature,* 429**,** 771-6.

Poirier, O., Nicaud, V., Cambien, F. & Tiret, L. 2000. The Pro12Ala polymorphism in the peroxisome proliferator-activated receptor gamma2 gene is not associated with postprandial responses to glucose or fat tolerance tests in young healthy subjects: the European Atherosclerosis Research Study II. *J Mol Med (Berl),* 78**,** 346-51.

Poulsen, P., Kyvik, K. O., Vaag, A. & Beck-Nielsen, H. 1999. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. *Diabetologia,* 42**,** 139-45.

Prendergast, K. (2014). *The Link Between Obesity and Diabetes*. [online] Baltimore Washington Medical Center. Available at: https://www.mybwmc.org/link-between-obesity-and-diabetes [Accessed 9 May 2015].

PrimerQuest® program, IDT, Coralville, USA. Retrieved 12 December, 2012. http://www.idtdna.com/Scitools

Prinseq.sourceforge.net, (2016). *PRINSEQ @ SourceForge.net*. [online] Available at: http://prinseq.sourceforge.net/manual.html [Accessed 12 Oct. 2015].

Puoane, T., Steyn, K., Bradshaw, D., Laubscher, R., Fourie, J., Lambert, V. & Mbananga, N. 2002. Obesity in South Africa: the South African demographic and health survey. *Obes Res,* 10**,** 1038-48.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.

Ramadan, R., Zaki, M., Desouky, L., Madkour, M. and Kamel, K. (2016). The C161T Polymorphism in Peroxisome Proliferator-Activated Receptor γ2, but Not Pro12Ala, Is Associated with Diabetic Retinopathy in Type 2 Diabetes Mellitus in an Egyptian Population. *Journal of Diabetes Mellitus*, 06(01), pp.1-9.

Raza, S. T., Abbas, S., Ahmed, F., Fatima, J., Zaidi, Z. H. & Mahdi, F. 2012. Association of MTHFR and PPARgamma2 gene polymorphisms in relation to type 2 diabetes mellitus cases among north Indian population. *Gene,* 511**,** 375-9.

Ristow, M., Muller-Wieland, D., Pfeiffer, A., Krone, W. & Kahn, C. R. 1998. Obesity associated with a mutation in a genetic regulator of adipocyte differentiation. *N Engl J Med,* 339**,** 953-9.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011. Integrative genomics viewer. *Nat Biotech,* 29**,** 24-26.

Robitaille, J., Despres, J. P., Perusse, L. & Vohl, M. C. 2003. The PPAR-gamma P12A polymorphism modulates the relationship between dietary fat intake and components of the metabolic syndrome: results from the Quebec Family Study. *Clin Genet,* 63**,** 109-16.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* 475**,** 348-52.

Sanger, F., Nicklen, S. & Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74**,** 5463-7.

Sanghera, D. K., Demirci, F. Y., Been, L., Ortega, L., Ralhan, S., Wander, G. S., Mehra, N. K., Singh, J., Aston, C. E., Mulvihill, J. J. & Kamboh, I. M. 2010. PPARG and ADIPOQ gene

polymorphisms increase type 2 diabetes mellitus risk in Asian Indian Sikhs: Pro12Ala still remains as the strongest predictor. *Metabolism,* 59**,** 492-501.

Sanghera, D. K., Ortega, L., Han, S., Singh, J., Ralhan, S. K., Wander, G. S., Mehra, N. K., Mulvihill, J. J., Ferrell, R. E., Nath, S. K. & Kamboh, M. I. 2008. Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk. *BMC Med Genet,* 9**,** 59.

Savkur, R. S. & Miller, A. R. 2006. Investigational PPAR-gamma agonists for the treatment of Type 2 diabetes. *Expert Opin Investig Drugs,* 15**,** 763-78.

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., Hughes, T. E., Groop, L., Altshuler, D., Almgren, P., Florez, J. C., Meyer, J., Ardlie, K., Bengtsson Bostrom, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H. N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Rastam, L., Speliotes, E. K., Taskinen, M. R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjogren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S. B., Chirn, G. W., Ma, Q., Parikh, H., Richardson, D., Ricke, D. & Purcell, S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science,* 316**,** 1331-6.

Schmieder, R. & Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics,* 27**,** 863-4.

Schwarz, P. E. H., Towers, G. W., van der Merwe, A., Perez-Perez, L., Rheeder, P., Schulze, J., Bornstein, S. R., Licinio, J., Wong, M. L., Schutte, A. E. & Olckers, A. 2008. Global meta-analysis of the C-11377G alteration in the ADIPOQ gene indicates the presence of population-specific effects: challenge for global health initiatives. *Pharmacogenomics J,* 9**,** 42-48.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., Prokunina-Olsson, L., Ding, C. J., Swift, A. J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X. Y., Conneely, K. N., Riebow, N. L., Sprau, A. G., Tong, M., White, P. P., Hetrick, K. N., Barnhart, M. W., Bark, C. W., Goldstein, J. L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T. A., Watanabe, R. M., Valle, T. T., Kinnunen, L., Abecasis, G. R., Pugh, E. W., Doheny, K. F., Bergman, R. N., Tuomilehto, J., Collins, F. S. & Boehnke, M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science,* 316**,** 1341-5.

Shah, N. R. & Braverman, E. R. 2012. Measuring adiposity in patients: the utility of body mass index (BMI), percent body fat, and leptin. *PLoS One,* 7**,** e33308.

Shahrjerdi, A., Jadhav, B. & Jamkhedkar, S. 2013. Study of single-nucleotide polymorphism within candidate genes associated with type 2 diabetes in Western Indian population. *Journal of Pharmacy Research,* 6**,** 233-238.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. & Church, G. M. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science,* 309**,** 1728-32.

Sher, T., Yi, H. F., McBride, O. W. & Gonzalez, F. J. 1993. cDNA cloning, chromosomal mapping, and functional characterization of the human peroxisome proliferator activated receptor. *Biochemistry,* 32**,** 5598-604.

Shrestha, R. K., Lubinsky, B., Bansode, V. B., Moinz, M. B., McCormack, G. P. & Travers, S. A. 2014. QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics,* 15**,** 33.

Signorello, L. B., Schlundt, D. G., Cohen, S. S., Steinwandel, M. D., Buchowski, M. S., McLaughlin, J. K., Hargreaves, M. K. & Blot, W. J. 2007. Comparing diabetes prevalence

between African Americans and Whites of similar socioeconomic status. *Am J Public Health,* 97**,** 2260-7.

Simen, B. B., Yin, L., Goswami, C. P., Davis, K. O., Bajaj, R., Gong, J. Z., Peiper, S. C., Johnson, E. S. & Wang, Z. X. 2015. Validation of a next-generation-sequencing cancer panel for use in the clinical laboratory. *Arch Pathol Lab Med,* 139**,** 508-17.

Singh, R. R., Patel, K. P., Routbort, M. J., Reddy, N. G., Barkoh, B. A., Handal, B., Kanagal-Shamanna, R., Greaves, W. O., Medeiros, L. J., Aldape, K. D. & Luthra, R. 2013. Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes. *J Mol Diagn,* 15**,** 607-22.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C. & Froguel, P. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature,* 445**,** 881-5.

Small, K. S., Hedman, A. K., Grundberg, E., Nica, A. C., Thorleifsson, G., Kong, A., Thorsteindottir, U., Shin, S. Y., Richards, H. B., Consortium, G., Investigators, M., Consortium, D., Soranzo, N., Ahmadi, K. R., Lindgren, C. M., Stefansson, K., Dermitzakis, E. T., Deloukas, P., Spector, T. D., McCarthy, M. I. & Mu, T. C. 2011. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat Genet,* 43**,** 561-4.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. & Hood, L. E. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature,* 321**,** 674-9.

Soriguer, F., Morcillo, S., Cardona, F., Rojo-Martínez, G., de la Cruz Almaráz, M., de la Soledad Ruiz de Adana, M., Olveira, G., Tinahones, F. & Esteva, I. 2006. Pro12Ala Polymorphism of the PPARG2 Gene Is Associated with Type 2 Diabetes Mellitus and Peripheral Insulin Sensitivity in a Population with a High Intake of Oleic Acid. *The Journal of Nutrition,* 136**,** 2325-2330.

Stead, L. F., Sutton, K. M., Taylor, G. R., Quirke, P. & Rabbitts, P. 2013. Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: applications in tumor subclone resolution. *Hum Mutat,* 34**,** 1432-8.

Steyn, K., Kazenellenbogen, J. M., Lombard, C. J. & Bourne, L. T. 1997. Urbanization and the risk for chronic diseases of lifestyle in the black population of the Cape Peninsula, South Africa. *J Cardiovasc Risk,* 4**,** 135-42.

Stumvoll, M. & Haring, H. 2002. The peroxisome proliferator-activated receptor-gamma2 Pro12Ala polymorphism. *Diabetes,* 51**,** 2341-7.

Svendsen, P. A., Lauritzen, T., Soegaard, U. & Nerup, J. 1982. Glycosylated haemoglobin and steady-state mean blood glucose concentration in Type 1 (insulin-dependent) diabetes. *Diabetologia,* 23**,** 403-5.

Tahara, Y. & Shima, K. 1995. Kinetics of HbA1c, Glycated Albumin, and Fructosamine and Analysis of Their Weight Functions Against Preceding Plasma Glucose Level. *Diabetes Care,* 18**,** 440-447.

Tai, E. S., Corella, D., Deurenberg-Yap, M., Adiconis, X., Chew, S. K., Tan, C. E. & Ordovas, J. M. 2004. Differential effects of the C1431T and Pro12Ala PPARgamma gene variants on plasma lipids and diabetes risk in an Asian population. *J Lipid Res,* 45**,** 674-85.

Takada, I. & Makishima, M. 2015. Therapeutic application of vitamin D receptor ligands: an updated patent review. *Expert Opin Ther Pat*, 25, 1373-83.

Tavares, V., Hirata, R. D., Rodrigues, A. C., Monte, O., Salles, J. E., Scalissi, N., Speranza, A. C. & Hirata, M. H. 2005. Association between Pro12Ala polymorphism of the PPAR-

gamma2 gene and insulin sensitivity in Brazilian patients with type-2 diabetes mellitus. *Diabetes Obes Metab,* 7**,** 605-11.

Thomson, G. 1984. HLA DR antigens and susceptibility to insulin-dependent diabetes mellitus. *Am J Hum Genet,* 36**,** 1309-17.

Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics,* 14**,** 178-192.

Tillil, H. & Kobberling, J. 1987. Age-corrected empirical genetic risk estimates for first-degree relatives of IDDM patients. *Diabetes,* 36**,** 93-9.

Tong, J. Y., Huang, N., Wang, L., Yi, Y. S., Pan, X. H. & Lu, Y. 2012. [Relationship between PPARgamma2 Pro12Ala polymorphism and type 2 diabetes mellitus in Chinese Han population: a Meta-analysis]. *Zhonghua Yu Fang Yi Xue Za Zhi,* 46**,** 359-63.

Tonjes, A., Scholz, M., Loeffler, M. & Stumvoll, M. 2006. Association of Pro12Ala polymorphism in peroxisome proliferator-activated receptor gamma with Pre-diabetic phenotypes: meta-analysis of 57 studies on nondiabetic individuals. *Diabetes Care,* 29**,** 2489-97.

Tontonoz, P., Hu, E., Graves, R. A., Budavari, A. I. & Spiegelman, B. M. 1994. mPPAR gamma 2: tissue-specific regulator of an adipocyte enhancer. *Genes Dev,* 8**,** 1224-34.

Trombetta, M., Bonetti, S., Boselli, M. L., Miccoli, R., Trabetti, E., Malerba, G., Pignatti, P. F., Bonora, E., Del Prato, S. & Bonadonna, R. C. 2013. PPARG2 Pro12Ala and ADAMTS9 rs4607103 as "insulin resistance loci" and "insulin secretion loci" in Italian individuals. The GENFIEV study and the Verona Newly Diagnosed Type 2 Diabetes Study (VNDS) 4. *Acta Diabetol,* 50**,** 401-8.

Tuei, V. C., Maiyoh, G. K. & Ha, C. E. 2010. Type 2 diabetes mellitus and obesity in sub-Saharan Africa. *Diabetes Metab Res Rev,* 26**,** 433-45.

Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A. P. 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res,* 36**,** e25.

Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D. P., Holmkvist, J., Borch-Johnsen, K., Jorgensen, T., Sandbaek, A., Lauritzen, T., Hansen, T., Nurbaya, S., Tsunoda, T., Kubo, M., Babazono, T., Hirose, H., Hayashi, M., Iwamoto, Y., Kashiwagi, A., Kaku, K., Kawamori, R., Tai, E. S., Pedersen, O., Kamatani, N., Kadowaki, T., Kikkawa, R., Nakamura, Y. & Maeda, S. 2008. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet,* 40**,** 1098-102.

Valve, R., Sivenius, K., Miettinen, R., Pihlajamaki, J., Rissanen, A., Deeb, S. S., Auwerx, J., Uusitupa, M. & Laakso, M. 1999. Two polymorphisms in the peroxisome proliferator-activated receptor-gamma gene are associated with severe overweight among obese women. *J Clin Endocrinol Metab,* 84**,** 3708-12.

Verge, C. F., Gianani, R., Kawasaki, E., Yu, L., Pietropaolo, M., Chase, H. P., Eisenbarth, G. S. & Jackson, R. A. 1996. Prediction of Type I Diabetes in First-Degree Relatives Using a Combination of Insulin, GAD, and ICA512bdc/IA-2 Autoantibodies. *Diabetes,* 45**,** 926-933.

Vergotine, Z., Kengne, A. P., Erasmus, R. T., Yako, Y. Y. & Matsha, T. E. 2014a. Rare mutations of peroxisome proliferator-activated receptor gamma: frequencies and relationship with insulin resistance and diabetes risk in the mixed ancestry population from South Africa. *Int J Endocrinol,* 2014**,** 187985.

Vergotine, Z., Yako, Y. Y., Kengne, A. P., Erasmus, R. T. & Matsha, T. E. 2014b. Proliferator-activated receptor gamma Pro12Ala interacts with the insulin receptor substrate 1 Gly972Arg and increase the risk of insulin resistance and diabetes in the mixed ancestry population from South Africa. *BMC Genet,* 15**,** 10.

Vijayaraghavan, K. 2010. Treatment of dyslipidemia in patients with type 2 diabetes. *Lipids in Health and Disease,* 9**,** 144-144.

Vimaleswaran, K. S., Radha, V., Jayapriya, M. G., Ghosh, S., Majumder, P. P., Rao, M. R. & Mohan, V. 2010. Evidence for an association with type 2 diabetes mellitus at the PPARG locus in a South Indian population. *Metabolism,* 59**,** 457-62.

Volgyi, E., Tylavsky, F. A., Lyytikainen, A., Suominen, H., Alen, M. & Cheng, S. 2008. Assessing body composition with DXA and bioimpedance: effects of obesity, physical activity, and age. *Obesity (Silver Spring),* 16**,** 700-5.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. & Burge, C. B. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature,* 456**,** 470-6.

Wang, X., Liu, J., Ouyang, Y., Fang, M., Gao, H. & Liu, L. 2013. The association between the Pro12Ala variant in the PPARgamma2 gene and type 2 diabetes mellitus and obesity in a Chinese population. *PLoS One,* 8**,** e71985.

Ward, C. L., Dempsey, M. H., Ring, C. J., Kempson, R. E., Zhang, L., Gor, D., Snowden, B. W. & Tisdale, M. 2004. Design and performance testing of quantitative real time PCR assays for influenza A and B viral load measurement. *J Clin Virol,* 29**,** 179-88.

Wareham, N. J., Franks, P. W. & Harding, A. H. 2002. Establishing the role of gene-environment interactions in the etiology of type 2 diabetes. *Endocrinol Metab Clin North Am,* 31**,** 553-66.

Wei, G. S., Coady, S. A., Goff, D. C., Jr., Brancati, F. L., Levy, D., Selvin, E., Vasan, R. S. & Fox, C. S. 2011. Blood pressure and the risk of developing diabetes in african americans and whites: ARIC, CARDIA, and the framingham heart study. *Diabetes Care,* 34**,** 873-9.

Weinger, K., Jacobson, A. M., Draelos, M. T., Finkelstein, D. M. & Simonson, D. C. 1995. Blood glucose estimation and symptoms during hyperglycemia and hypoglycemia in patients with insulin-dependent diabetes mellitus. *Am J Med,* 98**,** 22-31.

Weiss, E. P., Kulaputana, O., Ghiu, I. A., Brandauer, J., Wohn, C. R., Phares, D. A., Shuldiner, A. R. & Hagberg, J. M. 2005. Endurance training-induced changes in the insulin response to oral glucose are associated with the peroxisome proliferator-activated receptor-gamma2 Pro12Ala genotype in men but not in women. *Metabolism,* 54**,** 97-102.

WHO. 2003. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care,* 26 Suppl 1**,** S5-20.

WHO. 2015. *Diabetes: the cost of diabetes* [Online]. World Health Organization Available: http://www.who.int/mediacentre/factsheets/fs236/en/ [Accessed 17/03 2015].

Wilkin, T. J. 2001. The accelerator hypothesis: weight gain as the missing link between Type I and Type II diabetes. *Diabetologia,* 44**,** 914-22.

Wommack, K. E., Bhavsar, J. & Ravel, J. 2008. Metagenomics: read length matters. *Appl Environ Microbiol,* 74**,** 1453-63.

Wong, M. L. & Medrano, J. F. 2005. Real-time PCR for mRNA quantitation. *Biotechniques,* 39**,** 75-85.

World Health Organization, (2011). *Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus*. Abbreviated Report of a WHO Consultation. [online] Switzerland: WHO Press. Available at: http://www.who.int/diabetes/publications/report-hba1c_2011.pdf [Accessed 22 Sep. 2014].

Xuan, J., Yu, Y., Qing, T., Guo, L. & Shi, L. 2013. Next-generation sequencing in the clinic: promises and challenges. *Cancer Lett,* 340**,** 284-95.

Yanase, T., Yashiro, T., Takitani, K., Kato, S., Taniguchi, S., Takayanagi, R. & Nawata, H. 1997. Differential Expression of PPAR γ1 and γ2 Isoforms in Human Adipose Tissue. *Biochemical and Biophysical Research Communications,* 233**,** 320-324.

Ye, E., Yang, H., Chen, L., Chen, Q., Sun, M., Lin, Z., Yu, L., Peng, M., Zhang, C. & Lu, X. 2014. Adiponectin and peroxisome proliferator-activated receptor-gamma gene polymorphisms and gene-gene interactions with type 2 diabetes. *Life Sci,* 98**,** 55-9.

Yen, C. J., Beamer, B. A., Negri, C., Silver, K., Brown, K. A., Yarnall, D. P., Burns, D. K., Roth, J. & Shuldiner, A. R. 1997. Molecular scanning of the human peroxisome proliferator activated receptor gamma (hPPAR gamma) gene in diabetic Caucasians: identification of a Pro12Ala PPAR gamma 2 missense mutation. *Biochem Biophys Res Commun,* 241**,** 270-4.

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R., Rayner, N. W., Freathy, R. M., Barrett, J. C., Shields, B., Morris, A. P., Ellard, S., Groves, C. J., Harries, L. W., Marchini, J. L., Owen, K. R., Knight, B., Cardon, L. R., Walker, M., Hitman, G. A., Morris, A. D., Doney, A. S., Wellcome Trust Case Control, C., McCarthy, M. I. & Hattersley, A. T. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science,* 316**,** 1336-41.

Zhou, X. P., Smith, W. M., Gimm, O., Mueller, E., Gao, X., Sarraf, P., Prior, T. W., Plass, C., von Deimling, A., Black, P. M., Yates, A. J. & Eng, C. 2000. Over-representation of PPARgamma sequence variants in sporadic cases of glioblastoma multiforme: preliminary evidence for common low penetrance modifiers for brain tumour risk in the general population. *J Med Genet,* 37**,** 410-4.

Zhu, Y., Alvares, K., Huang, Q., Rao, M. S. & Reddy, J. K. 1993. Cloning of a new member of the peroxisome proliferator-activated receptor gene family from mouse liver. *J Biol Chem,* 268**,** 26817-20.

Zouari Bouassida, K., Chouchane, L., Jellouli, K., Cherif, S., Haddad, S., Gabbouj, S. & Danguir, J. 2005. The peroxisome proliferator activated receptorgamma2 (PPARgamma2) Pro12Ala variant: lack of association with type 2 diabetes in obese and non obese Tunisian patients. *Diabetes Metab,* 31**,** 119-23.

APPENDIX A (1)

Informed consent document example as used in study ECUFS 162/2012 (the same consent document was used in collection of both cohorts, and was translated into Afrikaans and South Sotho):
Informed Consent for Genetic polymorphisms in black South Africans with Type 2 Diabetes Mellitus from the central Free State area.

I, _____(Principle Investigator name), am conducting a study to evaluate if there is a genetic cause of Type 2 diabetes mellitus (sugar disease) in the black population, for the purpose of obtaining a higher degree. To do this study, human blood is needed to assess genes inside the blood (genes are what you inherit from your mother and father – therefore it will be in your blood). In the study, weight, height and body composition will also be needed.

You are invited to participate in this study by volunteering to give a small amount of your blood, and allowing me to measure your weight and height. If you agree, we would also like to keep your blood for further genetic tests to do future research.

All blood will be drawn by a qualified professional nurse at the diabetes clinic. Three small tubes of blood (15ml) will be drawn from a vein in your arm. Your weight and height will be taken by a qualified dietician, trained to take these measurements.

During the blood drawing process, you may experience discomfort and bruising at the site where the blood is drawn.

You will not benefit financially from participating in this study, and there will be no costs to you. You will receive a light snack after blood has been drawn. Participating in this study will not have direct bearing on your normal medical treatment. No feedback can be given on genetic testing because the role of this gene on the population is still unknown. By participating in this study, you will make a contribution to help find if a gene generally linked to type 2 diabetes in black South Africans and thereby assist to help treat Type 2 diabetes mellitus in the future.

The researcher will keep records of all the genetic information, weight, and height in a secure database. Only the researchers will know the identity of the study participants, because the blood will be marked with a number and not names.

The results of this study will be published once the study is completed. Should you feel that your participation in this study has been detrimental to you, please contact me at _____ (Principle investigator phone number).

Your signature on this form means that you understand the information given to you and that- you are volunteering to participate in this study. It also means that we may use your blood for other genetic studies after this study is completed. You can withdraw from this study at any time. If you are unhappy to have your blood stored for future research – it will be disposed of at the end of the study. If you have any sensitivity on how your blood should be disposed of please indicate how, when the blood is taken. These will be recorded and taken into account at the time of disposal. Your routine medical treatment will not be compromised in any way if you do.

_____ _____

Signature of Participant Signature of researcher

Contact information for Ethical Committee if there are any ethical concerns.
Contact number: 0514052812

APPENDIX A (2)

Genetic Informed Consent example as used in study ECUFS 162/2012 (the same consent document was used in collection of both cohorts, and was translated into Afrikaans and South Sotho):

Information document for genetic research for the study Genetic polymorphisms in black South Africans with Type 2 Diabetes Mellitus from the central Free State area.

We are planning a research project on the genetic cause of type 2 diabetes mellitus and request your permission to draw your blood (10ml) and to use your DNA present in blood for further laboratory tests.
Genes are what you inherit from your parents. They are found in every part of your body and therefore they will be present in blood.
The findings of this study will not have direct bearing on your health management, but may eventually benefit others in terms of prevention or treatment of health conditions.
You are free to refuse consent and you do not have to give reasons for doing so.

Privacy and Confidentiality
The following arrangements have been made to ensure privacy and confidentiality of your genetic information:

- Your blood sample will be marked with a code and not your name. Only the Principle Investigator will therefore be able to identify the sample, but not technicians working with the sample.

Results of research
It is not intended to provide feedback because the association of the gene is still not clear.
If research generates information about you which may be of relevance to the health of other family members, your consent will be sought before offering to disclose such information to the family members concerned.

Family members
Information about family members, in addition to that provided by you, is not required for the research.
Your material and information will not be released for other uses other than research without consent, unless required by law.

Storage

We would like to retain your blood and DNA for possible future research.

The duration of storage will be maximum fifteen years.

If you are unhappy to have your blood stored for future research, your genetic material and information will be disposed of at the end of this study, once the sample storage and record-keeping requirements of good research practice have been met.

Do you have any sensitivity on how your blood should be disposed of? If so, what are they?

_____

These will be recorded and taken into account at the time of disposal.

We can dispose of your genetic material even after the research has started since the samples are stored in an identifiable form.

Voluntary Participation

You do not have to agree to take part in this research and you are free to withdraw from the research at any time. Your routine medical treatment will not be compromised in any way if you do not participate.

---	---
Signature of participant:	Signature of researcher
Name :	Name:
Date:	Date

APPENDIX B

General/Terminology

a. Command line

The command line is a text interface which can be used to perform various computational tasks such as creating files and directories, navigating through files and folders, to execute software etc. Commands are executable programs written in Shell, Perl, Python, Ruby, etc.

b. Shell script

The Shell is a program that takes commands from the keyboard and passes them to the operating system. A shell script is a sequence of commands for which you have repeated use. Shell scripts can be created using any text editor such as vim, emacs, nano or gedit. On most Linux systems a program called bash (Bourne Again SHell) acts as the shell interpreter.

c. A "terminal"

A "terminal" is a program called a terminal emulator. This programs opens a window and allows you to interact with the shell. In Linux the gnome-terminal is commonly used. Gedit is the editor supplied with the Gnome desktop environment with a graphical interface.

d. ASCII Table

ASCII stands for American Standard Code for Information Interchange. ASCII was first introduced in 1968 as a method of encoding alphabetical and numerical data in digital format. ASCII code is standardized allowing computers and other electronic devices to exchange data with each other.

e. FASTQ file

FASTQ file format is a text-based format that is used for storing nucleotide sequence and its Phred quality scores. The quality score of each sequence letter is encoded with a single ASCII character (offset of 33). It provides a simple extension to the FASTA format.

```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
3 +
4 !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

The first line starts with the “@” character which is followed by the sequence identifier or sequence name. The second line 2 contains the raw sequence letters. The next line contains a single “+” character. The last line (line 4) encoding the quality values according to the ASCII code for the sequence in line 2.

f. Phred

Phred is a base calling program (Ewing *et al.*, 1998). The Phred quality scores are used for assessing sequence quality, identification and removal of low-quality sequence (end trimming), and determination of accurate consensus sequences.

Phred Quality Score	Base call accuracy	Probability of incorrect base call
10	90%	1 in 10
20	99%	1 in 100
30	99.9%	1 in 1 000
40	99.99%	1 in 10 000
50	99.999%	1 in 100 000
60	99.9999%	1 in 1 000 000

g. For Loop

The **for** loop allows you to iterate (repeat a process) over a series of 'words' within a string. It controls a sequence of repetitions.

A basic *for* structure:

```
1  #/bin/bash
2  for i in $[1..8]
3      do
4              …
5      done
```

The second line, “i”, is declared to be a variable that can take different values contained in “$”, followed by the commands in line 3 to 5. The loop is initialized in

line 3 with “do” and upon fulfilment of the for loop is concluded in line 5 with “done”. In line 5 ‘done’ indicates that the code written in the loop is finished. The bounds of “i” is defined in “$[1..8]”.

h. Nested Loops

If one loop is placed inside the body of another loop it is called nesting. If two loops are “nested”, the outer loop takes control of the number of complete repetitions of the inner loop.

i. **“Echo” command**

The “echo” command is a shell built-in that prints out its text arguments to standard output. Thus, the command that you want to execute can be displayed before it is evaluated (before the shell acts on it).

j. “**Eval” command**

The “eval” (evaluate) command executes the command in the current script.

APPENDIX C

Table 1: Table containing all the software used for NGS data analysis.

NGS Software	Reference	Software Availability	Open Source	License	File format required	Output file	URL
BamTools	Barnett et al., 2011	Free download	Yes	MIT license	BAM files	FASTQ	http://github.org/pezmaster31/bamtools.
PRINSEQ	Shrestha et al., 2014	Free download	Yes	GNU Public License (GPL)	FASTQ	FASTQ	http://prinseq.sourceforge.net/
Bowtie2	Langmead et al., 2009	Free download	Yes	GPLv3 license	FASTQ	SAM file	http://bowtie.cbcb.umd.edu
SamTools	Li et al., 2009	Free download	Yes	BSD License, MIT License	SAM	BAM/mpileup	http://www.htslib.org/download/
Varscan 2	Koboldt et al., 2012	Free download	Yes	Non-Profit OSL 3.0	mpileup	VCF/CSV	http://varscan.sourceforge.net.
Integrative Genomics Viewer (IGV)	Robinson et al., 2011	Free download	Yes	GNU Lesser General Public License (LGPL)	Sorted BAM	None	http://www.broadinstitute.org/igv/download.

APPENDIX D

1. File conversion (BamTools)

The basic format of a command used in BamTools:

```
>bamtools convert –format fastq –in (file name).bam –out (file name).fastq
```

2. Quality Control (PRINSEQ)

a) Prinseq-lite without quality control parameters

Table 2: Options used in the prinseq-lite command when no parameters were included.

Input Option
-fastq **< filename-C$i.fastq / filename-P$i.fastq >**
Output Option
–out_good **<null>**
-out_bad **<null>**

Command for running prinseq-lite for controls and patients, without any parameters:

```
for e in C P
      for i in {1..8}
      do
            qc_graphs_c="prinseq-lite.pl -verbose -fastq filename-
            $e$i.fastq -out_good null -out_bad null -graph_data
            $e$i-orig.gd"
            echo $qc_graphs_c
            eval $qc_graphs_c

      done
done
```

b) Prinseq-graph without quality control parameters

Table 3: The input and output options used in the prinseq-graph command when no parameters were included.

Input Option
-i **<C$i-orig.gd/ P$i-orig.gd>**
Output Options
-o **<C$i-orig_graph/ P$i-orig_graph>**
-png all
-html all

Command for running prinseq-graph for controls and patients for data without QC.

```
For e in C P

      for i in {1..8}
      do
             qc_graphs_c="prinseq-graphs.pl -verbose -i $e$i-
             orig.gd -o $e$i-orig_graph -png_all -html_all"
             echo $qc_graphs_c
             eval $qc_graphs_c

      done
done
```

c) Prinseq-lite with quality control parameters

Table 4: The parameters that was included when prinseq-lite was run.

Input Option
–fastq **<filename-ki.fastq>**
Output Option
–out_good **<ki-QCed>**
-out_bad **null**
Filter Option
-min_len <**40>**

Trim Option
-trim_qual_right <**30>**
-trim_qual_type <**min>**
-trim_qual_rule <**lt>**
-trim_qual_window <**10>**
-trim_qual_step <**3>**

The command for running prinseq-lite with specific parameters included for quality control:

```
for e in C P
do
        for i in {1..8}
        do
                trim_cmd="prinseq-lite.pl -verbose -fastq filename-
                $e$i.fastq -out_good $e$i-QCed -out_bad null -
                min_len 40 -trim_qual_right 30 -trim_qual_type min
                -trim_qual_rule lt -trim_qual_window 10 -
                trim_qual_step 3"
                echo $trim_cmd
                eval $trim_cmd
        done
done
```

d) Prinseq-graph with quality control parameters

Table 5: The input and output options when prinseq-graph was run.

Input Option
-i **<C$i-QCed.gd/ P$i-QCed.gd>**
Output Options
-o **<C$i-QCed_graph / P$i-QCed_graph >**
-png all
-html all

The command for running prinseq-graph with specific parameters included for quality control:

```
for e in C P

      for i in {1..8}
      do
             qc_graphs_c="prinseq-graphs.pl -verbose -i C$i-QCed.gd -
             o C$i-QCed_graph -png_all -html_all"
             echo $qc_graphs_c
             eval $qc_graphs_c

      done
done
```

3. Mapping

The basic command running Bowtie2

```
bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>}
-S [<hit>]
```

The basic command for running Bowtie-build

```
bowtie2-build [options]* <reference_in> <bt2_base>
```

3.1 Mapping to Chromosome 3

Command for building indexes for Chr3:

```
do
      bt2_build="bowtie2-build chr3_ref_hg19.fa chr3"
      echo $bt2_build
      eval $bt2_build
done
```

Command for running Bowtie2 using Chr3 as reference:

```
for e in C P
do
	for i in {1..8}
	do
		bt_cmd="bowtie2 --end-to-end --sensitive -a -p 16 -q
		$e$i-QCed.fastq -x chr3 -S $e$i-bt2_pparg_chr3.sam"
		samtools_conv="samtools view -Sb $e$i-
		bt2_pparg_chr3.sam > $e$i-bt2_pparg_chr3.bam"
		samtools_sort="samtools sort $e$i-bt2_pparg_chr3.bam
		$e$i-bt2_pparg_chr3_sorted"
		samtools_index="samtools index $e$i-
		bt2_pparg_chr3_sorted.bam"
		echo $bt_cmd
		eval $bt_cmd
		echo $samtools_conv
		eval $samtools_conv
		echo $samtools_sort
		eval $samtools_sort
		echo $samtools_index
		eval $samtools_index

	done
done
```

4. Conversion and Sorting (SAMtools mpileup)

Mpileup files are generated with the following basic command:

```
samtools mpileup -f [reference sequence] [BAM file(s)]
>myData.mpileup
```

Commands for building and sorting mpileup files from sorted BAM files:

```
for e in C P
do
	for i in {1..8}
	do
		samtools_cmd="samtools mpileup -f chr3_ref_hg19.fa
		$e$i-bt2_pparg_chr3_sorted.bam > $e$i-
		bt2_pparg_chr3_sorted.mpileup"
		echo $samtools_cmd
		eval $samtools_cmd
	done
done
```

5. Variant calling

The commands for calling SNPs by generating vcf and csv files using Varscan:

```
for e in C P
do
        for i in {1..8}
        do
                vcf_cmd="java -jar VarScan.v2.4.1.jar mpileup2snp
                $e$i-bt2_pparg_chr3_sorted.mpileup --output-vcf 1 >
                $e$i-bt2_pparg_chr3_sorted.vcf"
                csv_cmd="java -jar VarScan.v2.4.1.jar mpileup2snp
                $e$i-bt2_pparg_chr3_sorted.mpileup --output-vcf 0 >
                $e$i-bt2_pparg_chr3_sorted.csv"
                echo $vcf_cmd
                eval $vcf_cmd
                echo $csv_cmd
                eval $csv_cmd
        done
done
```

APPENDIX E

Table 6: Table showing all patient samples collected. Included is the participant number, age, weight, height, BMI, HbA1c level and genotypes obtained from qPCR results. The individuals selected for NGS are shaded in grey.

Participant Number	Age (years)	Weight (kg)	Height (cm)	BMI (kg/m^2)	HbA1c (%)	Genotype
P1	46	69.8	156	28.7	12.6	C/C
P2	49	84.3	159	33.3	8.1	C/C
P3	42	69.8	160	27.3	7.1	C/C
P4	41	79.3	157	32.2	12.6	C/C
P5	40	53.1	153	22.7	6.5	C/C
P6	49	74.6	157	30.2	13.3	C/C
P7	47	100.4	156	41.3	10.0	C/C
P8	41	117.5	163	44.2	12.6	C/C
P9	53	64.2	154	27.2	6.0	C/C
P10	53	117.4	157	47.8	7.7	C/C
P11	54	109.9	154	46.4	6.2	C/C
P12	55	86.6	160	34.0	6.7	C/C
P13	55	108.2	162	41.2	9.9	C/C
P14	43	120.6	165	44.3	7.6	C/C
P15	52	83.0	158	33.2	8.3	C/C
P16	45	116.1	165	42.9	7.5	C/C
P17	51	82.9	149	37.3	8.9	C/C
P18	43	82.0	162	31.2	5.8	C/C
P19	55	94.4	161	36.3	11.1	C/C
P20	57	104.0	152	45.0	7.2	C/C
P21	41	106.4	168	37.9	12.8	C/C
P22	57	94.2	153	40.5	9.3	C/C
P23	54	80.7	153	34.4	10.6	C/C
P24	52	115.0	157	46.8	8.5	C/C
P25	59	73.5	151	32.1	8.0	C/C
P26	57	125.5	157	51.2	7.6	C/C
P27	41	70.5	153	30.1	6.5	C/C
P28	52	53.4	142	26.6	6.0	C/C
P29	48	105.6	155	44.0	8.4	C/C
P30	46	87.6	161	34.0	6.5	C/C
P31	54	84.7	161	32.6	8.7	C/C
P32	49	56.3	159	22.4	6.6	C/C
P33	51	96.8	166	35.3	12.6	C/C
P34	57	123.0	159	48.9	6.5	C/C
P35	59	87.0	156	36.0	6.1	C/C
P36	61	97.5	158	39.0	8.5	C/C

P37	60	82.5	157	33.4	7.7	C/C
P38	46	52.3	156	21.5	12.3	C/C
P39	56	128.2	166	46.5	8.5	C/C
P40	60	121.4	162	46.3	8.3	C/C
P41	50	60.9	152	26.3	13.8	C/C
P42	52	79.4	157	32.1	11.7	C/C
P43	50	130.6	161	50.2	8.2	C/C
P44	51	101.4	151	44.4	6.8	C/C
P45	49	106.0	160	41.6	5.3	C/C
P46	59	93.3	152	40.4	8.3	C/C
P47	56	84.8	172	28.7	5.6	C/C
P48	38	90.6	159	35.9	11.7	C/C
P49	44	79.2	152	34.4	7.7	C/C
P50	48	82.3	158	32.9	8.4	C/C
P51	60	80.1	150	35.8	7.4	C/C
P52	53	101.9	160	39.8	9.6	C/C
P53	51	93.9	147	43.5	8.7	C/C
P54	57	58.5	152	25.5	7.8	C/C
P55	46	86.1	162	32.7	9.5	C/C
P56	57	81.6	153	34.8	13.2	C/C & C/G
P57	47	73.3	164	27.3	6.8	C/C
P58	35	97.3	159	38.5	10.0	C/C
P59	40	127.2	156	52.2	7.8	C/C
P60	39	66.7	153	28.5	12.3	C/C
P61	60	121.2	157	49.4	8.6	C/C
P62	56	67.7	145	32.2	5.7	C/C
P63	44	52.4	148	23.9	5.8	C/C
P64	54	121.6	169	42.6	6.7	C/C
P65	49	96.2	158	38.7	10.9	C/C
P66	55	74.7	164	27.9	12.5	C/C
P67	45	83.8	163	31.5	10.9	C/C
P68	57	70.8	150	31.5	8.9	C/C
P69	53	78.3	155	32.8	6.4	C/C
P70	60	109.1	150	48.8	10.7	C/C
P71	41	73.5	152	31.8	5.5	C/C
P72	60	101.8	157	41.3	6.8	C/C
P73	56	63.0	146	29.6	7.1	C/C
P74	48	67.0	161	25.8	10.6	C/C
P75	59	90.3	156	37.1	6.1	C/C
P76	46	81.6	162	31.1	6.2	C/C
P77	59	125.3	162	47.7	9.2	C/C
P78	57	60.0	155	25.0	7.3	C/C
P79	49	100.2	166	36.3	10.3	C/C
P80	47	100.9	167	36.2	13.1	C/C
P81	50	84.4	157	34.2	9.8	C/C
P82	51	79.9	154	33.7	5.9	C/C

P83	49	60.6	152	26.2	15.8	C/C
P84	42	123.6	159	48.9	7.9	C/C
P85	59	89.2	163	33.8	8.4	C/C
P86	43	75.5	149	34.0	10.0	C/C
P87	59	74.9	152	32.6	8.0	C/C
P88	53	72.9	170	25.2	9.6	C/C
P89	41	69.6	164	25.9	8.2	C/C
P90	44	98.0	160	38.3	6.0	C/C
P91	47	89.0	155	37.0	7.1	C/C
P92	41	60.0	153	25.6	11.3	C/C
P93	60	82.0	145	39.0	9.4	C/C

Table 7: Table showing all control samples collected. Included is the participant number, age, HbA1c level (percentage), weight, height and BMI. The individuals selected for NGS are shaded in grey.

Participant Number	Age (years)	Weight (kg)	Height (cm)	BMI (kg/m^2)	HbA1c (%)	Genotype
C1	45	74.3	158	29.8	6.1	C/C
C2	48	85.5	157	34.7	4.9	C/C
C3	40	65.8	157	26.7	5.4	C/C
C4	41	81.3	161	31.4	5.0	C/C
C5	40	53.7	160	21.0	5.6	C/C
C6	49	71.0	151	31.1	4.8	C/C
C7	46	96.0	152	41.6	5.8	C/C
C8	41	114.0	162	43.4	5.8	C/C
C9	46	96.6	156	39.7	5.5	C/C
C10	49	69.5	154	29.3	4.3	C/C
C11	57	76.4	153	32.6	5.9	C/C
C12	59	109.6	164	40.7	5.6	C/C
C13	58	75.2	169	26.3	5.5	C/C
C14	54	95.8	163	36.1	4.9	C/C
C15	54	64.1	162	24.4	4.5	C/C
C16	59	99.5	151	43.6	5.5	C/C
C17	46	75.8	166	27.5	4.8	C/C
C18	39	67.4	148	30.8	5.2	C/C
C19	41	59.7	149	26.9	5.0	C/C
C20	40	79.0	163	29.7	5.9	C/C
C21	53	79.7	157	32.3	5.9	C/C
C22	58	82.9	168	29.4	6.1	C/C
C23	59	102.3	169	35.8	5.0	C/C
C24	53	93.7	152	40.5	5.4	C/C
C25	45	77.3	157	31.4	5.8	C/C
C26	57	83.9	149	37.8	5.8	C/C
C27	52	86.9	153	37.1	5.2	C/C

C28	43	83.4	160	32.6	5.3	C/C
C29	52	67.8	146	31.8	5.7	C/C
C30	59	52.0	146	24.4	5.6	C/C
C31	62	82.2	164	30.6	5.6	C/C
C32	46	81.7	149	36.8	5.8	C/C
C33	52	100.0	156	41.1	6.1	C/C
C34	53	95.3	166	34.6	5.3	C/C
C35	45	73.0	156	30.0	5.2	C/C
C36	57	111.9	159	44.3	5.4	C/C
C37	50	78.1	158	31.3	5.7	C/C
C38	51	64.0	148	29.2	5.4	C/C
C39	43	107.1	153	45.8	5.4	C/C
C40	58	79.0	158	31.7	5.5	C/C
C41	53	85.9	151	37.7	5.7	C/C
C42	41	105.0	149	47.3	5.8	C/C
C43	57	101.3	164	37.7	6.3	C/C
C44	54	89.0	165	32.7	5.8	C/C
C45	42	91.1	154	38.4	5.9	C/C
C46	52	81.2	168	28.8	5.6	C/C
C47	45	87.7	157	35.6	6.1	C/C
C48	61	77.1	155	32.1	5.8	C/C
C49	49	113.0	163	42.5	5.4	C/C
C50	48	63.1	164	23.5	5.6	C/C
C51	51	78.7	157	31.0	5.4	C/C
C52	62	70.8	162.2	28.0	4.9	C/C
C53	37	89.4	154	37.0	5.3	C/C
C54	44	105.9	174.4	33.0	6.4	C/C
C55	40	116.9	158.5	41.0	5.6	C/C
C56	42	87.4	159	35.0	5.4	C/C
C57	55	62.5	157	25.0	5.8	C/C
C58	37	101.0	150.3	45.0	4.9	C/C
C59	47	85.2	155.3	36.0	5.6	C/C
C60	37	91.0	167.2	32.0	5.4	C/C
C61	61	65.0	151.5	29.0	5.6	C/C
C62	48	89.9	160.5	35.0	5.5	C/C
C63	53	62.8	150	28.0	5.2	C/C
C64	62	59.2	145.2	28.0	5.4	C/C
C65	40	85.2	164.7	32.0	5.1	C/C
C66	60	99.3	154.5	41.0	5.6	C/C
C67	46	105.0	158.8	42.0	5.2	C/C
C68	56	110.0	156.4	44.0	5.3	C/C
C69	55	71.8	154.8	29.0	5.0	C/C
C70	52	58.3	146.8	26.0	5.6	C/C
C71	58	71.9	159	28.0	6.1	C/C
C72	52	83.0	154	34.0	5.5	C/C
C73	36	82.7	154	34.9	5.4	C/C

C74	40	65.0	152	28.1	4.8	C/C
C75	38	55.3	157	22.4	5.4	C/C
C76	49	120.1	164	44.7	5.6	C/C
C77	41	68.7	158	27.5	6.3	C/C
C78	38	97.0	168	34.3	5.2	C/C
C79	52	81.0	176	26.5	5.2	C/C
C80	53	86.5	160	33.8	5.3	C/C
C81	46	100.5	155	41.8	5.9	C/C
C82	62	65.9	171	22.5	5.7	C/C
C83	41	81.0	157	32.9	4.9	C/C
C84	53	70.3	159	27.8	6.0	C/C
C85	36	106.0	160	41.4	5.8	C/C
C86	44	64.5	152	27.0	5.4	C/C
C87	38	77.1	160	30.1	5.9	C/C
C88	49	72.0	151	31.6	5.8	C/C
C89	41	103.0	163	38.8	5.1	C/C
C90	44	103.9	157	42.2	5.7	C/C
C91	57	73.0	163	27.5	6.4	C/C

APPENDIX F

Inqaba Biotechnical Industries (Pty) Ltd
P.O. Box 14356, Hatfield 0028, South Africa
Tel: 012 343 5829
Fax: 012 343 0287
E-mail: info@inqaba.com

SYNTHESIS REPORT

04 Nov 2013

Client Detail: Charne Oosthuizen
University of the Free State
Department of Haematology and Cell Biology
Room nr. 409, Francois Retief building
Bloemfontein
9300
South Africa

Name:	positive control		**Barcode:** C4204	**Length:** 121 bases	
Sequence:	ATTCCATGCTGTTATGGGTGAAACTCTGGGAGATTCTCCTATTGACGCAGAAAGCGAT TCCTTCACTGATACACTGTCTGCAAACATATCACAAGGTAAAGTTCCTTCCAGATACGGCTAT				
OD	3.8285	**MW min \ max**	37224.6\37224.6	**5' Mod**	None
nmoles	2.91	**GC % min \ max**	42.98\42.98	**3' Mod**	None
Tm min \ max	78.39\78.39			**Purification**	Cartridge
For a 100 µM stock solution add 29.07 **µl water or buffer** pmol/ul				**PAGE QC Image >>**	
Comments					

RECOMMENDATIONS FOR HANDLING AND STORAGE OF OLIGOS

- Lyophilized oligo pellets might become displaced from the bottom of the tube during shipment. Briefly centrifuge each tube before opening to prevent the loss of the pellet.
- Prepare stock solution of oligos (e.g. 100 uM = 100 pmole per ul) preferably with a sterile buffered solution such as TE (10 mM Tris, pH 7.5 to 8.0, 1 mM EDTA). If sterile distilled water is used, make sure that the pH is above 7.0 since acidic solutions favours oligo depurination and subsequent loss of activity.
- Working solutions might be diluted from the stock solution with sterile, nuclease-free water to prevent inhibition of enzymatic reactions (e.g. PCR) by EDTA.
- Store the oligos as concentrated stock solution or lyophilized at -20° C.
- Avoid frequent freeze-thaw cycles by dividing the stock solution into smaller aliquots for long term storage and to prevent accidental contamination.
- Dye-modified oligos are light sensitive and should always be stored in the dark.
- Re-suspend modified oligos preferably in a slightly basic solution (i.e., TE at pH 8.0). However, Cy dye modified oligos are best kept at pH 7.0 at -20° C.
- Preferably store the modified oligos as dried aliquots at -20° C.

APPENDIX G

1. Quality Control Results

1.1 Input Information

Tables 8 and 9 show the Input Information for Control 1 (C1) consisting of the file name (Filename-C1.fastq) followed by the file format, the difference in number sequences and total bases.

Table 8: Input Information before QC.

Input file(s):	**Gerda-C1.fastq**
Input format(s):	**FASTQ**
# Sequences:	**214,906**
Total bases:	**48,384,147**

Table 9: Input Information after QC.

Input file(s):	**C1-QCed.fastq**
Input format(s):	**FASTQ**
# Sequences:	**191,373**
Total bases:	**36,556,410**

1.2 Length Distribution

The graphs generated in PRINSEQ show the mean length (M) and the length for one and two standard deviations (1SD and 2SD). The M, 1SD and 2SD were indicators of which parameters to select for the data pre-processing (Figure 1 and 2). Table 10 and 11 gives the length distribution statistics before and after QC was performed. The graphical representation of the length distribution before and after is illustrated in Figure 1 and 2.

Table 10: Length Distribution before QC.

Mean sequence length:	**225.14 ± 81.35 bp**
Minimum length:	**25 bp**
Maximum length:	**524 bp**
Length range:	**500 bp**
Mode length:	**212 bp with 5,237 sequences**

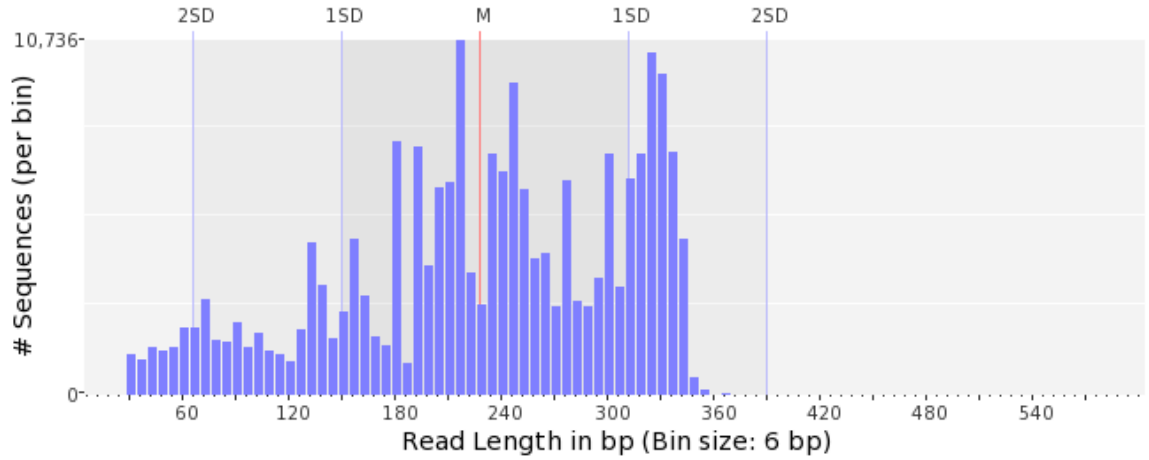


Figure 1: Length distribution in a graphical format before QC was performed.

Table 11: Length Distribution after QC.

Mean sequence length:	**191.02 ± 75.15 bp**
Minimum length:	**40 bp**
Maximum length:	**452 bp**
Length range:	**413 bp**
Mode length:	**177 bp with 3,029 sequences**

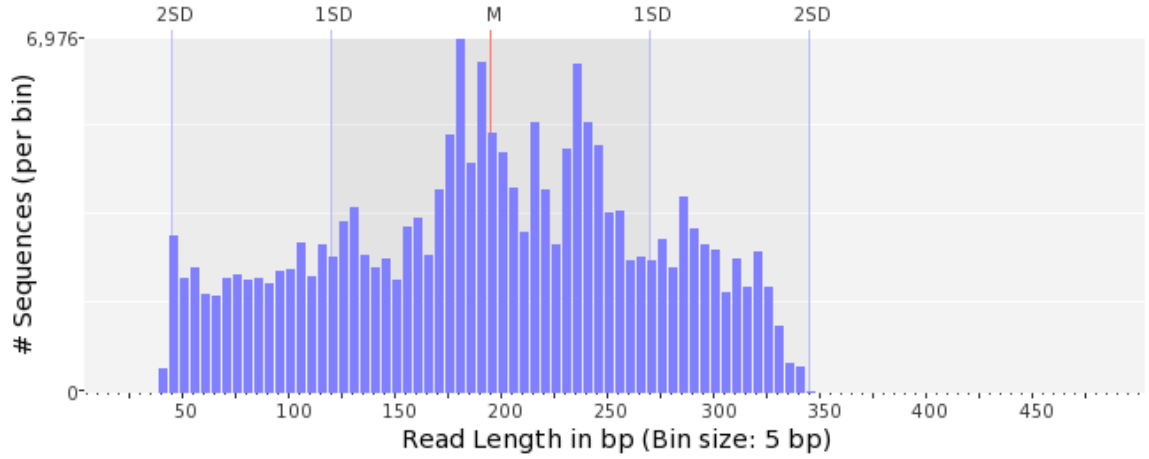


Figure 2: Length distribution in graphical format after QC was performed.

1.3 Base qualities

Figure 3 and 5 below indicate the quality score across the reads position (in %) before and after QC. PRINSEQ provides an additional plot that shows the distribution of sequence mean quality scores of a dataset, indicated by Figure 4 (before QC) and Figure 5 (after QC).

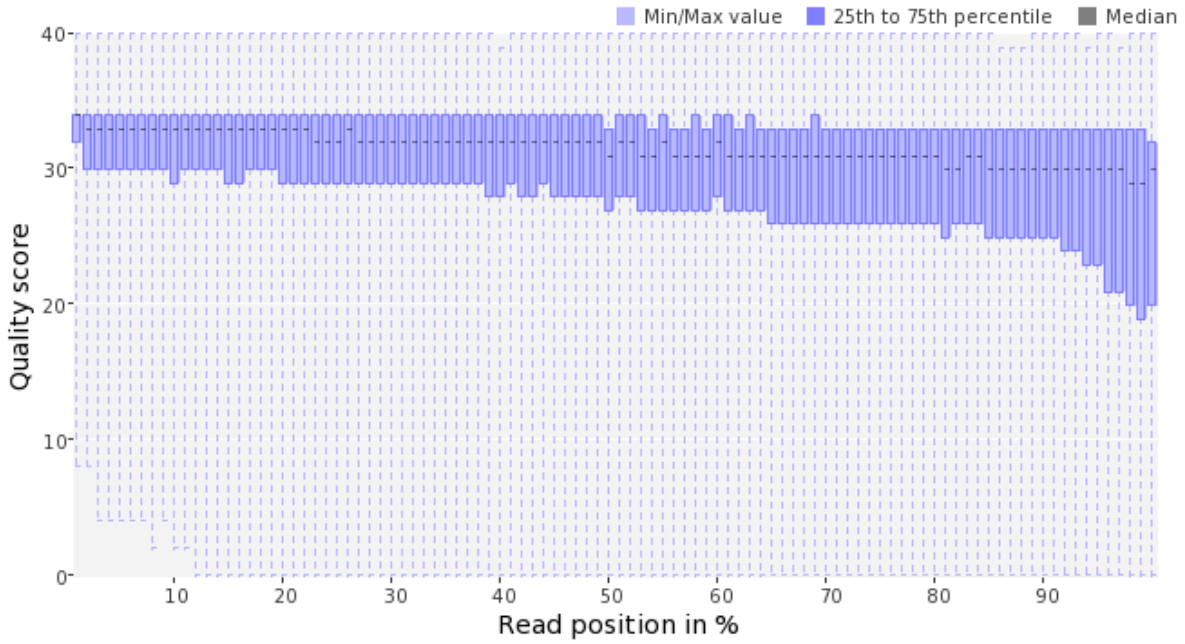


Figure 3: This graph shows the quality scores of the sequences across the read length before QC.

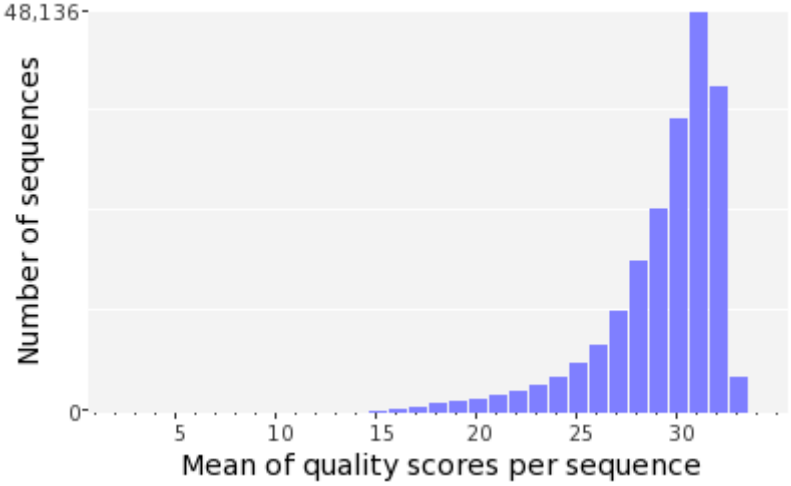


Figure 4: The distribution of mean sequence quality scores of a dataset before QC.

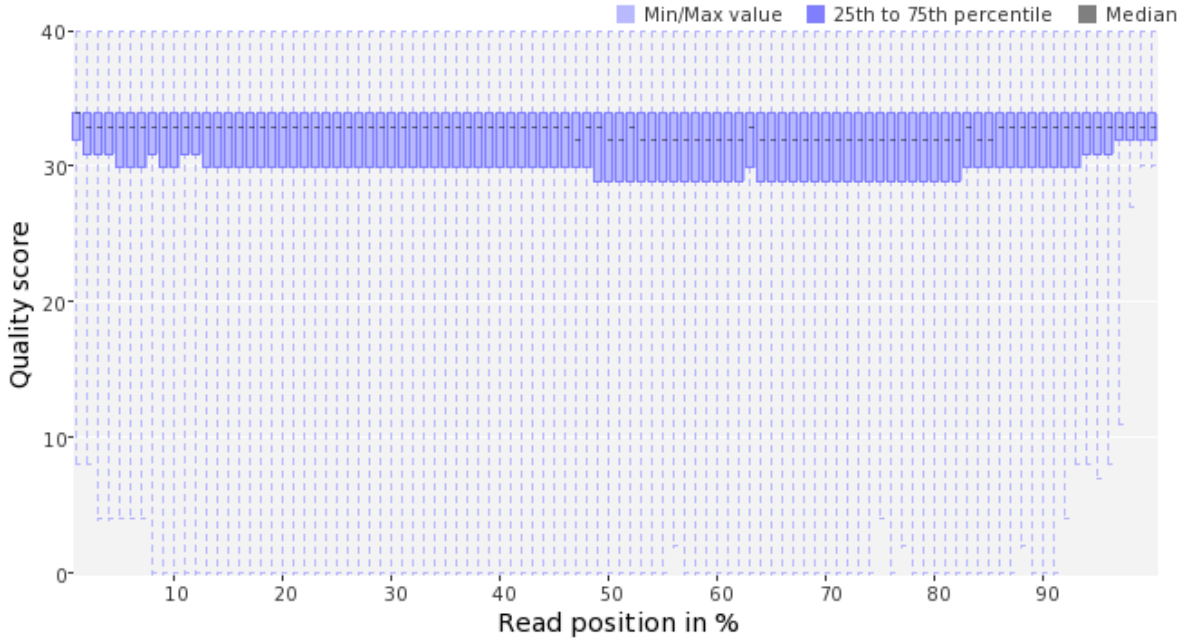


Figure 5: This graph shows the quality scores of the sequences across the read length after QC.

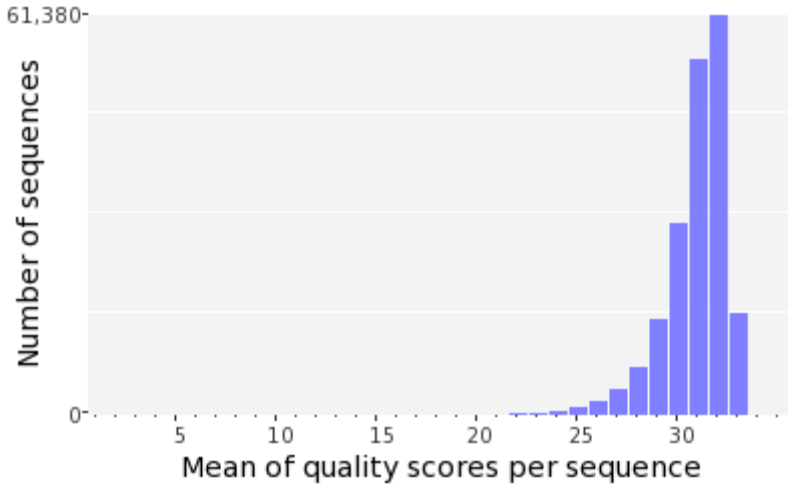


Figure 6: Graph that shows the distribution of mean sequence quality scores of a dataset after QC.

1.4 Sequence Duplication

The following plots show the number of sequence duplicates for different lengths before and after QC. Table 4.11 and 4.12 provides a summary of the data depicted in Graph 4.9 and 4.10. The tables show the different duplicated which were detected as well as the total amount of duplicates in the sequence data and how it changed after QC was performed.

Table 12: Sequence Duplication before QC.

	# Sequences	Max duplicates
Exact duplicates:	**129,403 (60.21 %)**	**2094**
Exact duplicates with reverse complements:	**1,338 (0.62 %)**	**3**
5' duplicates	**13,506 (6.28 %)**	**11**
3' duplicates	**2,048 (0.95 %)**	**6**
5'/3' duplicates with reverse complements	**935 (0.44 %)**	**2**
Total:	**147,230 (68.51 %)**	**-**

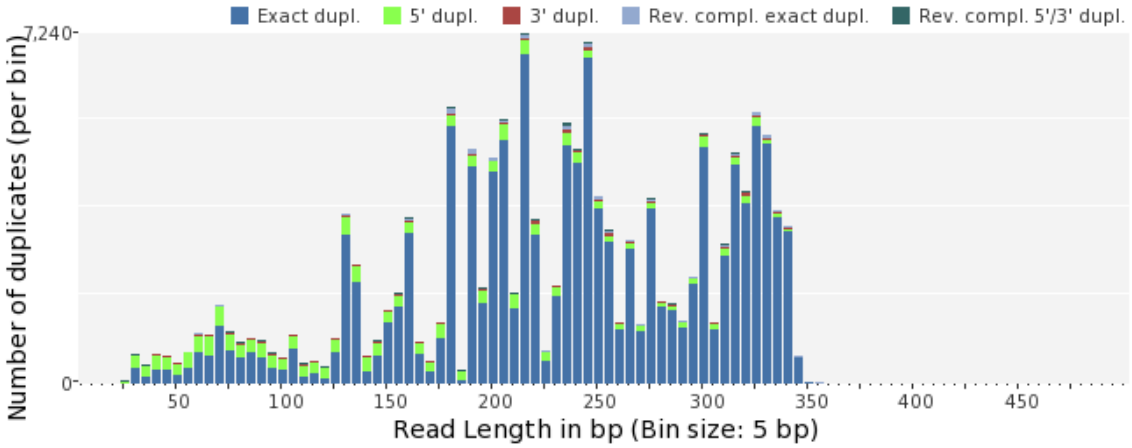


Figure 7: Graph showing the number of duplicates for the different read lengths before QC. The different duplicates are indicated by the different colours seen in the heading.

Table 13: Sequence Duplication after QC.

	# Sequences	Max duplicates
Exact duplicates:	**128,501 (67.15 %)**	**1281**
Exact duplicates with reverse complements:	**107 (0.06 %)**	**2**
5' duplicates	**28,223 (14.75 %)**	**14**
3' duplicates	**957 (0.50 %)**	**4**
5'/3' duplicates with reverse complements	**1,018 (0.53 %)**	**2**
Total:	**158,806 (82.98 %)**	**-**

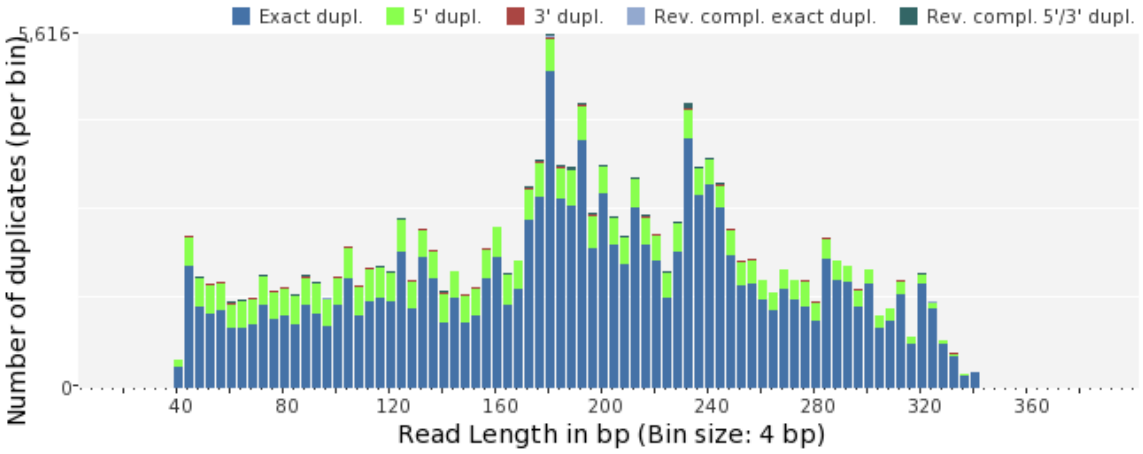


Figure 8: Graph showing the number of duplicates for the different read lengths after QC. The different duplicates are indicated by the different colours seen in the heading.

1.5 Tag Sequences

Figure 9 shows the base frequencies across the reads present. If the distribution of nucleotides is uneven, it could indicate some residual tag sequences. The equal distribution of the different nucleotides indicates that no sequence tags were present. This was expected due to pre-analysis with Ion Torrent Suite™.

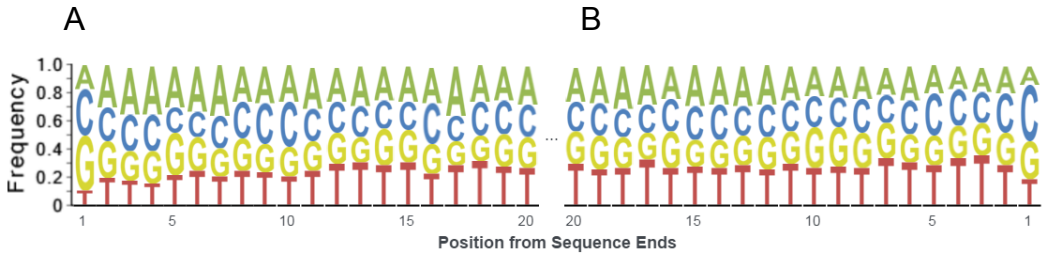


Figure 9: Illustrate the frequency of each base at the specific position at the 5' end (A) and 3' end (B).

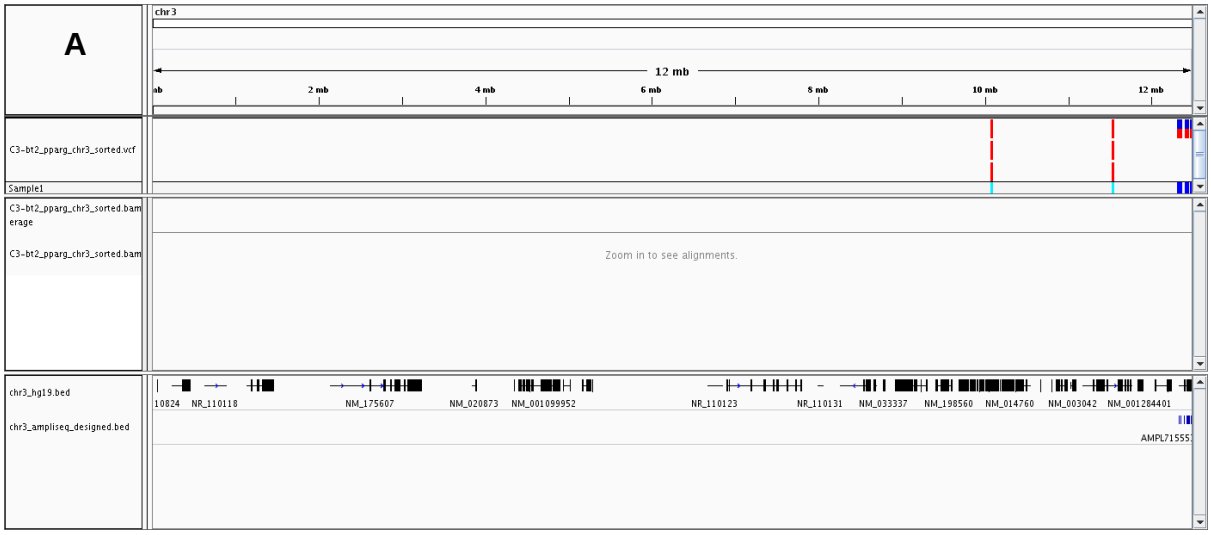
2. Visualization of NGS data

Figure 10 shows a series of images saved as PNG files from the IGV panel during visualization of analysed data. These figures illustrate different aspects of the visualization process and give an oversight of how the SNP's in each sample were visually analysed and confirmed. Figure 10 shows a succession of images from one

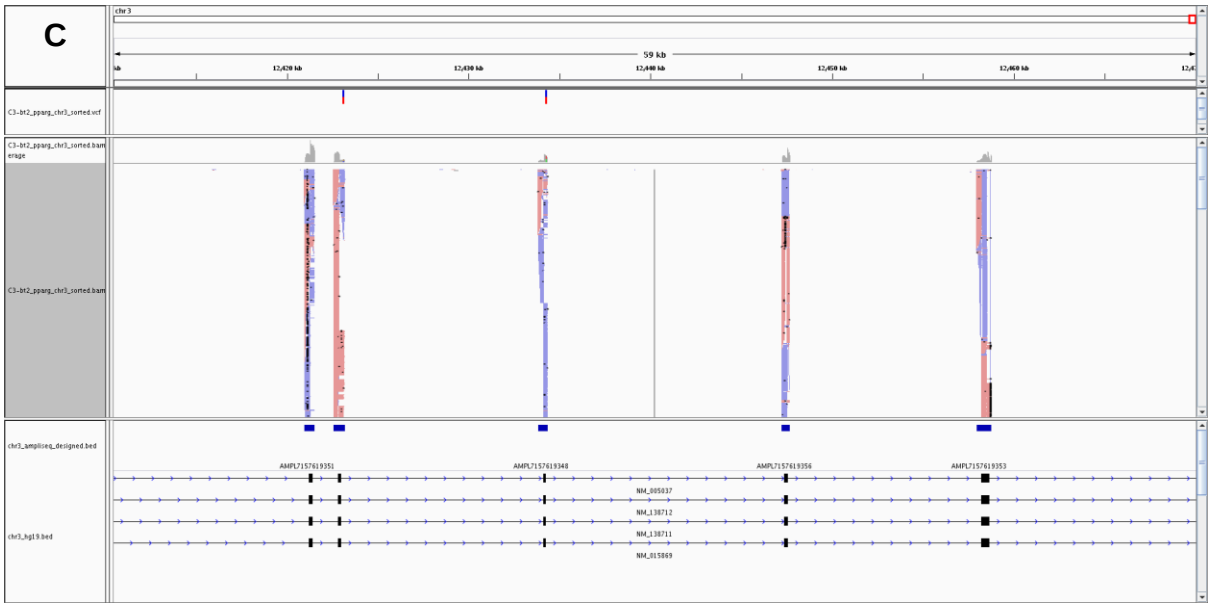
sample, Control 3 (C3) mapped against Chr3 (FASTA file) with increased magnification. C3 was selected as an example because this was one of three individuals containing a SNP within the coding region of *PPARG*.

BED files were created in UCSC of each reference file that was used for mapping. The BED files were loaded as additional tracks in IGV to allow for easy visualization and comparison of the read sequences to the references. For this example, a FASTA file of Chr3 was loaded from our file system to IGV as a reference.

Figure 10 Image A shows a 12 Mb section of Chr3 with the different genes present on that part of the chromosome. The *PPARG* gene is located at the end of this part of Chr3.The AmpliSeq™ design is indicated in blue below the Chr3 BED file. The VCF file indicates the mutations that were called by Varscan. The sample track is only visible with a higher magnification. A closer view, Image B shows an expanded view of the *PPARG* gene with the different isoforms. The AmpliSeq™ design is clear on this image, showing the regions the design covers. The AmpliSeq™ design perfectly aligns to the exons seen on the different isoforms. Image C displays the mapped reads the *PPARG* gene. The forward and reverse reads are depicted in pink and purple respectively. The read depth is indicated in grey above the mapped reads. Image D is a view of the mapping across the last exon (exon 6) and the 3' UTR of *PPARG*. Two SNP's are called on this image, the first is present in exon 6 and the second outside the *PPARG* gene. This image also show the 100 bp padding added by the AmpliSeq™ design where the ampliseq track overlaps. Image E is a close view of an SNP (variant) called by Varscan. As indicated by the combination of red and blue this is a heterozygous SNP. The yellow colour variation across the reads indicate where the sequence is different and specifically illustrate an A to G change. The translation is indicated in all three frames.


A
chr 3
12 mb
2 mb
4 mb
6 mb
8 mb
10 mb
12 mb
C3-bt2_pparg_chr3_sorted.vcf
Sample1
C3-bt2_pparg_chr3_sorted.bam
erage
C3-bt2_pparg_chr3_sorted.bam
Zoom in to see alignments.
chr3_hg19.bed
chr3_ampliseq_designed.bed
10824
NR_110118
NM_175607
NM_020873
NM_001099952
NR_110123
NR_110131
NM_033337
NM_198560
NM_014760
NM_003042
NM_001284401
AMPL71555


B
chr 3
238 kb
12,260 kb
12,280 kb
12,300 kb
12,320 kb
12,340 kb
12,360 kb
12,380 kb
12,400 kb
12,420 kb
12,440 kb
12,460 kb
12,480 kb
C3-bt2_pparg_chr3_sorted.vcf
Sample1
C3-bt2_pparg_chr3_sorted.bam
erage
C3-bt2_pparg_chr3_sorted.bam
Zoom in to see alignments.
chr3_ampliseq_designed.bed
chr3_hg19.bed
AMPL7155530187
AMPL7155530183
AMPL7157619349
AMPL7157619351
AMPL7157619356
AMPL7156749757
NM_005037
NM_138712
NM_138711
NM_015869


C
chr 3
59 kb
12,420 kb
12,430 kb
12,440 kb
12,450 kb
12,460 kb
C3-bt2_pparg_chr3_sorted.vcf
C3-bt2_pparg_chr3_sorted.bam
erage
C3-bt2_pparg_chr3_sorted.bam
chr3_ampliseq_designed.bed
chr3_hg19.bed
AMPL7157619351
AMPL7157619348
AMPL7157619356
AMPL7157619353
NM_005037
NM_138712
NM_138711
NM_015869

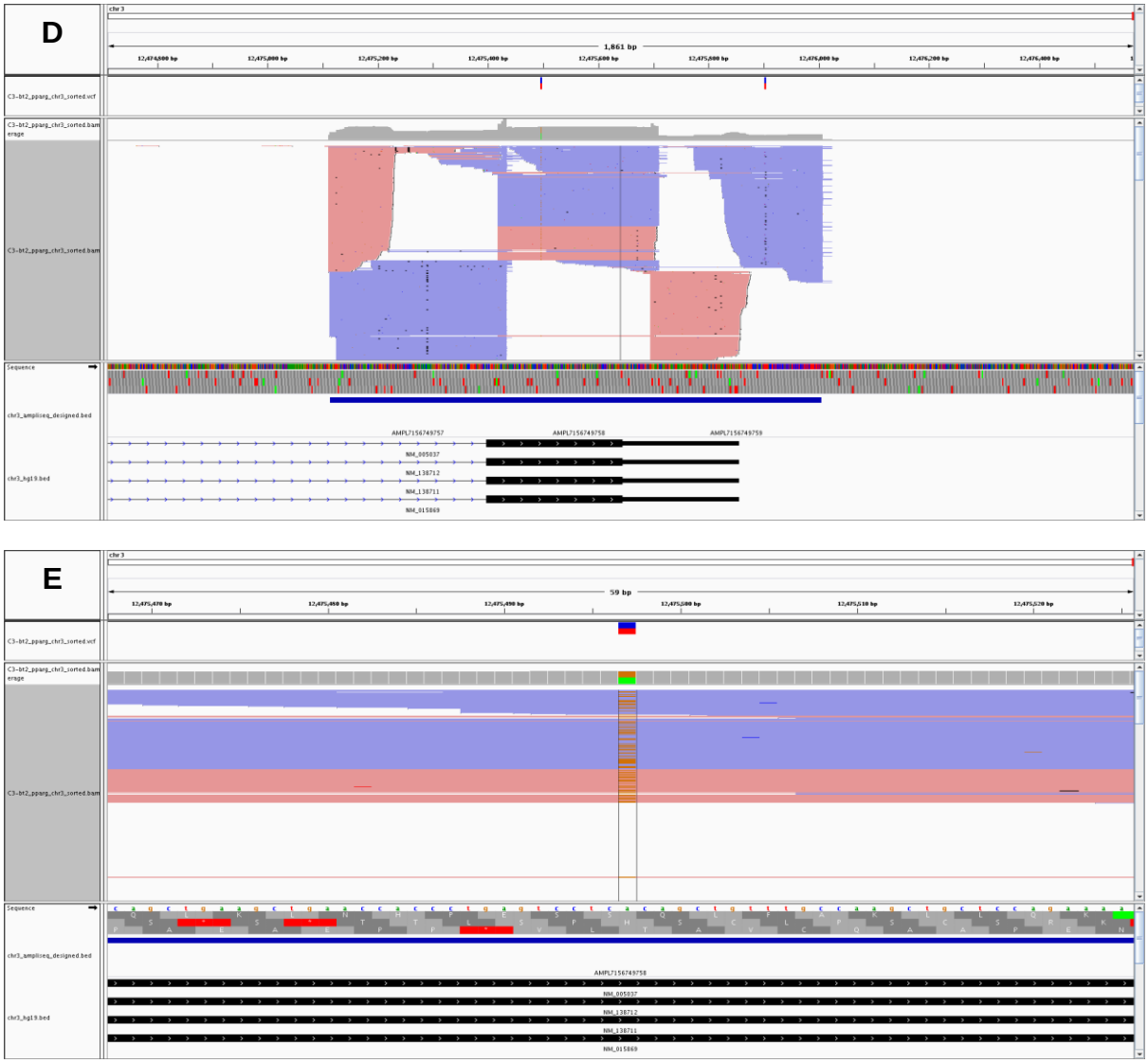


Figure 10 (A-E): This figure shows a series of PNG images from the IGV server illustrating different features and information regarding data analysis.

A- A 12 Mb section of Chr3. The Chr3 BED file is loaded in the IGV track and the AmpliSeq™ design track is indicated in blue below the BED file.

B- A closer view, showing the three *PPARG* isoforms in the track, with the AmpliSeq™ design corresponding to the *PPARG* exons. The variants called are indicated in red and blue.

C- The reads are displayed mapping to the *PPARG* exons**.**

D- An enlarged view of the 3' end of the *PPARG* gene**.** The read depth is indicated in grey with forward and reverse reads depicted in pink and purple respectively.

E- A magnified image illustrating an SNP called by Varscan. The translation of the sequence is shown in all three frames**.**

APPENDIX H

Table 14: List of all mutations found in the *PPARG* gene. These include mutations found in the exons, intronic regions, UTR's and 100bp padding outside (o/s) the *PPARG* gene. The mutations discussed in the results section is highlighted in grey.

Participant	Number	Position	Reference	Alternate	Heter/Homo	ADP (read depth)	Intron, Exon, UTR, 100bp padding	Exon
C1	1	12329562	G	C	Heter	276	Intron	
	2	12475904	T	C	Heter	47	100 bp padding o/s PPARG	
C2	1	12330143	T	G	Heter	145	Intron	
	2	12330574	A	T	Heter	401	Untranslated Exon	A1
	3	12354030	T	C	Heter	340	Intron	
	4	12421445	T	C	Heter	1343	Intron	
	5	12475794	A	G	Heter	163	3' UTR	
	6	12475904	T	C	Heter	29	100 bp padding o/s PPARG	
C3	1	12329562	G	C	Heter	77	Intron	
	2	12353993	G	A	Heter	317	Intron	
	3	12423113	G	C	Heter	222	Intron	
	4	12434272	A	T	Heter	471	Intron	
	5	12475497	A	G	Heter	561	Exon	Exon 6
	6	12475904	T	C	Heter	25	100 bp padding o/s PPARG	
C4	1	12475162	A	C	Heter	331	Intron	
C5	1	12353993	G	A	Heter	75	Intron	
	2	12423113	G	C	Heter	46	Intron	
	3	12475162	A	C	Heter	154	Intron	
	4	12475794	A	G	Heter	69	3' UTR	
C6	1	12353993	G	A	Heter	673	Intron	
	2	12423113	G	C	Heter	707	Intron	
	3	12475557	C	T	Heter	400	Exon	Exon 6

	4	12475904	T	C	Heter	45	100 bp padding o/s PPARG	
C7	1	12475162	A	C	Heter	1042	Intron	
	2	12475927	C	T	Heter	209	100 bp padding o/s PPARG	
C8	1	12330143	T	G	Heter	283	Intron	
	2	12330574	A	T	Heter	196	Untranslated Exon	A1
	3	12475794	A	G	Heter	439	3' UTR	
	4	12475904	T	C	Heter	67	100 bp padding o/s PPARG	
P1	1	12330143	T	G	Heter	225	Intron	
	2	12330574	A	T	Heter	428	Untranslated Exon	A1
	3	12423113	G	C	Heter	629	Intron	
	4	12434070	G	C	Heter	441	Intron	
P2	1	12475904	T	C	Heter	49	100 bp padding o/s PPARG	
	2	12475927	C	T	Heter	97	100 bp padding o/s PPARG	
P3	1	12329562	G	C	Heter	264	Intron	
	2	12353993	G	A	Heter	631	Intron	
	3	12423113	G	C	Heter	718	Intron	
	4	12434070	G	C	Heter	563	Intron	
P4	1	12330143	T	G	Heter	342	Intron	
	2	12330574	A	T	Heter	520	Untranslated Exon	A1
	3	12353993	G	A	Heter	774	Intron	
	4	12421089	C	T	Heter	1322	Intron	
	5	12423113	G	C	Heter	728	Intron	
	6	12475497	A	G	Heter	1579	Exon	Exon 6
P5		NONE						
P6		NONE						
P7	1	12434272	A	T	Heter	2161	Intron	
	2	12475904	T	C	Heter	33	100 bp padding o/s PPARG	
P8		NONE						

END OF DISSERTATION